N°d'ordre: .....

République Algerienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la Recherche scientifique

Université Djillali Liabès de Sidi Bel Abbès
Faculté des Sciences Exactes
Département d'informatique
Laboratoire EEDIS

Thèse de Doctorat

Domaine : Mathématiques Informatique
Filière : Informatique
Spécialité : Technologie de l'information

Par

Mlle DIF Nassima

L'apprentissage profond pour le traitement d'images

Soutenue le 23-12-2020 devant le jury :

Dr.	Farah Ben naoum	Université Djillali Liabes	Président du jury
Dr.	Djamel Berrabah	Université Djillali Liabes	Examinateur
Pr.	Ghalem belalem	Université d'Oran 1 Ahmed Ben Bella	Examinateur
Pr.	Elberrichi Zakaria	Université Djillali Liabes	Directeur de thèse

Année Universitaire : 2020 - 2021

Je dédie ce modeste travail à :
Mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,
Ma chère soeur Amina pour son encouragement permanent, et son soutien moral et financier, c'est grâce à elle que j'ai pu avoir une machine puissante pour réaliser toutes les expérimentations de cette thèse,
Ma chère soeur Wafaa, pour son amour, et son soutien moral.
Mes chers frères Mustapha, Abderrahmane, et Moussa pour leur appui et leur encouragement,
Mes chéres amies Asma et Narimene pour leur soutien et leur envouragement,
Que ce travail soit l'accomplissement de vos voeux tant allégués, et le fuit de votre soutien infaillible,
« MERCI » d'être toujours là pour moi.

Remerciement

Il me sera très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme.

Je voudrais tout d'abord remercier grandement mon directeur de thèse, Pr. Zakaria Elberrichi, pour toute son aide. Je suis ravi d'avoir travaillé en sa compagnie car outre son appui scientifique, il a toujours été là pour me soutenir et me conseiller au cours de l'élaboration de cette thèse.

J'adresse mes remerciements à Messieurs Ghalem Belalem, Djamel Berrabah ainsi que Madame Farah Ben Naoum de m'avoir fait l'honneur d'examiner ce travail de thèse et d'avoir accepté de faire partie des membres du jury. J'apprécie l'intérêt qu'ils ont porté à mes travaux.

Je tiens à remercier le staff du département d'informatique de Sidi Bel Abbes. Je pense notamment au Dr Sofiane Boukli Hacene le directeur du laboratoire pour ces précieux services, son soutien, sa disponibilité. Je remercie également Mr. Rafik Belkhodja pour ces précieux conseils et encouragements.

Je tiens à remercier particulièrement mon frère et mon mentor Dr. Mustapha Mahmoud DIF pour toutes nos discussions et ses conseils qui m'ont accompagné tout au long de mon cursus.

Je remercie toutes les personnes avec qui j'ai partagé mes études et notamment ces années de thèse.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Dans cette introduction nous présenterons le contexte de notre étude à savoir les méthodes d'apprentissage profond. Nous nous focaliserons sur le domaine de vision par ordinateur qui constitue le noyau de plusieurs systèmes automatiques. Nous nous intéresseront notamment à la catégorisation des images histopathologiques par différentes techniques d'apprentissage profond. Nous exposerons essentiellement la problématique concernant l'optimisation des réseaux de neurones convolutif pour résoudre les différents problèmes liés à la classification de ces images. Nous passerons ensuite à une analyse des besoins justifiant notre contribution à travers ce travail de thèse.

Contexte de l'étude

La vision par ordinateur est une branche de l'intelligence artificielle. Elle permet à un ordinateur d'analyser, de traiter, et de comprendre les images. Les systèmes de vision sont exploités pour extraire des informations pertinentes à partir des entrées visuelles (image ou vidéo) afin de les utiliser dans d'autres tâches de recommandation.

La reconnaissance des entrées visuelles par les humains nécessite moins d'effort par rapport aux systèmes automatiques. Avec le développement de l'internet et des réseaux sociaux, la quantité des images a rapidement augmenté. Par conséquent, le traitement de cette quantité d'information par les êtres humains devient impossible, car ils ne peuvent pas traiter efficacement autant de données. Ces informations sont donc traitées automatiquement à l'aide des systèmes de vision par ordinateur.

Le but de la vision par ordinateur est de développer des méthodes pour reproduire des systèmes qui ont une capacité équivalente à la vision humaine. Malgré les efforts faits, le domaine de vision par ordinateur a plusieurs défis liés à la compréhension limitée des systèmes de vision biologiques et leurs complexités très élevées par rapport aux machines actuelles.

Les systèmes de prédiction en vision par ordinateur sont basés sur les algorithmes d'apprentissage automatique (ML) et d'apprentissage profond (DL). Ils permettent d'analyser les entrées visuelles prises par un système d'acquisition. Ces algorithmes sont entrainés sur des données pour produire des modèles en sortie. Les modèles générés sont exploités ensuite dans la phase de prédiction.

L'apprentissage profond (DL) est une branche de l'apprentissage automatique basée sur les réseaux de neurones artificiels (ANN). Ces réseaux ont été exploités en apprentissage supervisé et non supervisé. Afin

d'optimiser les coûts de stockage et de calcul des réseaux DL classiques, plusieurs types d'architectures ont été proposés en apprentissage

supervisé (les réseaux de neurones convolutif (CNN), les réseaux de neurones récurrents (RNN), mémoire à long-court terme (LSTM)) et non supervisé (les auto-encodeurs empilés (SAE), réseaux de croyances profondes (DBN), les réseaux contradictoires génératifs GAN). Dans cette thèse, nous nous intéressons aux CNN.

Contrairement aux méthodes ML classiques, les méthodes DL et particulièrement les CNN sont plus adaptés aux données complexes, car ils intègrent la phase d'extraction des caractéristiques dans le processus de l'apprentissage. Ces réseaux sont caractérisés par des couches de convolution et de pooling par rapport aux réseaux DL classiques. Ces couches introduisent des liens partiels pour réduire le nombre des paramètres et renforcer le partage des caractéristiques communes.

Récemment, les réseaux de neurones convolutif ont été largement exploités en vision par ordinateur grâce à leur stratégie de réduction de paramètres et la disponibilité des grands volumes de données. En plus, l'évolution de la capacité de stockage et de calcul a encouragé la communauté de vision par ordinateur à proposer d'autres architectures de type CNN plus profondes. En classification, les architectures proposées optimisent les couches de convolution classiques. Le but principal de cette variation est de réduire le nombre des paramètres et d'ajouter des couches supplémentaires qui permettent d'améliorer la non-linéarité. En détection et en segmentation, le but principal était d'adapter les architectures proposées aux applications en temps réel.

En raison du progrès important des architectures CNN, ils ont été exploités dans plusieurs applications du monde réel, comme : la classification des images, la détection et localisation des objets, la segmentation sémantique, la reconnaissance d'action et d'activité, l'estimation de la pose humaine, la reconnaissance faciale, et le traitement des images médicales.

En imagerie médicale, les images numérisées ont plusieurs types comme : ultrason (US), rayon-X, tomodensitométrie (CT) et imagerie par résonance (MRI), tomographie par émission de positrons (PET), et lames histologiques. Les CNN sont entrainés sur ces images pour résoudre différentes tâches en imagerie médicale : classification, localisation, détection, et segmentation.

Les méthodes proposées dans cette thèse s'intéressent à la résolution des problèmes liés à la classification des images médicales et notamment les images histopathologiques par les réseaux CNN.

Problématique

L'histopathologie est une branche de l'histologie, où les tissus et les cellules biologiques malades sont examinés sous le microscope. Récemment, les biopsies sont numérisées sous forme d'images en champ large (Whole slide images(WSI)) par les scanners de lame entière (Whole slide digital scanners (WSD)). Ces scanners sont des outils puissants pour la numérisation, l'acquisition et le partage des WSI. L'analyse des images histopathologiques est une étape non triviale dans le diagnostic des cancers. D'autre part, l'examen manuel de ces images a plusieurs enjeux liés à

la subjectivité des décisions des pathologistes et l'apparence variable des images issues de différents laboratoires. Afin d'aider les pathologistes et d'éviter les décisions subjectives, les systèmes d'aide au diagnostic (CAD) sont exploités.

Les CAD sont basés sur les méthodes ML et DL. L'avantage des méthodes DL par rapport aux méthodes ML, notamment dans l'analyse des images histopathologiques, est leur capacité d'extraction des caractéristiques. En revanche, les méthodes ML exigent un prétraitement d'extraction de caractéristiques (handcrafted features), d'où le besoin d'un expert dans le domaine pour décider des attributs discriminants. En plus, les caractéristiques diffèrent d'un sous domaine à un autre, et cela complique la conception d'un outil standard pour le traitement des images médicales.

Malgré les avantages des méthodes DL par rapport aux méthodes ML et le succès des CNN dans différents domaines, ces derniers ont plusieurs défis liés aux problèmes de sur-apprentissage sur les volumes limités de données, leurs exigences en termes de stockage et de capacité de calcul, et le temps considérable d'inférence en détection et en segmentation. Il est intéressant de noter que malgré la disponibilité des WSD et leur avantage dans la collecte de données, le nombre des images histopathologiques disponibles reste limité pour les applications de type DL. En plus, l'annotation de ces données et surtout en segmentation nécessite un effort considérable par les pathologistes, d'où le besoin des méthodes automatiques pour la résolution de ces problèmes.

Plusieurs efforts ont été faits dans l'état de l'art pour adapter les méthodes DL, notamment les CNN, en apprentissage sur les volumes limités de données histopathologiques :

— **L'apprentissage transféré à partir des modèles pré-entrainés sur la base d'apprentissage ImageNet :** cette technique réduit les différents problèmes liés au sur-apprentissage, car seulement un sous ensemble de couches est entrainé sur la nouvelle tâche. En plus, ce traitement permet de diminuer le temps considérable d'apprentissage des méthodes DL.

— **Les méthodes d'augmentation de données :** cette technique permet de générer plusieurs images à partir d'une seule image originale. L'augmentation de données a été largement exploitée dans l'analyse des images histopathologiques en raison de leur grande résolution et les formes similaires des structures biologiques sur les WSI. Elle permet de générer une quantité considérable de données et donc elle réduit les problèmes de sur-apprentissage.

— **Les méthodes de normalisation de couleurs :** la normalisation de couleurs est utilisée afin de transformer les images générées de différents laboratoires et traitées dans différentes conditions à un espace normalisé. Cette technique permet de réduire la variabilité entre les laboratoires et d'améliorer la généralisation des modèles entrainés sur ces images.

— **L'hybridation entre les méthodes ML et DL :** dans ce cadre, plusieurs travaux ont proposé l'exploitation des CNN pour l'extraction des caractéristiques et des algorithmes ML pour la classification. D'autres investigations suggèrent l'hybridation entre les attri-

buts CNN et les handcrafted features. L'objectif de cette hybridation est de prendre avantage des réseaux CNN en raison de leur capacité d'extraction des caractéristiques et des méthodes ML en raison de leur adéquation aux volumes limités de données, et donc elle réduit les différents problèmes de sur-apprentissage. En plus, elle diminue les exigences des réseaux DL en termes de capacité de stockage et de temps d'apprentissage.

— **Méthodes de régularisation :** les méthodes de régularisation comme : les régularisations L1 et L2, la régularisation par abandon (Dropout), l'augmentation de données, et l'arrêt prématuré ont été largement exploité dans les applications DL. Ces techniques permettent de réduire la grande variance entre les performances sur les données d'apprentissage et de validation et donc elles présentent un bon outil pour l'amélioration de la généralisation des modèles entrainés.

— **Apprentissage semi-supervisé :** L'apprentissage semi-supervisé présente une autre solution au manque de données annotées. Cette technique permet d'entrainer des modèles sur des données étiquetées et d'autres non étiquetées. Cette spécificité encourage l'exploitation de cette technique comme alternative aux méthodes d'apprentissage supervisé dans le cas des quantités limitées de données classifiées.

Malgré les efforts faits, plusieurs enjeux subsistent dans le domaine de l'analyse des images histopathologiques, comme :

— Le choix de l'architecture optimale pour la résolution de la tâche en question.

— Le choix des hyper-paramètres convenables.

— Le choix du modèle approprié parmi plusieurs modèles enregistrés durant le processus d'apprentissage itératif.

— La réduction d'intra-variabilité et d'inter-variabilité entre les décisions de différentes architectures de type CNN.

— Le lien entre la base d'apprentissage ImageNet et les bases d'apprentissage histopathologiques en apprentissage transféré.

Ainsi la problématique abordée peut être résumée en plusieurs grandes interrogations :

— **Question 1 :** Comment choisir l'architecture convenable pour éviter le problème de sur-apprentissage lié au manque de données histopathologiques ?

— **Question 2 :** Comment guider le processus d'apprentissage en suivant plusieurs méthodes de régularisation ?

— **Question 3 :** Comment exploiter efficacement la décision de plusieurs modèles de type CNN, et notamment sélectionner les modèles qui ont plus d'impact sur la décision de l'ensemble ?

— **Question 4 :** Comment exploiter différemment les techniques d'apprentissage transféré dans le cadre des images histopathologiques ?

Contribution

Afin de répondre à la problématique exposée ci-dessus, nous proposons à travers la présente thèse une contribution théorique, et trois contributions expérimentales. Le but de la contribution théorique est de présenter le processus complet de la détection automatique de la mitose à partir des échantillons histopathologiques [Dif & Elberrichi 2020a]. D'autre part, les contributions expérimentales proposent un framework de régularisation, de nouvelles méthodes ensemblistes, et une méthode d'apprentissage transféré dans le cadre de classification des images histopathologiques. Il est intéressant de noter que les deux premières contributions expérimentales[Dif & Elberrichi 2020c, Dif & Elberrichi 2020b] exploitent les techniques de régularisation, d'apprentissage transféré classique, et des méthodes ensemblistes. Tandis que la troisième contribution [Dif & Elberrichi 2020d] s'intéresse à l'apprentissage transféré entre les bases d'apprentissage histopathologiques.

Contribution théorique : Les méthodes d'apprentissage profond pour la détection de la mitose à partir des images histopathologiques du cancer du sein : un aperçu complet

Le but de cette contribution [Dif & Elberrichi 2020a] est de présenter les connaissances médicales liées à la détection de la mitose à la communauté de l'intelligence artificielle, d'expliquer de comparer les méthodes DL proposées pour la détection de la mitose sur les images histopathologiques, et enfin de discuter les défis et les perspectives. En résumé, les travaux proposés exploitent les stratégies suivantes :

— Les stratégies de régularisation pour réduire les problèmes de sur-apprentissage.

— Les techniques de l'apprentissage transféré, fine tuning, et l'exploitation des CNN en tant qu'extracteur de caractéristiques afin de réduire la complexité temporelle en apprentissage et de résoudre les problèmes de sur-apprentissage.

— L'exploitation du réseau fully convolutional network (FCN) et les méthodes d'apprentissage profond en détection pour optimiser la complexité de détection en inférence.

— les réseaux de régression pour réduire le temps d'inférence.

— L'apprentissage multi-échelles pour améliorer le processus de détection.

— Les stratégies d'apprentissage en deux phases pour résoudre le problème du taux élevé des faux positifs.

Contribution expérimentale 1 : Un framework de régularisation pour la classification des images histopathologiques à l'aide des réseaux de neurones convolutifs

Cette contribution [Dif & Elberrichi 2020b] est un framework de régularisation pour la classification des images histopathologiques. Ce travail introduit un framework de régularisation qui regroupe un maximum de

méthodes de régularisation afin de réduire le problème de la variance élevée des réseaux CNN, et cela permet de diminuer le problème sur-apprentissage sur les volumes limités de données histopathologiques. Cette contribution répond sur les trois premières problématiques exposées dans la section précédente de la manière suivante :

Réponse à la question 1 : le sur-apprentissage dépend du nombre de paramètres du réseau, et généralement, les modèles peu profonds sont exploités en apprentissage sur les volumes limités de données, car ces derniers sont caractérisés par un nombre réduit de paramètres par rapport aux réseaux plus profonds. D'autre part, les réseaux plus profonds sont caractérisés par leur efficacité dans la résolution des problèmes complexes. Afin de réaliser un compromis entre la profondeur et le nombre de paramètres, nous avons choisi d'exploiter des architectures qui ont été proposées spécialement pour les systèmes de vision intégrés (MobileNetV1, MobileNetV2). Ces architectures introduisent des techniques de réduction de dimensionnalité à travers de nouveaux modules de convolution et les deux hyper-paramètres : multiplicateur de largeur et multiplicateur de résolution.

Réponse à la question 2 : nous avons choisi d'exploiter un maximum de méthodes de régularisation, comme : l'augmentation de données, les petits (small) modèles (MobileNetV1, MobileNetV2), la sélection de la méthode d'optimisation appropriée en apprentissage (SGD, RmsProp), et les méthodes ensemblistes. Afin de sélectionner un modèle efficace en généralisation, nous avons commencé par la comparaison entre les performances de la méthode d'apprentissage transféré et de la technique d'apprentissage à partir des initialisations aléatoires. Ensuite, nous avons exploité la technique d'apprentissage appropriée pour comparer entre les performances des modèles MobileNetV1 entrainés à base des optimiseurs SGD et RmsProp. Dans l'étape qui suit, nous avons choisit l'optimiseur approprié en apprentissage pour comparer entre les performances des modèles MobileNetV1 et MobileNetV2. Ensuite, nous avons comparé entre les performances du meilleur modèle avec et sans dropout. Enfin, nous avons exploité ce modèle dans l'étape suivante d'apprentissage ensembliste.

Réponse à la question 3 : les méthodes ensemblistes sont parmi les solutions utilisées pour combiner la décision de plusieurs modèles. Les réseaux d'apprentissage profond sont caractérisés par leur instabilité et dépendance de plusieurs conditions initiales. L'objectif des méthodes ensemblistes est d'exploiter les prédictions de différents modèles afin de réduire leur sensibilité et d'assurer la stabilité des prédictions faites. Pour sélectionner les modèles qui ont plus d'impact sur la décision de l'ensemble, dans cette contribution, nous avons choisi de voir l'effet des méthodes de sélection statique sur les résultats. La première méthode combine entre les N derniers modèles enregistrés dans différents points d'apprentissage, tandis que la deuxième méthode combine entre les N meilleurs modèles.

Contribution expérimentale 2 : Une nouvelle méthode de sélection dynamique des modèles d'apprentissage profond pour la classification du cancer colorectal

Notons que chaque contribution dans cette thèse exploite les conclusions ainsi que les avantages et les inconvénients de la contribution précédente. Dans la contribution précédente, nous avons constaté que la diversité et la bonne coopération entre les derniers et les meilleurs modèles n'est pas assurées, car, ces modèles peuvent avoir presque les mêmes erreurs et donc leur combinaison ne conduit pas toujours aux meilleurs résultats. La motivation de la deuxième contribution [Dif & Elberrichi 2020c] est de proposer une nouvelle méthode de sélection dynamique à base de la metaheuristique optimisation par essaim particulaire (PSO). L'objectif de la sélection dynamique est d'accorder plus d'attention à la qualité de l'ensemble sélectionné au lieu de considérer séparément la qualité de chaque modèle appartenant à l'ensemble. Cette contribution répond sur la première et la troisième question exposées dans la section précédente de la manière suivante :

Réponse à la question 1 : dans cette contribution, nous avons utilisé les CNN comme des modules d'extraction de caractéristiques. L'exploitation des CNN en extraction des caractéristiques est l'une des méthodes utilisées pour éviter les problèmes de sur-apprentissage liés au manque de données. Afin de choisir l'architecture appropriée dans la tâche traitée, nous avons commencé par une étude comparative entre la performance de 7 architectures de type CNN. Ensuite, nous avons sélectionné l'architecture la plus performante dans le reste des expérimentations.

Réponse à la question 3 : Nous avons choisi d'opter pour les méthodes ensemblistes pour combiner la décision de plusieurs modèles CNN. Pour sélectionner les modèles qui ont plus d'impact sur la décision de l'ensemble, nous avons proposé une nouvelle méthode de sélection dynamique basée sur la métaheuristique PSO pour sélectionner les membres appropriés dans le groupe. Cette sélection permet de générer un ensemble pertinent en fonction de la coopération du groupe plutôt que sur la force de chaque membre appartenant à cet ensemble.

Contribution expérimentale 3 : Une nouvelle stratégie de fine-tuning entre les bases d'apprentissage histopathologiques en apprentissage profond.

Dans les deux contributions précédentes, nous avons exploité les notions d'apprentissage ensembliste et d'apprentissage transféré pour résoudre le problème de manque de données histopathologiques. Malgré les avantages des méthodes ensemblistes, ces derniers sont caractérisées par un temps de traitement élevé et qui peut poser un grand problème dans le cadre des méthodes DL en raison de leur exigence en termes de capacité de calcul. Pour réduire ce temps, dans la première contribution, nous avons combiné entre des modèles générés à partir d'un seul apprentissage. Cette méthode permet d'effecteur un seul apprentissage au lieu de N apprentissage pour générer un ensemble composé de N modèles.

Tandis que dans la deuxième contribution, nous avons exploité la technique d'apprentissage transféré pour accélérer le temps de génération de l'ensemble.

L'inconvénient de ces deux approches est la capacité de stockage exigé pour stocker les modèles sélectionnés en inférence, et cela peut poser un problème pour les systèmes de vision intégrés. En plus, les méthodes ensemblistes exigent d'effectuer plusieurs prédictions et de combiner l'ensemble des prédictions afin d'obtenir la décision finale, et cela augmente le temps total d'inférence. La troisième approche proposée vise à éviter tous ces inconvénients à travers une stratégie simple et originale.

La troisième proposition [Dif & Elberrichi 2020d] exploite les notions d'apprentissage transféré entre les bases d'apprentissage histopatologiques et répond principalement à la quatrième question exposée dans la section précédente de la manière suivante :

Réponse à la question 4 : les méthodes de l'état de l'art basées sur l'apprentissage transféré proposent de transférer de la connaissance de la base d'apprentissage naturel ImageNet à des bases d'apprentissage histopathologiques. Malgré le succès de cette technique, il n'existe pas de principes théoriques sur le fonctionnement interne de cette stratégie et beaucoup de questions se posent sur la relation entre la base ImageNet et les bases d'apprentissage histopathologiques. Pour cela, nous avons proposé une méthode qui se base sur l'apprentissage transféré entre des bases d'apprentissage de même domaine au lieu de transférer la connaissance à partir de la base ImageNet.

Organisation de la thèse

Le contenu de la thèse est organisé en chapitre comme suit :

Les deux premiers chapitres présentent le background lié aux méthodes DL, notamment aux techniques exploitées dans cette thèse. Le troisième chapitre détaille l'état de l'art des domaines d'application des méthodes DL, et surtout en traitement des images médicales. Le quatrième chapitre présente le background lié au domaine d'application traité dans cette thèse. Le cinquième chapitre expose la contribution théorique ainsi que l'état de l'art des travaux liés à nos contributions. Enfin, le dernier chapitre et qui désigne le corps de cette thèse présente l'ensemble des contributions expérimentales.

Le chapitre 1 introduit les notions théoriques liées à l'apprentissage profond (DL). Il détaille l'origine et l'historique des réseaux de neurones ainsi que leur évolution. Ensuite, il présente l'architecture générale d'un réseau de neurones profond et les techniques exploitées dans l'apprentissage. Enfin, il illustre la configuration de quelques architectures connues en apprentissage supervisé et non supervisé.

Le chapitre 2 s'intéresse aux réseaux de neurones convolutifs (CNN). Premièrement, il présente les origines et le développement de ce réseau. Ensuite, il introduit l'architecture générale ainsi que les différentes techniques employées dans l'apprentissage de ces réseaux. Enfin, il explique et il compare les différentes architectures connues de type CNN en classification, en détection, et en segmentation.

Le chapitre 3 présente l'état de l'art lié à l'application des CNN en vision par ordinateur, comme : la classification des images, la détection et la localisation des objets, la segmentation sémantique, la reconnaissance d'action et d'activité, l'estimation de la pose humaine, et la reconnaissance faciale. Ensuite, il détaille l'état de l'art lié à la classification de différentes modularités d'images médicales.

Le chapitre 4 présente les différentes étapes effectuées pour l'acquisition des tissus et la numérisation des lames histologiques. Ensuite, il expose les méthodes d'augmentation utilisées dans le prétraitement des images histopathologiques. Enfin, il décrit la structure des bases d'apprentissage histopathologiques exploitées dans les contributions de cette thèse.

Le chapitre 5 présente les travaux liés à l'application des réseaux DL en histopathologie et l'optimisation de ces architectures par les méthodes ensemblistes et l'apprentissage transféré. Premièrement, il expose une synthèse sur les méthodes proposées pour la détection de la mitose [Dif & Elberrichi 2020a]. Cette synthèse est une référence importante pour les travaux futurs en détection de la mitose. Elle fournit une idée sur les techniques à éviter et elle présente quelques perspectives importantes pour résoudre les problèmes et les enjeux non traités en traitement des images histopathologiques et précisément en détection de la mitose. Ensuite, ce chapitre détaille les méthodes de régularisation liées aux méthodes ensemblistes et des techniques d'apprentissage transféré en apprentissage profond. L'objectif de cette partie est de discuter les approches précédemment proposées afin de justifier la contribution et les apports des approches proposées dans cette thèse.

Le chapitre 6 détaille les trois contributions expérimentales proposées dans cette thèse. Il s'intéresse à la résolution des différentes limitations liées à la classification des images histopathologiques. Dans ce cadre, nous avons proposé trois travaux de recherche [Dif & Elberrichi 2020d, Dif & Elberrichi 2020c, Dif & Elberrichi 2020b]. Les deux premières contributions exploitent les techniques de régularisation, d'apprentissage transféré, et des méthodes ensemblistes [Dif & Elberrichi 2020c, Dif & Elberrichi 2020b]. La dernière contribution [Dif & Elberrichi 2020d] est une nouvelle méthode de fine tuning entre les bases d'apprentissage histopathologiques.

Les deux derniers chapitres ont donné lieu à un ensemble de posters dans une journée doctorale et un atelier international, un chapitre dans un livre, et un ensemble d'articles dans une conférence internationale et des journaux internationaux.

Posters

— Dif, N., & Elberrichi, Z. (2019). A New Intra-domain fine tuning framework for histopathological images classification. Journées doctorales de la faculté des sciences exactes (JDFSE).

— Dif, N., & Elberrichi, Z. (2019). How Transferable are Features Between Histopathological Datasets . Les Ateliers SACONET.

Chapitre

— Dif, N., & Elberrichi, Z. (2020). Deep Learning Methods for Mitosis Detection in Breast Cancer Histopathological Images : A Compre-

hensive Review. In Artificial Intelligence and Machine Learning for Digital Pathology (pp. 279-306). Springer, Cham.

Conférence Internationale

— Dif, N., & Elberrichi, Z. (2018). IICBU 2008 Lymphoma Dataset : Overview and Deep Learning Application. Third International Conference on Multimedia Information Processing (CITIM), Mascara, Algeria.

Journaux Internationaux

— Dif, N., & Elberrichi, Z. (2020). A New Intra Fine-Tuning Method Between Histopathological Datasets in Deep Learning. International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 11(2), 16-40.

— Dif, N., & Elberrichi, Z. (2020). A New Deep Learning Model Selection Method for Colorectal Cancer Classification. International Journal of Swarm Intelligence Research (IJSIR), 11(3), 72-88.

— Dif, N., & Elberrichi, Z. (2020). Efficient Regularization Framework for Histopathological Image Classification Using Convolutional Neural Networks. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 14(4), 62-81.

L'apprentissage profond

1

Sommaire

L'apprentissage profond (DL) est une branche de l'apprentissage automatique qui est basée sur les réseaux de neurones. Les réseaux DL ont été largement exploités dans différents domaines grâce à leur capacité dans la résolution des problèmes complexes. Ces réseaux ont prouvé leur efficacité par rapport aux méthodes d'apprentissage classique dans plusieurs applications récentes. Malgré leurs avantages, ils ont plusieurs défis liés aux problèmes de sur-apprentissage, de dégradation de gradient, et de la complexité temporelle élevée. Afin de réduire ces problèmes, plusieurs architectures optimisées ont été proposées en apprentissage supervisé et non supervisé. Le but de ce chapitre est de présenter l'évolution des réseaux de neurones aux réseaux d'apprentissage profond, d'expliquer le processus d'apprentissage dans un réseau de neurones, de détailler les architectures connues de type DL, et de présenter les problèmes de chaque architecture avec les solutions proposées.

Mots clés : Apprentissage automatique, Perceptron, Perceptron multicouche, Apprentissage profond, Optimisation, Régularisation.

1.1 Introduction à l'apprentissage automatique

L'apprentissage automatique (ML) est une branche de l'intelligence artificielle (IA). Ce champ d'étude permet d'extraire de la connaissance exploitable et pertinente à partir des grands volumes de données. L'objectif est réalisé à l'aide des algorithmes de l'apprentissage automatique et une base d'apprentissage. Le processus d'apprentissage génère un modèle pour prédire les classes ou les groupes des nouvelles instances.

Le but principal de l'apprentissage automatique est d'automatiser une tâche de classification, régression, association, ou regroupement. Ces traitements sont catégorisés en apprentissage supervisé (classification et régression) et non supervisé (associassions et regroupement). Le traitement supervisé est caractérisé par une base d'apprentissage $D = \{x, y\}_{n=1}^{N}$, où x sont les instances et y l'ensemble des classes. Ces classes sont discrètes en cas de classification et continues en cas de régression. Il existe plusieurs types d'algorithmes d'apprentissage supervisé comme : les machines à vecteur de support (SVM), les plus proches voisins (KNN), les arbres de décision (DT) et la régression linéaire (LR). Contrairement à l'apprentissage supervisé, dans un apprentissage non supervisé, la base d'apprentissage $D = \{x\}_{n=1}^{N}$ est sans classes. Par exemple en regroupement, l'algorithme d'apprentissage regroupe l'ensemble des instances sous forme de catégories en se basant sur un ensemble défini de critères.

L'apprentissage profond est une branche de l'apprentissage automatique basée sur les réseaux de neurones artificiels (ANN). Ces réseaux ont été exploités en apprentissage supervisé [Ciregan *et al.* 2012] et non supervisé [Radford *et al.* 2015].

Ce chapitre détaille les différentes notions liées aux algorithmes DL du plus petit module (neurone ou perceptron) jusqu'au fonctionnement total d'un réseau multicouche.

1.2 Origine et inspiration

Les réseaux de neurones artificiels s'inspirent du fonctionnement des réseaux de neurones biologiques du cerveau. Le cerveau humain contient un nombre important de neurones biologiques massivement connectés. Ces neurones échangent les messages par des signaux transmis à travers les synapses. Ces systèmes biologiques sont caractérisés par leur grande complexité par rapport aux systèmes artificiels à cause du nombre élevé de neurones massivement interconnectés et qui ne peuvent pas être traités par une machine.

La figure 1.1 illustre la différence entre un neurone biologique et un neurone artificiel. Un neurone biologique est composé de plusieurs structures comme les dendrites, le noyau, le corps cellulaire et l'axone. Ces neurones échangent les messages par les signaux qui sont transmis entre les différents composants. Premièrement, les dendrites reçoivent les signaux de la part des synapses des neurones de leur voisinage. Ensuite, ces signaux sont regroupés et traités par le corps cellulaire. Ce dernier génère une impulsion qui est transmise à l'axone si le signal reçu dépasse un certain seuil. Enfin, le signal généré est transmis aux autres neurones

de voisinage à travers les synapses. Ce signal est modifié en fonction de la nature des synapses. Les synapses excitatrices augmentent ce signal, tandis que les inhibitrices réduisent sa valeur.

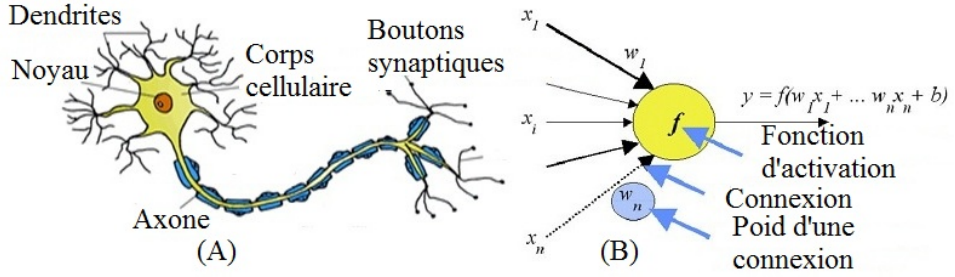

FIGURE 1.1 – *La différence entre (A) un neurone biologique et (B) un neurone artificiel.*

Un neurone artificiel imite un neurone biologique dans certaines fonctions. Ces systèmes neuronaux sont présentés à l'aide des notions mathématiques. Les signaux en entrée sont formulés par les valeurs $x_i, i \in [1, N], i \in \mathbb{N}$, les dendrites sont les poids w_i, le corps cellulaire c'est la fonction de combinaison, et l'axone c'est la valeur en sortie. Un neurone artificiel peut être composé d'un ensemble de nœuds ou de neurones qui sont associés à une valeur x_i et un poids w_i. Ces valeurs sont combinées par une fonction de combinaison et le résultat est traité par une fonction d'activation f.

1.3 HISTORIQUE

Les recherches sur les ANN ont commencé à partir des années 1940 quand les développements en neurobiologie ont encouragé les chercheurs à formuler le comportement des neurones biologiques.

En 1943, [McCulloch & Pitts 1943] ont développé un neurone artificiel pour la résolution des fonctions booléennes. L' indépendance de cette approche de l'apprentissage et son traitement manuel pour le calcule des poids w_i ont limité son utilisation dans d'autres domaines d'applications.

En 1957, [Rosenblatt 1957] a introduit la notion de perceptron pour la classification binaire. Contrairement à l'approche booleenne proposée précedement, ce modéle est basé sur l'apprentissage pour l'ajustement des poids w_i.

En 1968, [Minsky & Papert 2017] ont démontré les limites du perceptron dans la résolution d'une simple fonction booléenne XOR à cause de son adaptation au séparations linéaires. Tous ces facteurs ont décourager la communauté de l'IA à poursuivre la recherche dans ce domaine. Cette limitation a été résolu par l'introduction de la notion du perceptron multicouche (MLP) qui offre plus de non linéarité à travers les couches cachées supplémentaires. Ensuite, les MLP ont été développé en réseaux d'apprentissage profond (DNN) qui sont caractérisés par plus de deux couches cachées. Malgré la puissance des DNN dans les séparations non linéaire, les algorithmes d'apprentissage automatique comme SVM ont connu plus de sucées grâce à leur complexité temporelle optimisée par rapport aux DNN. Jusqu'à 2012 où l'architecture d'apprentissage profond AlexNet a marqué

un taux d'erreur impressionnant sur la base d'apprentissage ImageNet [Krizhevsky *et al.* 2012]. L'architecture proposée et la technologie des processeurs graphiques (GPU) ont encouragé les chercheurs d'exploiter les méthodes de l'apprentissage profond dans d'autres domaines d'application.

1.4 Perceptron

Le perceptron ou neurone formel [Rosenblatt 1957] est un algorithme d'apprentissage automatique qui appartient à la catégorie des classificateurs linéaires. Il permet de prédire les valeurs des poids w_{ij} d'un neurone artificiel. Le but du perception est de résoudre les problèmes linéairement séparables à deux classes. Pour une base d'apprentissage $D = \{X, y\}_{n=1}^{N}, y_n \in \{0, 1\}$ où $X_n = \{x_1, x_2, \ldots, x_m\}$ sont les instances et y_n sont les classes, il existe un hyperplan qui permet de séparer les différentes instances en deux classes (figure 1.2).

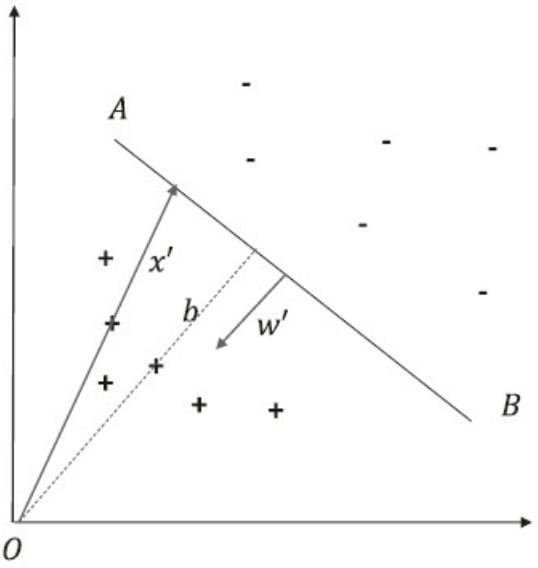

Figure 1.2 – *L'hyperplan AB séparant deux classes +,-.*

Le perceptron est caractérisé par n entrées et une seule sortie (figure 1.1 : B). Le neurone reçoit les valeurs en entrée qui sont associées à des poids w_i. Ensuite, les entrées sont combinées linéairement avec les poids à travers une fonction de combinaison et la somme pondérée en sortie est fournie à la fonction d'activation F (équation 1.1). L'équation 1.2 illustre le processus de calcul des classes, où $\widehat{y}_j$ est la classe prédite dans l'itération j et b est le biais qui présente une distance approximative de l'origine.

$$F(Z) = \begin{cases} 1 & si\ z > 0 \\ 0 & si\ non \end{cases} \tag{1.1}$$

$$\widehat{y}_j = F(b + \sum_{i=1}^{n} w_{ij} x_i) \tag{1.2}$$

Le perceptron prédit les poids $\{W_i\}_{i=0}^{m}$ de l'hyperplan $(W^T X + b) = 0$ qui permet de classifier correctement les différentes instances dans un processus itératif. Le critère d'arrêt dépend de la convergence des perfomances ou d'un nombre défini d'itérations.

Algorithme 1 : Le perceptron

Initialiser aléatoirement les poids w_i et le biais b;
while $i \leq iterarions$ **do**
  Prédire les classes y_n selon l'équation 1.2;
  Calculer les nouveau poids $W^{(t)}$ selon l'équation 1.3;

$$W^{(t)} = W^{(t-1)} + \alpha(y_n - \widehat{y_n})x_n \tag{1.3}$$

end

Le rôle du perceptron est de construire un seul hyperplan pour séparer entre deux classes, et donc il n'est pas capable de résoudre les problèmes non linéairement séparables comme la fonction logique XOR. Les problèmes non linéaires exigent plus de deux hyperplans pour réaliser une séparation correcte. Afin d'assurer cette séparation, les perceptrons multicouches ont été introduit. Cette non linéarité est assurée à travers les fonctions d'activations non linéaires au niveau des couches cachées.

1.5 Perceptron multicouche

1.5.1 L'architecture générale

Un perceptron multicouche est un réseau de neurones composé de plusieurs types de couches : d'entrée, cachée et de sortie.

La figure 1.3 illustre l'architecture d'un perceptron multicouche. Les nœuds dans la couche d'entrée reçoivent les valeurs à partir de la base d'apprentissage. Ensuite, ces valeurs sont propagées aux nœuds des autres couches cachées. Enfin, la dernière couche en sortie prédit les classes des instances en entrée.

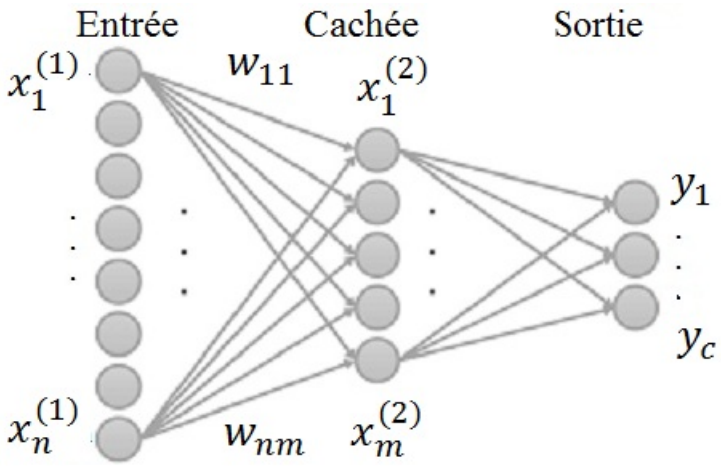

Figure 1.3 – *L'architecture d'un perceptron multicouche.*

Dans un MLP (figure 1.3) chaque nœud est caractérisé par une valeur x_i et contient des connexions aux couches adjacentes qui sont présentées par des poids w_{ij}. Chaque valeur x_i de la couche courante présente une entrée à la couche suivante. Le nombre de nœuds dans la couche d'entrée dépend du nombre d'attributs dans la base d'apprentissage. En revanche, la topologie appropriée aux couches cachées est choisie aléatoirement ou selon une procédure d'optimisation afin de maximiser les performances du problème traité. Les valeurs x_i des couches cachées sont calculées selon l'équation 1.4, où $x_i^{(k)}$ est la valeur du nœud i de la couche k, F est la fonction d'activation, w_{ij} sont les poids associés à x_i, et m est le nombre de nœuds dans la couche suivante.

$$x_i(k) = F^{(k)}(\sum_{j=1}^{m} w_{ij} x_j^{(k-1)} + b_j) \tag{1.4}$$

Un MLP est caractérisé par son adaptation aux problèmes non linéaires. L'apprentissage dans un MLP consiste à ajuster équitablement les poids w_{ij} en minimisant la valeur d'une fonction du coût C. Cette fonction mesure l'erreur entre les classes prédites et les classes réelles. La méthode d'apprentissage la plus connue en apprentissage supervisé à base d'un MLP est la rétropropagation. La partie suivante détaille les fonctions d'activation et les techniques d'apprentissage dans un MLP.

1.5.2 Les fonctions d'activation

Le but des fonctions d'activation est de convertir un signal en entrée d'un nœud à un signal en sortie. Cette conversion permet d'introduire des propriétés non linéaires au MLP afin qu'il puisse résoudre les problèmes non linéaires et traiter les données complexes, comme les images et les vidéos. Il existe deux types de fonctions d'activations : des fonctions linéaires (identité) et des fonctions non linéaires (sigmoïde logistique, softmax, tangente hyperbolique, et l'unité linéaire rectifiée). Généralement, les couches cachées utilisent la même fonction d'activation (tangente hyperbolique, sigmoïde ou l'unité linéaire rectifiée) et la couche en sortie est basée soit sur la fonction softmax, soit sur la fonction sigmoïde, en fonction du type de classification. Parmi les fonctions connues pour la résolution des problèmes non linéaires nous avons : la sigmoïde logistique, la tangente hyperbolique, l'unité linéaire rectifiée, et la Softmax.

L'identité

Cette fonction est caractérisée par sa linéarité (équation 1.5).

$$F(z) = z \tag{1.5}$$

La sigmoïde logistique

La Sigmoïde logistique est une fonction d'activation utilisée par les couches cachées ou de sortie (équation 1.6). Elle permet d'introduire plus de non linéarité aux couches cachées et de prédire les probabilités des

classes en sortie. Cette fonction est utilisée généralement dans les tâches de classification binaires.

$$F(z) = \frac{1}{(1+e^{-z})}, F(z) \in [0,1], z \in R \tag{1.6}$$

La tangente hyperbolique (tanh)

Cette fonction permet d'introduire de la non linéarité dans les couches cachées (équation équation 1.7). Elle est caractérisée par sa bonne précision en reconnaissance par rapport à la fonction sigmoïde logistique [Karlik & Olgac 2011].

$$F(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}, F(z) \in [-1,1], z \in R \tag{1.7}$$

L'unité linéaire rectifiée (ReLu)

L'unité linéaire rectifiée est une fonction d'activation non linéaire utilisée par les couches cachées. Selon l'équation 1.8, cette fonction neutralise les valeurs négatives à 0. ReLu a prouvé son efficacité par rapport aux fonctions sigmoïde et tanh grâce à sa simplicité. En plus, elle a permis d'accélérer le temps d'apprentissage des réseaux de neurones convolutif (CNN) [Krizhevsky *et al.* 2012]. Tous ces critéres ont fait de cette fonction la plus utilisée en apprentissage profond [Ramachandran *et al.* 2017].

$$F(z) = max(0,z), F(z) \geq 0, z \in R \tag{1.8}$$

Softmax

La fonction softmax est une fonction non linéaire utilisée par la couche en sortie. Elle permet de calculer l'ensemble des probabilités associées à chaque classe. Cette fonction est une généralisation de la fonction sigmoïde et qui est adaptée aux problèmes de classification à K classes, où K>2 (équation 1.9).

$$F(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, j \in \{1,...,K\}, F(z) \in [0,1], z \in R \tag{1.9}$$

1.5.3 L'apprentissage dans un perceptron multicouche

L'apprentissage dans un MLP est un problème d'optimisation qui consiste à ajuster les poids w_{ij} dans un processus itératif afin d'améliorer les performances du modèle. Le but principal est de minimiser la perte qui est présentée par une fonction du coût C. Cette fonction permet de mesurer la différence entre les classes prédites et les classes réelles. Les fonctions du coût les plus utilisées en apprentissage supervisé dans un MLP sont l'entropie croisée (CE) en classification (équation 1.10) et l'erreur moyenne quadratique (MSE) en régression (équation 1.11). y est le vecteur des classes réelles, $\widehat{y}$ est le vecteur des classes prédites, q est le vecteur des distributions réelles, p est le vecteur des distributions prédites,

et K est le nombre des classes. Les distributions présentent la probabilité d'appartenance de l'instance x à la classe i.

$$CE = \begin{cases} -q_1 log(p_1) - (1 - q_1)(1 - log(p_2)) & si\ k = 2 \\ -\sum_{i=1}^{K} q_i log(p_i) & si\ k > 2 \end{cases} \tag{1.10}$$

$$MSE = \frac{1}{K}\sum_{i=1}^{K}(y_i - \widehat{y}_i)^2 \tag{1.11}$$

L'apprentissage en MLP consiste à ajuster l'ensemble des paramètres θ (poids w_ij et biais b) afin de minimiser une fonction du coût C. Pour optimiser ces paramètres, la méthode de descente de gradient (GD) est exploitée. GD est la méthode d'optimisation la plus utilisée dans les réseaux DL. L'équation 1.12 illustre le processus d'ajustement des paramètres, où η est le taux d'apprentissage, $\theta^{(t)}$ sont les valeurs des paramètres dans l'itération t, et $-\nabla C(\theta^{(t)})$ est l'inverse du gradient de la fonction du coût. Cet inverse permet de définir la direction vers la valeur minimale de C. Le minimum est atteint lorsque $\nabla C(\theta) = 0$, ce qu'il engendre une convergence ($\theta^{(t+1)} = \theta^{(t)}$) et donc la fin du processus d'optimisation. Généralement, le critère d'arrêt dépend d'un nombre défini d'itérations, car la convergence au minimum global n'est pas assurée par les méthodes d'optimisations stochastiques. Le taux d'apprentissage η représente un paramètre très important dans la convergence, où les grandes valeurs de η peuvent causé un problème d'oscillation autour du minimum. Tandis que les petites valeurs peuvent engendrer une convergence très lente.

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla C(\theta^{(t)}) \tag{1.12}$$

En descente de gradient, le gradient de la fonction du coût est calculé par la méthode de rétropropagation. [Rumelhart *et al.* 1986] ont démontré la rapidité de cette méthode en apprentissage par rapport aux méthodes précédemment exploitées.

Durant l'apprentissage, chaque itération est composée de deux traitements principaux : la propagation vers l'avant et la rétropropagation. En premier, les valeurs $x_i^{(k)}$de chaque neurone sont calculées suivant l'équation 1.4. Ensuite, l'erreur ou la fonction du coût est mesurée en fonction des valeurs prédites et réelles. Cette erreur est rétropropagée aux poids des couches précédentes où le gradient de la fonction du coût est estimé dans un processus itératif. Cette procédure est suivie par la mis à jour des poids w_ij selon l'équation 1.12 en se basant sur les valeurs du gradient calculées précédemment.

1.5.4 Les techniques d'optimisation

L'ajustement des hyperparamètres est parmi les techniques importantes qui permettent d'améliorer les performances d'un MLP. L'optimiseur est un hyperparamètre qui a une grande influence sur la performance et la convergence d'un MLP.

Dans la partie précédente, nous avons détaillé le processus d'optimisation dans un MLP qui est basé principalement sur la méthode de descente

de gradient. Afin d'améliorer et d'accélérer l'apprentissage, différentes variantes de GD ont été proposées dans la littérature comme : la descente de gradient stochastique (SGD) et la descente de gradient à mini-lots (BGD). En plus, plusieurs méthodes ont été développées pour optimiser la GD comme : la descente avec inertie, descente de gradient accélérée de Nesterov (NAG), Adagrad, AdaDelta, Adam, et RmsProp.

Descente de gradient stochastique (SGD)

Dans la méthode GD, les gradients sont calculés en se basant sur toute la base d'apprentissage, cela limite son utilisation sur les grands volumes de données à cause de l'espace limité de la mémoire et le temps d'apprentissage très élevé. Afin de résoudre ces problèmes, la méthode de descente de gradient stochastique (SGD) est exploitée.

En SGD, le calcule des gradients et la mise à jour des poids est appliquée à chaque instance dans la base d'apprentissage. Ces instances sont choisies aléatoirement durant le processus d'optimisation. La sélection aléatoire et l'exploitation d'une seule instance permettent d'accélérer le temps d'apprentissage et d'améliorer la généralisation. Malgré ces avantages, les mises à jour fréquentes des poids w_ij à chaque itération peuvent causer une grande variance et une convergence instable. Afin de minimiser le nombre des mises à jour, la méthode de descente de gradient stochastique à mini-lot est proposée.

Descente de gradient à mini-lots (BGD)

Dans la méthode descente de gradient à mini-lots, la mise à jour des paramètres dépend d'un lot de données. Ce lot est composé d'un nombre défini d'instances qui sont choisies aléatoirement dans chaque itération. Ensuite, le processus de mise à jour des paramètres est effectué en fonction du lot de données sélectionnées. Les mini-lots permettent de réduire la variance entre les anciens et les nouveaux poids durant la mise à jour. Cela engendre une convergence plus stable par rapport à SGD. Actuellement, cette méthode est fréquemment utilisée dans les travaux d'apprentissage profond et référencée par SGD au lieu de BGD. Le processus d'optimisation dans la méthode SGD a plusieurs défis liés à :

— Le choix de la valeur optimale du taux d'apprentissage η et sa valeur fixe durant l'apprentissage. Les grandes valeurs de η peuvent causer une convergence rapide vers un minimum local. En revanche, les petites valeurs engendrent une convergence lente. En plus, la valeur de η doit varier et dépendre de la fréquence des attributs en entrée.

— Le risque de se faire piéger dans un minimum local à cause de la nature non convexe de la fonction du coût. Une fonction est non convexe (figure 1.4) si elle admet un minimum global ainsi que d'autres minimums locaux. Ce problème peut piéger la descente de gradient vers divers minimums locaux.

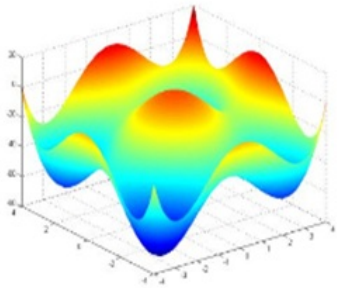

FIGURE 1.4 – *Fonction non convexe.*

La descente avec inertie

La descente avec inertie [Qian 1999] est une version optimisée de la méthode SGD. Cette technique permet d'accélérer la convergence et de réduire les différents problèmes liés à la nature non convexe de la fonction du coût. Cette méthode introduit la notion de vitesse qui dépend du paramètre d'inertie α. Cela permet de réduire les mises à jour des paramètres lorsque le gradient change de signe et de les accélérer si le gradient est dans la même direction de v. L'équation 1.13 illustre le processus de calcule des paramètres où α est le paramètre d'inertie et η est le taux d'apprentissage.

$$\begin{cases} v^{(t+1)} = \alpha v^{(t)} - \eta \nabla C(\theta^{(t)}) \\ \theta^{(t+1)} = \theta^{(t)} + v^{(t+1)} \end{cases} \tag{1.13}$$

Descente de gradient accélérée de Nesterov (NAG)

Descente de gradient accélérée de Nesterov [Nesterov 1983] est une version optimisée de la méthode de descente avec inertie. La technique d'inertie a des limites qui sont liées aux mises à jour importantes des poids w (figure 1.5 : 4a). Cela peut empêcher le processus d'optimisation à détecter le minimum global. Afin d'éviter ces sauts importants, NAG propose de calculer la vitesse en fonction du gradient de l'étape suivante au lieu de l'étape courante. Selon l'équation 1.14, NAG commence par une mise à jour partielle (4a) afin de calculer le paramètre intermédiaire $\theta^{(t+\frac{1}{2})}$. Ensuite, le paramètre final $\theta^{(t+1)}$ est ajusté en fonction du gradient du paramètre intermédiaire (4b). Le changement du signe du gradient de la fonction du coût entre les paramètres $\theta^{(t+\frac{1}{2})}$ et $\theta^{(t+1)}$ permet d'assurer des pas en arrière en réduisant la magnitude des mises à jour.

$$\begin{cases} \theta^{(t+\frac{1}{2})} = \theta^{(t)} + \alpha v^{(t)} \\ v^{(t+1)} = \alpha v^{(t)} - \eta \nabla C(\theta^{(t+\frac{1}{2})}) \\ \theta^{(t+1)} = \theta^{(t)} + \alpha v^{(t+1)} \end{cases} \tag{1.14}$$

Adagrad

En SGD, le taux d'apprentissage η est fixe et appliqué d'une façon équitable à tous les poids. En effet, les caractéristiques moins fréquentes

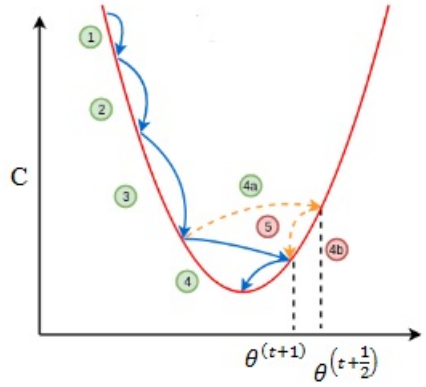

FIGURE 1.5 – *La descente de gradient accélérée de Nesterov.*

nécessitent des mises à jour plus importantes. Pour répondre à cette problématique, la méthode Adagrad [Duchi *et al.* 2011] propose d'adapter le taux d'apprentissage η à chaque paramètre. Cela rend cette technique plus convenable aux bases d'apprentissage creuses en DL. L'équation 1.15 illustre la procédure de calcul des paramètres $\eta^{(t+1)}$.

$$\begin{cases} \theta^{(t+1)} = \theta^{(t)} - \eta G_{(t)}^{-1} \nabla C(\theta^{(t)}) \\ G_{(t)} = \begin{bmatrix} \sqrt{\sum_{\tau=1}^{t} \theta_1^{(\tau)^2} + \epsilon} & \dots & \dots \\ & \dots \sqrt{\sum_{\tau=1}^{t} \theta_i^{(\tau)^2} + \epsilon} & \dots \\ \dots & \dots & \sqrt{\sum_{\tau=1}^{t} \theta_n^{(\tau)^2} + \epsilon} \end{bmatrix} \end{cases} \tag{1.15}$$

AdaDelta

Dans la méthode Adagrad, le taux d'apprentissage $\eta G_{(t)}^{(-1)}$ diminue d'une manière continue à cause de la somme accumulée $\sqrt{\sum_{\tau=1}^{t} \theta_1^{(\tau)^2} + \epsilon}$ des itérations précédentes. Cela engendre une convergence très lente en raison des petites valeurs du taux d'apprentissage. La méthode AdaDelta [Zeiler 2012] est une version optimisée de Adagrad et qui résout le problème de dégradation du taux d'apprentissage en considérant la moyenne mobile exponentiellement des gradients carrés. L'équation 1.16 illustre la procédure de calcul des paramètres $\theta^{(t+1)}$, où γ est la constante de décroissement exponentielle, η le taux d'apprentissage et $g_{ij}^{(t)}$ est la racine carrée moyenne des gradients.

$$\begin{cases} g_{ij}^{(t)} = \gamma g_{ij}^{(t-1)} + (1-\gamma)(\nabla C(\theta^{(t)}))^2 \\ \theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{\sqrt{g_{ij}^{(t)} + \epsilon}} \nabla C(\theta^{(t)}) \end{cases} \tag{1.16}$$

Adam

Adam [Kingma & Ba 2015] appartient à la catégorie des méthodes qui proposent un taux d'apprentissage variant comme Adagrad et AdaDelta. Dans la mise à jour du taux d'apprentissage, cette méthode utilise la

moyenne mobile exponentiellement des gradients carrés passés v_t et des gradients passés m_t. Les variables v_t et m_t présentent le premier et le deuxième moment du gradient (équation 1.17) où β_1 et β_2 sont les taux de décroissance. Pour éviter la convergence des moments v_t et m_t vers 0, les moments corrigés du biais $\widehat{m_{ij}^{(t)}}$ et $\widehat{v_{ij}^{(t)}}$ sont utilisés (équation 1.18). Enfin, les poids $w_{ij}^{(t+1)}$sont mis à jour selon l'équation 1.19.

$$\begin{cases} m_{ij}^{(t)} = \beta_1 m_{ij}^{(t-1)} + (1-\beta_1)\frac{\partial C^{(t)}}{\partial w_{ij}} \\ v_{ij}^{(t)} = \beta_2 v_{ij}^{(t-1)} + (1-\beta_2)(\frac{\partial C^{(t)}}{\partial w_{ij}})^2 \end{cases} \tag{1.17}$$

$$\begin{cases} \widehat{m_{ij}^{(t)}} = \frac{m_{ij}^{(t)}}{(1-\beta_1^t)} \\ \widehat{v_{ij}^{(t)}} = \frac{v_{ij}^{(t)}}{(1-\beta_2^t)} \end{cases} \tag{1.18}$$

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \frac{\eta}{\sqrt{\widehat{v_{ij}^{(t)}} + \epsilon}} \widehat{m_{ij}^{(t)}} \tag{1.19}$$

RmsProp

RmsProp [Tieleman & Hinton 2012] est une variante de l'optimiseur Rprop [Riedmiller & Braun 1993] qui est adaptée à l'apprentissage à mini-lots. Cette méthode est considérée comme une combinaison des méthodes Rprop et SGD et connue par sa similarité à l'optimiseur AdaDelta. Le but principal de cette stratégie est de résoudre le problème de dégradation du taux d'apprentissage. RmsProp divise le taux d'apprentissage par la moyenne mobile exponentiellement des gradients carrés et fixe la constante de décroissement γ de AdaDelta à 0.9. L'équation 1.20 illustre le processus de mise à jour des poids $\theta^{(t+1)}$, ou η est le taux d'apprentissage et $g_{ij}^{(t)}$ est la racine carrée moyenne des gradients.

$$\begin{cases} g_{ij}^{(t)} = 0.9 g_{ij}^{(t-1)} + 0.1(\nabla C(\theta^{(t)}))^2 \\ \theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{\sqrt{g_{ij}^{(t)} + \epsilon}} \nabla C(\theta^{(t)}) \end{cases} \tag{1.20}$$

1.5.5 Les techniques de régularisation

En apprentissage automatique, un modèle efficace est caractérisé par une bonne performance sur les données d'apprentissage et de test. Tandis que la conception d'un bon modèle en généralisation est l'un des grands défis en apprentissage automatique à cause des problèmes de sous-apprentissage et sur-apprentissage. Le sous-apprentissage présente les catégories des modèles qui souffrent d'une faible performance sur les données d'apprentissage et test. En revanche, le sur-apprentissage est défini par les modèles performants sur les données d'apprentissage et mal adaptés aux données tests, et cela engendre une grande variance entre les deux et une mauvaise généralisation.

En apprentissage profond, les réseaux sont caractérisés par une variance élevée et des biais faibles. La figure 1.6 illustre le problème de

sur-apprentissage qui est lié principalement à la complexité du modèle, où les modèles complexes comme les DL ont plus de risque de sur-apprentissage. Afin de réduire ce risque, il existe deux stratégies : la stabilisation structurelle et la régularisation [Bishop *et al.* 1995]. La stabilisation structurelle contrôle la complexité du réseau en fonction de la variation du nombre des paramètres. La méthode de régularisation consiste à ajouter un terme qui pénalise la fonction du coût. Ce terme est utilisé pour pénaliser la fonction du coût en cas de paramètres à grande magnitude. Il existe plusieurs méthodes de régularisations : les régularisations L1 et L2, la régularisation par abandon (Dropout), l'augmentation des données, et l'arrêt prématuré.

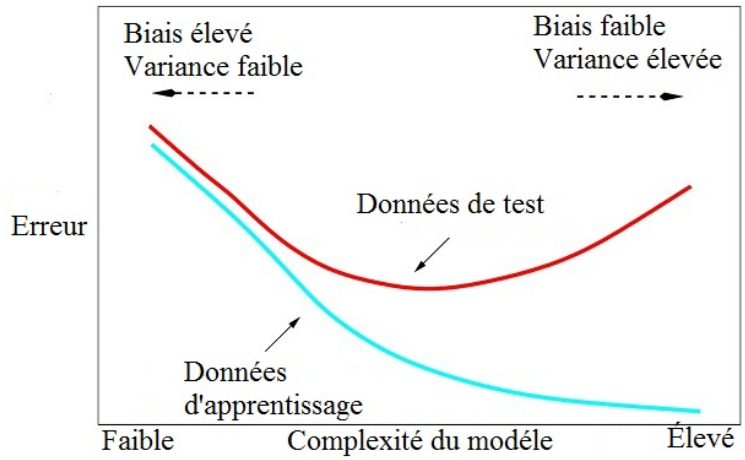

FIGURE 1.6 – *Modélisation du problème de sur-apprentissage.*

Les régularisations L1 et L2

Les régularisations L1 et L2 sont parmi les techniques connues en régularisation. Ces méthodes mettent à jour la fonction de coût en ajoutant un terme de régularisation. Leur but est de diminuer les valeurs des poids par l'ajout du terme de régularisation afin de générer des modèles simples et d'éviter les problèmes de sur-apprentissage .Les équations 1.21 et 1.22 présentent les termes de régularisation L1 et L2 respectivement, où λ est le paramètre de régularisation, $\| \theta \|$ et $\| \theta \|^2$ sont les normes l1 et l2 du vecteur des paramètres θ. La technique de régularisation L2 est aussi connue par la méthode de dégradation des poids (weight decay).

$$L_1 = \lambda \sum \| \theta \| \tag{1.21}$$

$$L_2 = \lambda \sum \| \theta \|^2 \tag{1.22}$$

La régularisation par abandon (Dropout)

La régularisation par abandon [Srivastava *et al.* 2014] est une méthode de régularisation qui consiste à négliger l'étape de l'apprentissage au niveau d'un sous ensemble de neurones. Ces neurones sont sélectionnés

aléatoirement selon un taux d'abandon et appartiennent aux couches d'entrée ou cachées. Cette technique permet de réduire le nombre important de liens entre les neurones et la spécialisation de certains au détriment d'autres, ce qui améliorera la généralisation et évitera le surapprentissage. Le processus d'abandon génère différentes architectures dans chaque itération et la performance totale est définie par la moyenne des performances de ces architectures. Cela permet de réduire la variance entre ces modèles et d'améliorer la généralisation.

L'augmentation des données

Contrairement aux méthodes d'apprentissage automatique classique, les méthodes d'apprentissage profond exigent un grand volume de données afin d'éviter le problème de sur-apprentissage. La collecte d'une grande quantité de données et leur annotation présentent un défi dans certains domaines comme le domaine médical. Afin de résoudre ces limites, les méthodes d'augmentation de données sont proposées. Cette augmentation peut être effectuée soit avant l'étape de l'apprentissage (hors ligne) ou durant l'apprentissage sur les mini-lots (en ligne). Il existe plusieurs méthodes d'augmentation de données. Parmi les techniques les plus utilisées, nous avons : la rotation, la translation, l'écaillage, le bruit gaussien, division en patch (figure 1.7), la normalisation et l'amélioration des couleurs, et la génération de nouvelles instances par les réseaux contradictoires génératifs (GAN).

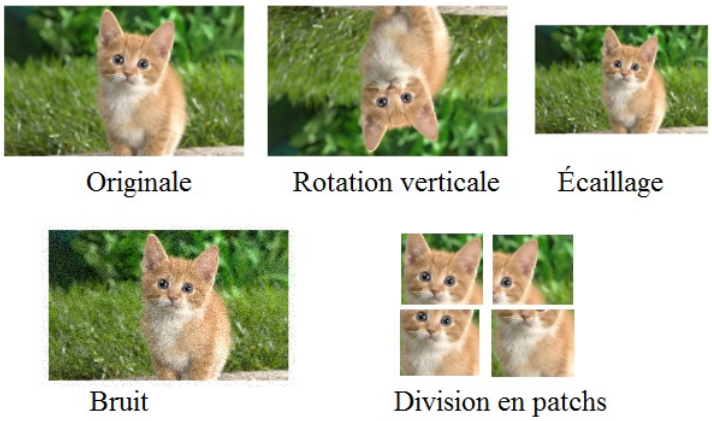

FIGURE 1.7 – *Les méthodes d'augmentation des données.*

L'arrêt prématuré

L'arrêt prématuré est une méthode de régularisation implicite [Zhang *et al.* 2016]. Dans cette méthode, le jeu de données est divisé en base d'apprentissage et de validation. Ensuite, la performance du modèle est évaluée durant l'apprentissage sur les deux bases en se basant sur un indicateur (l'erreur). Généralement, l'apprentissage est arrêté lorsque la valeur de l'erreur sur la base de validation commence à s'incrémenter (figure 1.6) à cause du problème de sur-apprentissage [Bishop 2006]. Enfin, le modèle qui minimise le taux d'erreur est stocké pour la phase de prédiction.

1.6 Les architectures d'apprentissage profond

L'apprentissage profond est une branche de l'apprentissage automatique qui est basée principalement sur les réseaux de neurones. Un MLP composé de plus de deux couches cachées est considéré comme un type de réseaux DL. Ces couches sont exploitées pour l'extraction et la transformation des caractéristiques.

Les caractéristiques de plus haut niveau sont construites par la combinaison des caractéristiques de plus bas niveau impliquant un apprentissage à multi-représentation. Par exemple, en vision par ordinateur, les neurones de la première couche représentent des caractéristiques simples comme les bornes. Ensuite, ces caractéristiques deviennent de plus en plus complexes dans les couches profondes. Ce processus donne avantage aux réseaux plus profonds à résoudre les problèmes complexes.

Selon la figure 1.8, les couches dans un réseau DL classique sont fortement connectées par des liens (poids w_{ij}). En revanche, le nombre élevé des paramètres peut conduire rapidement à des problèmes de surapprentissage et une complexité temporelle très élevée. Afin de réduire ces coûts et d'adapter les réseaux DL à des domaines d'applications spécifiques, plusieurs types d'architectures optimisées ont été proposés en apprentissage supervisé (CNN, les réseaux de neurones récurrents (RNN), mémoire à long-court terme (LSTM)) et non supervisé (les auto-encodeurs empilés (SAE), réseaux de croyances profondes (DBN), GAN). Le chapitre suivant présente en détail la structure du réseau CNN.

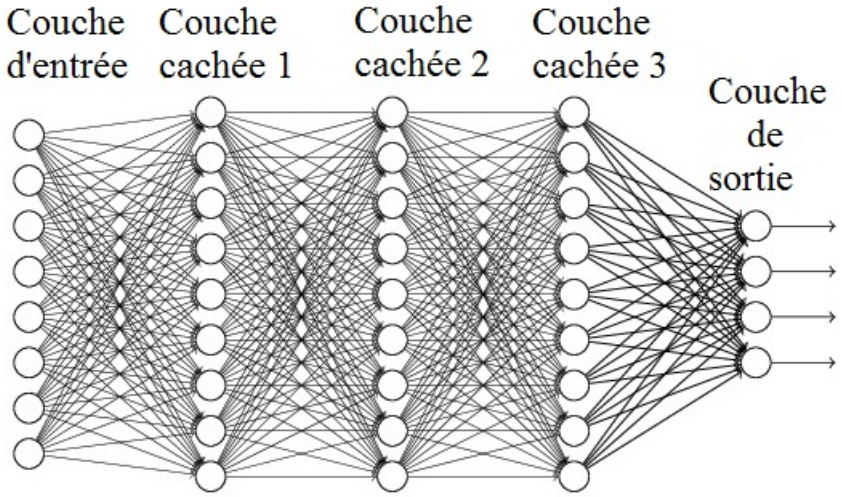

Figure 1.8 – *L'architecture d'un Fully connected NN.*

1.6.1 Les réseaux d'apprentissage supervisé

Les réseaux de neurones récurrents (RNN)

Les réseaux de neurones récurents [Pineda 1987] sont des réseaux d'apprentissage profond conçus pour l'apprentissage à partir des données séquentielles. Ils représentent un moyen pour le partage des poids au fil du temps.

Contrairement à MLP, dans un RNN, les activations dans les couches cachées dépendent des entrées actuelles et antérieures. Toute connexion permet de prendre en compte à l'étape courante une ou plusieurs informations prédites dans une étape précédente, où les mêmes éléments sont traités différemment selon la situation. De cette façon, les résultats de l'étape $t-1$ affectent les décisions de l'étape suivante t. Ces caractéristiques ont fait des RNN une bonne architecture dans les tâches du traitement du texte, car un mot peut avoir plusieurs sens en fonction de son positionnement dans une phrase.

La figure 1.9 illustre la structure basique d'un RNN, ce réseau est caractérisé par des cycles entre les différentes connexions pour traiter les séquences de tailles variables. Le nombre de couches présente le nombre de mots en entrée x_i. Ces neurones partagent les mêmes poids w_{ij} afin de réduire le nombre total des paramètres. L'équation 1.23 illustre le processus de calcul des sorties o_j, où h_t est la mémoire, w_{ij} sont les poids, et b_0 est le biais.

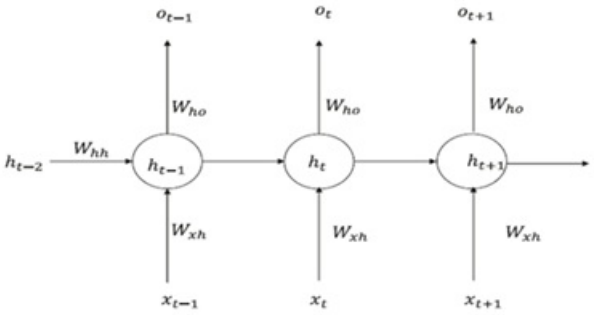

FIGURE 1.9 – *La structure d'un réseau de neurones récurrents.*

$$\begin{cases} h_t = (w_{hh}h_{t-1} + w_{xh}x_t) \\ o_t = w_{h0}h_t + b_0 \end{cases} \tag{1.23}$$

Le réseau de neurones récurent et ses variantes ont été largement exploités dans diverses applications en traitement des textes [Tang *et al.* 2015] et en vision par ordinateur : reconnaissance d'action [Du *et al.* 2015], génération de légendes [Vinyals *et al.* 2015], et segmentation des images [Xie *et al.* 2016].

Malgré les avantages des RNN en traitement des séquences temporelles, ces réseaux risquent le problème de dégradation du gradient à cause du nombre important des couches cachées. Ce nombre est lié au nombre des séquences en entrée qui varie selon la tâche traitée. Au fil des itérations, les gradients des séquences distantes ont tendance de converger vers 0, ce qui empêche le modèle à reconnaitre les associations entre ces séquences. Afin de résoudre ces limitations, le réseau mémoire à long-court terme (LSTM) [Hochreiter & Schmidhuber 1997] a été développé.

Mémoire à long-court terme (LSTM)

Le réseau mémoire à long-court terme [Hochreiter & Schmidhuber 1997] est une version optimisée des RNN. Contrairement à un RNN, ce réseau

permet de mémoriser les relations entre les séquences distantes. Un LSTM est composé d'un ensemble d'unités interconnectées de type LSTM. La figure 1.10 illustre la structure d'une unité LSTM, où C_t est l'état de cellule. Cet état est contrôlé par 3 portes : porte d'entrée i_t, porte d'oubli f_t et de sortie o_t.

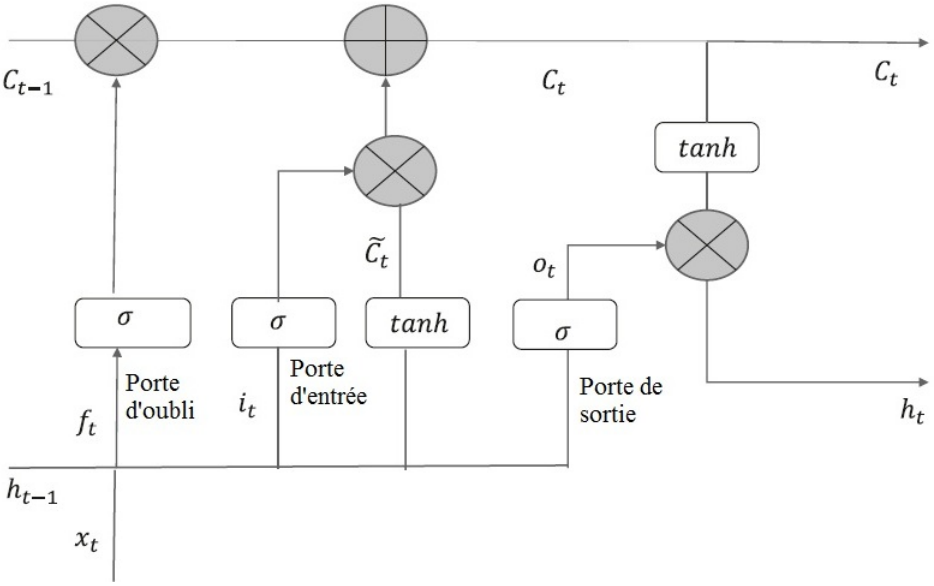

FIGURE 1.10 – *La structure d'une unité mémoire à long-court terme.*

Premièrement, la porte d'oubli (équation 1.24) contrôle les informations à oublier de l'état de cellule précédent $C_{(t-1)}$, où la valeur 0 indique l'oubli de l'état, tandis que la valeur 1 garde son historique. De la même façon, selon l'équation 1.25, la porte d'entrée décide la mis à jour des états de cellules précédentes. Ensuite l'état de mémoire C_t et la porte en sortie sont mises à jour selon les équations 1.26 et 1.27 respectivement, et enfin l'état caché h_t est calculé en fonction de la porte et l'état en sortie (équation 1.28). σ est la fonction d'activation sigmoïde. Les portes d'oubli et de sortie présentent deux paramètres importants qui permettent de garder seulement les informations pertinentes.

$$f_t = \sigma(W_f x_t + U_f h_{(t-1)}) \tag{1.24}$$

$$i_t = \sigma(W_f x_t + U_{if} h_{(t-1)}) \tag{1.25}$$

$$\begin{cases} \widetilde{C}_t = tanh(W_c x_t + U_c h_{(t-1)}) \\ C_t = f_t C_{t-1} + i_t \widetilde{C}_t \end{cases} \tag{1.26}$$

$$o_t = \sigma(W_0 x_t + U_0 h_{t-1}) \tag{1.27}$$

$$h_t = o_t \times tanh(C_t) \tag{1.28}$$

LSTM a été utilisé dans plusieurs domaines d'applications : amélioration des discours [Weninger *et al.* 2015], la détection d'activité vocale dans la vie réelle [Eyben *et al.* 2013] et le sous titrage automatique des images [Yao *et al.* 2017].

1.6.2 Les réseaux d'apprentissage non supervisé

les auto-encodeurs empilés (SAE)

Un auto-encodeur empilé est un réseau de neurones composé d'une succession d'auto-encodeurs (figure 1.11). Un auto-encodeur est composé de deux parties : encodeur et décodeur. L'encodeur est utilisé pour transformer les données en entrée x à une représentation compressée. Ensuite, le décodeur utilise cette représentation pour reconstruire les données en entrées $\widehat{x}$.

Le processus d'empilement des auto-encodeurs permet à SAE d'apprendre des structures plus complexes. L'apprentissage est divisé en deux parties : apprentissage non supervisé et apprentissage supervisé basé sur l'apprentissage transféré. Contrairement à un MLP, dans chaque itération, l'apprentissage dans un SAE est réalisé seulement au niveau d'une unité d'auto-encodeur, où chaque unité prend en entrée les sorties de l'unité précédente. Cela permet à chaque auto-encodeur de minimiser l'erreur de la couche précédente. Enfin, cette procédure est suivie par un apprentissage transféré dans les couches cachées.

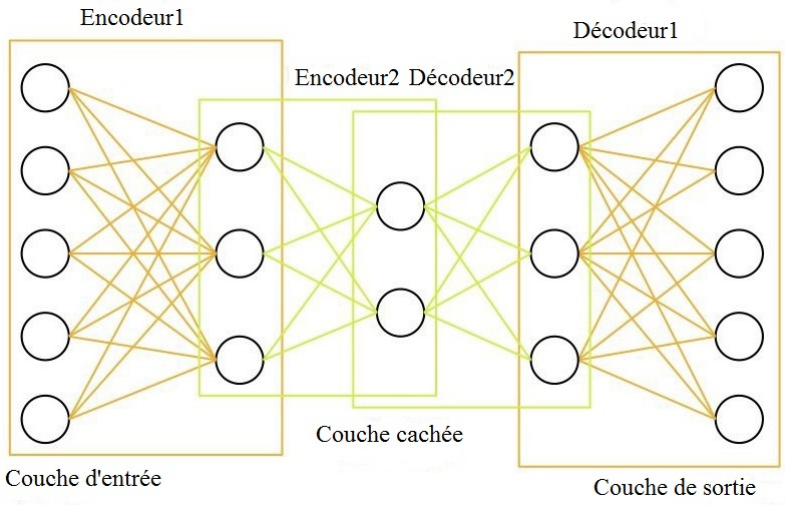

FIGURE 1.11 – *La structure d'un auto-encodeur empilé.*

Les SAE ont été exploité dans différentes applications comme : la compression des données [Tan & Eswaran 2011], la réduction de dimensionnalité [Zabalza *et al.* 2016], la réduction de bruit [Ishii *et al.* 2013] et la détection des anomalies [An & Cho 2015].

Les réseaux de croyances profondes (DBN)

Le réseau de croyance profond [Hinton *et al.* 2006] appartient à la catégorie des réseaux de neurones génératifs. Selon la figure 1.12, un DBN est composé d'un ensemble de réseaux empilés de type machine Boltzmann restreinte (RBM). L'apprentissage en DBN est effectué d'une manière successive dans chaque module RBM, où chaque module prend en entrée les sorties du module précédent. L'apprentissage est basé sur la méthode de divergence constructive.

DBN a été largement exploité dans plusieurs domaines d'applications comme la reconnaissance des images [Wu & Chen 2015], la prédiction et la classification du mouvement humain [Butepage *et al.* 2017], et la modélisation des séquences de vidéos [Gan *et al.* 2015].

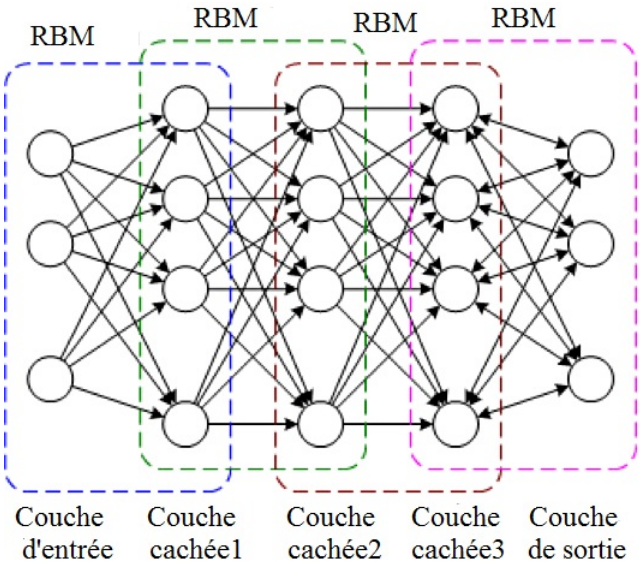

FIGURE 1.12 – *La structure d'un réseau de croyance profond.*

Les réseaux contradictoires génératifs (GAN)

GAN [Goodfellow *et al.* 2014] est un modèle génératif basé sur les réseaux neurones. La modélisation générative est une tâche d'apprentissage non supervisé en apprentissage automatique. Cette tâche permet de générer de nouvelles instances à partir des instances existantes. GAN est composé de deux modèles : générateur G et discriminateur D. Le modèle générateur permet de générer de nouvelles instances. Ensuite, ces instances sont classifiées en deux classes : réel ou truqué.

Les modèles générateur et discriminateur sont des réseaux de neurones de type DL et qui varient selon le domaine d'application, par exemple les réseaux de neurones convolutifs (CNN) sont exploités en traitement des images. Les réseaux contradictoires génératifs peuvent êtres utilisé dans différents domaines et applications comme : l'augmentation des données [Antoniou *et al.* 2018], la segmentation sémantique [Luc *et al.* 2016], la traduction image-à-image [Isola *et al.* 2017], et la génération de musique [Dong *et al.* 2018].

1.7 CONCLUSION

Dans ce chapitre, nous avons présenté les différents principes liés à l'apprentissage profond. Les réseaux d'apprentissage profond sont des algorithmes d'apprentissage automatique basés sur les réseaux de neurones.

Le perceptron a été le premier algorithme proposé pour l'apprentissage d'un neurone artificiel. Cet algorithme a rapidement montré ses limites dans la résolution des problèmes non linéaires. Afin de résoudre

cette limitation, le perceptron multicouche a été proposé. Ce dernier permet d'introduire la non-linéarité à travers les fonctions d'activation non linéaires dans les couches cachées. L'apprentissage dans un MLP est réalisé à l'aide de la méthode de rétroprobagation basée sur la technique d'optimisation SGD ou ses variantes optimisées. Les neurones dans un MLP profond sont fortement connectés, ce qu'il peut conduire à un problème de sur-apprentissage sur les volumes de données limités. Pour répondre à ce problème, d'autres variants de réseaux DL ont été proposés en apprentissage supervisé (RNN, LSTM, CNN) et non-supervisé (SAE, DBN, GAN), où chaque réseau est spécialisé dans des domaines d'application précis et répond aux limitations des autres réseaux.

Le chapitre suivant présente les principes liés à CNN qui est le sujet d'intérêt de cette thèse. Dans ce cadre, nous commencerons par la présentation de l'architecture générale d'un CNN. Ensuite, nous détaillerons les différentes architectures communes de type CNN en apprentissage supervisé.

Les réseaux de neurones convolutif

2

Sommaire

Un réseau de neurones convolutif (CNN) est un algorithme d'apprentissage profond. Ce réseau a été largement exploité en vision par ordinateur pour la classification et la détection des objets grâce à ses fonctionnalités inspirées du cortex visuel. Contrairement aux réseaux DL classiques, les CNN sont caractérisées par des couches de convolution et de pooling. Ces couches introduisent des liens partiels pour réduire le nombre des paramètres et renforcer le partage des caractéristiques communes. Malgré ces avantages, ils ont plusieurs défis liés aux problèmes de sur-apprentissage sur les volumes limités de données et une complexité de calcul élevée. Pour résoudre ces problèmes, des architectures connues de type CNN ont été proposées. Ces architectures sont basées sur des blocs convolutionnels optimisés qui permettent de générer des structures plus profondes et moins exigeantes en termes de capacité de calcule et de stockage. Le but de ce chapitre est de détailler la structure générale d'un CNN et de comparer entre les architectures communes de type CNN en classification, en détection des objets, et en segmentation.

Mots clés : Réseau de neurones convolutif, Classification, Détection des objets, Segmentation, Convolution, Pooling, ImageNet, PASCAL VOC.

2.1 Introduction

Les réseaux de neurones convolutif sont des réseaux d'apprentissage profond inspirés du cortex visuel [Hubel & Wiesel 1962]. Ces réseaux ont été utilisés dans les systèmes de recommandation [Ying *et al.* 2018], en traitement du langage naturel [Kim 2014], et en vision par ordinateur [Krizhevsky *et al.* 2012]. Leur exploitation en vision par ordinateur a connu un grand succès grâce à leurs caractéristiques inspirées des systèmes visuels naturels.

En vision par ordinateur, le processus de classification par les méthodes d'apprentissage classiques est basé sur 2 étapes principales : l'extraction des caractéristiques et l'apprentissage. Ces caractéristiques sont considérées comme des handcrafted features à cause de l'effort manuel et nécessaire dans l'étude des attributs discriminants. Les méthodes utilisées extraient ces caractéristiques d'une manière non supervisée. Cette séparation entre les modules d'extraction et de classification peut nuire la tâche de classification si certains attributs discriminants ont été négligés dans la phase de l'extraction.

Contrairement aux méthodes d'apprentissage classiques, les CNN réalisent implicitement le processus d'extraction des caractéristiques à travers les couches de convolution, où les premières couches représentent les caractéristiques simples. Ensuite, ces caractéristiques sont combinées pour former d'autres qui sont plus complexes dans les couches profondes. Cette spécificité a rendu les CNN un bon outils pour la classification des données non structurées comme les images et les textes. Malgré ces avantages, les CNN profonds risquent le problème de sur-apprentissage sur les volumes limités de données, car ils sont plus adaptés aux grands volumes à cause du problème de sur-apprentissage [Keshari *et al.* 2018].

Les CNN ont prouvé leur efficacité par rapport aux réseaux de neurones profonds classiques en termes de complexité temporelle et spatiale. Ces réseaux sont caractérisés par leur stratégie de partage des paramètres. Contrairement à un DNN, où les couches adjacentes sont fortement connectées, dans un CNN, chaque neurone de la couche courante est connecté seulement à un sous ensemble de neurones de la couche précédente (figure 2.1). Les CNN sont considérés comme des méthodes de stabilisation structurelle qui permettent de réduire les problèmes de sur-apprentissage à travers l'optimisation du nombre des paramètres.

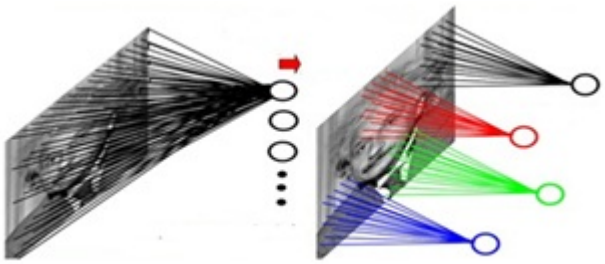

Figure 2.1 – *La différence entre le nombre de connexions dans une couche fortement connectée et une couche de convolution.*

Ce chapitre explique l'architecture générale d'un CNN et compare

entre les architectures connues de type CNN exploitées en classification, en détection des objets, et en segmentation.

2.2 Historique

Les réseaux de neurones convolutifs sont inspirés du cortex visuel. Le cortex visuel est la partie de cerveau responsable du traitement des informations provenant de l'œil.

En 1962, les chercheurs [Hubel & Wiesel 1962] ont inséré des électrodes dans des parties spécifiques du cortex visuel d'un chat afin de mesurer l'activation lorsque le chat observe quelques formes de base. Ils ont remarqué que les cellules simples répondent seulement aux barres horizontales en bas d'une image. Tandis que les cellules complexes sont caractérisées par une invariance spatiale, où ils peuvent répondre à ces barres dans différents emplacements dans l'image. Cette invariance est assurée par la combinaison des sorties des cellules simples.

En fonction de ces hypothèses, [Fukushima 1980] a développé un modèle (figure 2.2) composé de deux types de cellules neuronales : les cellules simples (S) et complexes (C). Les cellules S sont activées à la détection des formes basiques, tandis que les cellules C combinent les activations des cellules S. L'idée est de transformer les concepts biologiques introduits précédemment à des concepts mathématiques pour modéliser la tâche de reconnaissance visuelle des formes. Ce modèle a été exploité pour la reconnaissance des formes dans une approche non supervisée.

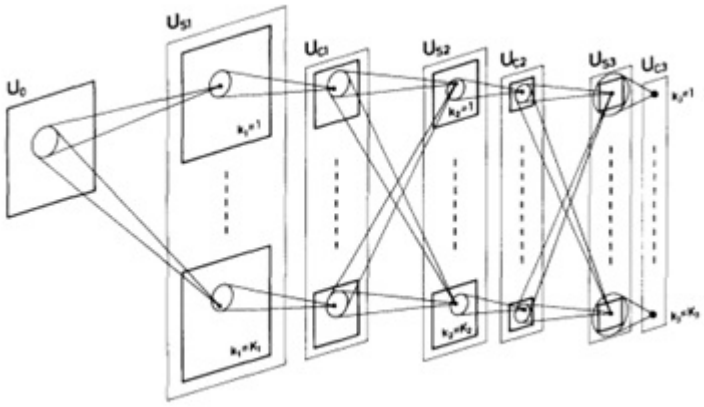

Figure 2.2 – *La structure du modèle proposé par [Fukushima 1980].*

En 1998, [LeCun *et al.* 1998] ont introduit le réseau de neurones convolutif qui est basé sur l'architecture proposée par Fukushima, où ils ont exploité la methode de rétropropagation afin d'accomplir une tâche de classification supervisée. Leur modèle a été testé sur la base d'apprentissage MNIST spécialisée en classification des caractères manuscrits.

Au début des années 2000, la recherche sur les réseaux CNN a stagné à cause de la puissance insuffisante des processeurs et des capacités des mémoires internes limités pour les besoins de tels algorithmes. Durant cette période, les algorithmes d'apprentissage automatique classiques ont

été largement exploités grâce à leurs exigences raisonnables en termes de complexité de calcul et espace de stockage.

En 2012, l'architecture AlexNet de type CNN a réussi le meilleur taux d'erreur dans l'état de l'art sur la base d'apprentissage ImageNet [Krizhevsky *et al.* 2012]. Cette bonne performance et la capacité des GPUs dans l'optimisation de la complexité temporelle ont encouragé la communauté de l'intelligence artificielle à proposer d'autres variantes optimisées de l'architecture CNN.

2.3 L'architecture générale d'un réseau de neurones convolutif

Un réseau de neurone convolutif est composé de trois types de couches : couche de convolution, couche de pooling et couche entièrement connectée. La figure 2.3 illustre l'architecture d'un CNN.

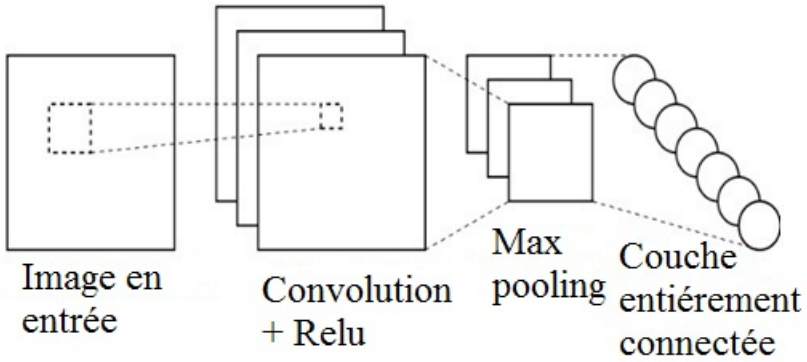

Figure 2.3 – *l'architecture générale d'un réseau de neurones convolutif.*

2.3.1 Couche de convolution

La couche de convolution est définie par le bloc de construction principal d'un CNN. Le but de cette couche est d'extraire implicitement les caractéristiques pertinentes des images en entrée durant l'apprentissage. Cette couche effectue une opération de convolution entre deux matrices, la première représente une sous partie des données en entrée (champ réceptif) et la deuxième représente un filtre qui contient les paramètres d'apprentissage. Une opération de convolution génère une troisième matrice référencée par la carte des caractéristiques. La figure 2.4 illustre une opération d'une convolution qui est réalisée par un produit scalaire entre le filtre et un champ réceptif. Ensuite, les résultats du produit sont additionnés pour produire un seul résultat présenté sous forme d'une case dans la carte des caractéristiques. Enfin, le filtre des poids est glissé par un pas S sur le reste des champs réceptifs de la matrice en entrée, et cette opération est répétée pour tous les autres champs.

Dans une convolution, la taille de la nouvelle carte des caractéristiques $N^{(t+1)}$ est calculée en fonction de 4 hyper-paramètres : la taille de l'ancienne carte des caractéristiques ou la matrice en entrée $N^{(t)}$, la taille du filtre F, la valeur du pas S et la valeur de la marge P (équation 2.1). La

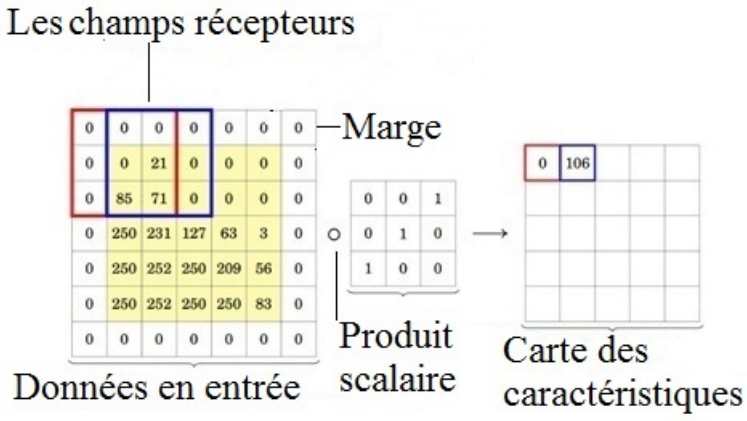

FIGURE 2.4 – *Une opération de convolution.*

marge représente des valeurs nulles qui entourent la matrice en entrée. Cette marge empêche le filtre de dépasser le cadre de cette matrice. L'application de N_c filtres sur les données en entrées résulte une carte des caractéristiques d'une taille de $N^{(t+1)} \times N^{(t+1)} \times N_c$, où N_c est sa profondeur. La concaténation des cartes des caractéristiques forme une couche de convolution.

$$N^{(t+1)} = \frac{N^{(t)} - F + 2P}{S} - 1 \tag{2.1}$$

Le processus d'une convolution illustre la stratégie de réduction de dimensionnalité d'un CNN, où chaque case (neurone) de la carte de caractéristique courante est connectée seulement à un sous ensemble des neurones en entrées (champs récepteurs). En plus, l'application du même filtre sur toute la carte des caractéristiques lui permet de découvrir les attributs précédemment détectés dans différentes zones de l'image.

À la fin de chaque opération de convolution, la fonction d'activation ReLu est appliquée sur la couche de convolution résultante afin d'améliorer la généralisation.

2.3.2 Couche de pooling

Le rôle de la couche de pooling est de réduire la dimensionnalité des couches de convolution résultantes. Le but de cette réduction est d'améliorer la précision par la sélection des attributs dominants. En plus, l'optimisation du nombre des paramètres permet de réduire la taille du modèle et d'optimiser la complexité temporelle.

La taille de la matrice résultante de l'opération de pooling est calculée par l'équation 2.1 avec $P = 0$. Il existe deux types d'opérations pooling : Max-pooling et Avg-pooling. L'opération de Max-pooling renvoie la valeur maximale du champ réceptif tandis que l'opération Avg-pooling renvoie la moyenne des valeurs. Max-pooling est la forme la plus utilisée dans la majorité des architectures de type CNN (figure 2.5).

2.3.3 Couche entièrement connectée

Dans un CNN, les couches entièrement connectées (FC) ont la même structure qu'un MLP. Le but de ces couches est d'apprendre les combi-

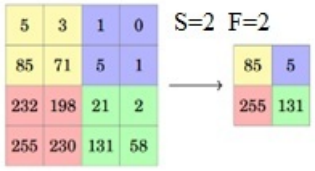

FIGURE 2.5 – *Une opération de Max-pooling.*

naisons non linéaires entre les caractéristiques extraites par les couches de convolution. Le résultat de la dernière couche de convolution $[N, N, N_c]$ est aplatit dans un vecteur de taille $[N \times N \times N_c]$. Ce vecteur présente la couche d'entrée à l'ensemble des couches entièrement connectées. En classification supervisée, la dernière couche est utilisée pour la prédiction en se basant sur la fonction d'activation Softmax.

2.4 Les types d'apprentissage

L'apprentissage dans un réseau de neurones convolutif peut être effectué de deux façons : apprentissage à partir des paramètres initialisés aléatoirement (training from scratch) ou apprentissage par transfert.

2.4.1 Apprentissage à partir des initialisations aléatoires

Dans le chapitre précédent, nous avons discuté le processus d'apprentissage dans un MLP par la méthode de rétropropagation. Les réseaux de neurones convolutif sont basés sur la même méthode. Le but principal des CNN est de réduire l'erreur de la fonction du coût par l'ajustement des filtres. Ces filtres représentent les paramètres w de l'apprentissage. Comme nous l'avons mentionné précédemment, les CNN sont caractérisés par le partage des paramètres. Cette spécificité permet de réduire le nombre des paramètres dans une couche de convolution et d'optimiser la complexité temporelle et spatiale. Ce partage et la représentation des paramètres dans des filtres exigent d'adapter la fonction de rétropropagation sur les couches de convolution et de pooling.

L'apprentissage dans un CNN commence par une propagation vers l'avant pour calculer la valeur de la fonction du coût en fonction des entrées. Ensuite, les filtres initialisés aléatoirement sont ajustés par un processus de rétropropagation. Ce processus est répété pour un certain nombre d'itérations jusqu'à atteindre le critère d'arrêt. Le critère peut dépendre d'un nombre fixe d'itérations, arrêt prématuré, ou une convergence.

2.4.2 Apprentissage par transfert et fine-tuning

L'apprentissage par transfert est une méthode d'apprentissage automatique qui permet de réutiliser un modèle développé précédemment pour l'apprentissage d'une tâche A dans une autre tâche B. Ces tâches peuvent être similaires ou différentes selon la nature du processus de l'apprentissage transféré.

La recherche en apprentissage transféré a commencé depuis les années 1995 [51], où ils étaient inspirés du comportement des humains dans leur méthode d'apprentissage basée sur la connaissance acquise précédemment. Ces méthodes sont catégorisées en : apprentissage par transfert inductif, tranductif, et non supervisé [50]. Dans la méthode de transfert inductif, les tâches source et cible sont différentes et appartiennent au même domaine. Tandis qu'en apprentissage tranductif, les tâches sont similaires et différentes dans la distribution des probabilités dans l'espace des attributs. En apprentissage non supervisé, les données de l'apprentissage ne sont pas catégorisées dans les domaines source et cible.

Les méthodes de l'apprentissage par transfert en apprentissage profond appartiennent à la catégorie de l'apprentissage inductif. Elles sont exploitées dans différents domaines comme le traitement du langage naturel [52] et la vision par ordinateur [53].

En apprentissage automatique classique, les algorithmes d'apprentissage sont caractérisés par leur dépendance de la distribution des attributs en entrée, où les données source et cible doivent avoir la même représentation des données. Le changement de la distribution des attributs exige de reprendre l'apprentissage à partir du début. Contrairement à ces algorithmes, en apprentissage profond, il est possible de transférer la connaissance à partir des modèles formés précédemment, où la distribution des paramètres (poids) des réseaux DL offre cette possibilité.

En vision par ordinateur, l'apprentissage par transfert à partir des réseaux CNN a connu un grand succès grâce à leur nature hiérarchique [197]. Les premières couches représentent des caractéristiques générales comme les filtres de Gabor. Ils permettent de détecter des formes basiques (courbes et bordures). En revanche, les couches profondes permettent de modéliser des caractéristiques plus complexes et liées au domaine d'application de la base d'apprentissage. Les caractéristiques communes des premières couches offre la possibilité d'effectuer un apprentissage transféré entre différentes tâches.

Nous avons détaillé précédemment les différents problèmes de surapprentissage liés au manque de données. L'apprentissage par transfert est l'une des méthodes proposées dans l'état de l'art pour résoudre cette limitation. Cette technique est généralement utilisée quand le volume de données de la tâche cible est limité. Elle permet de réutiliser les premières couches des modèles générés à partir des grands volumes de données. L'apprentissage par transfert à partir des modèles entrainés sur la base d'apprentissage ImageNet [54] a été largement exploité dans différents domaines grâce à son grand volume de données (15 millions) et son nombre important de catégories (22 000).

L'apprentissage par transfert en CNN peut être utilisé sous forme de trois méthodes : (a) l'exploitation du modèle source comme un module d'extraction de caractéristiques, (b) transférer un sous ensemble de couches et réajuster le reste, et (c) transférer et réajuster tous les couches [196].

Module d'extraction des caractéristiques

Cette stratégie est considérée comme une hybridation entre les algorithmes d'apprentissage automatique et les réseaux DL, où des réseaux de type CNN sont exploités pour l'extraction des caractéristiques et les algorithmes ML pour la classification. La figure 2.6 illustre ce processus, généralement, toutes les couches du modèle source sont transférées au modèle cible. Ensuite, la dernière couche est supprimée. Les données de la base d'apprentissage cible sont passées ensuite au modèle. Cela permet de générer une base d'apprentissage structurée sous forme d'attributs et d'instances. Enfin, cette base est classifiée par un algorithme de type ML.

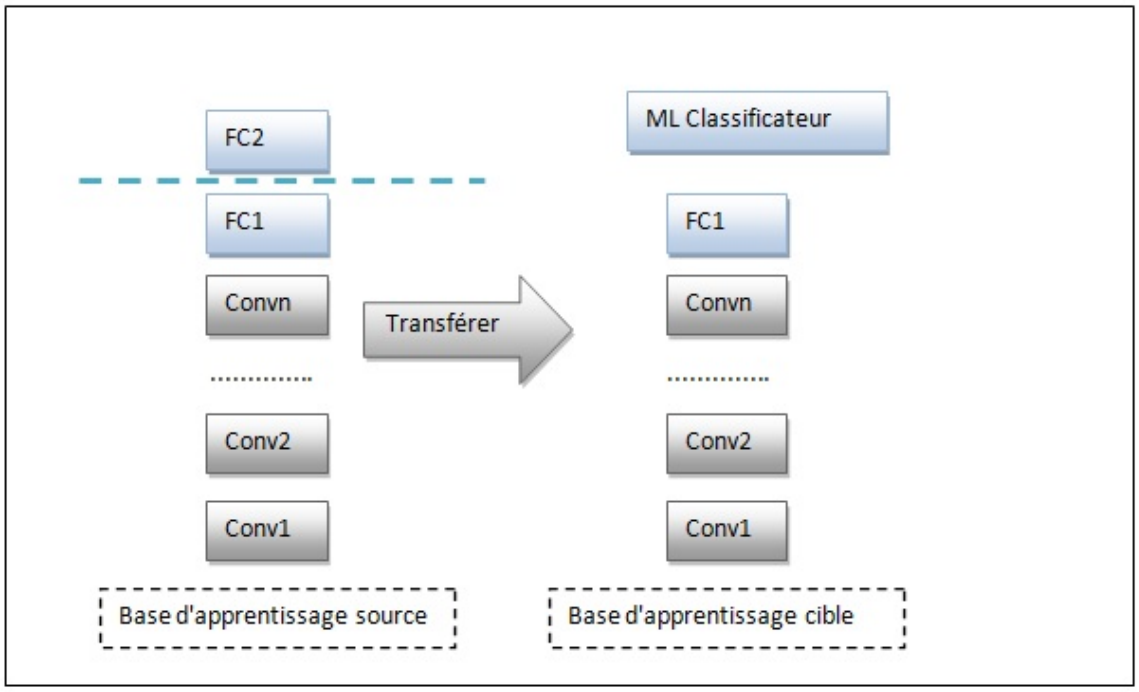

FIGURE 2.6 – *L'utilisation des réseaux de neurones convolutifs pour l'extraction des caractéristiques.*

Fine Tuning

Comme nous l'avons discuté auparavant, les premières couches permettent de représenter des caractéristiques générales, tandis que les couches profondes sont liées au domaine d'application source. Afin d'adapter ces couches au domaine d'application cible, un sous ensemble de couches profondes (convolutions et entièrement connecté) est réajusté par un processus d'apprentissage (figure 2.7).

2.5 Les architectures communes

Récemment, les réseaux de neurones convolutif ont été largement exploités en vision par ordinateur grâce à leur stratégie de réduction de paramètres et la disponibilité des grands volumes de données. En plus,

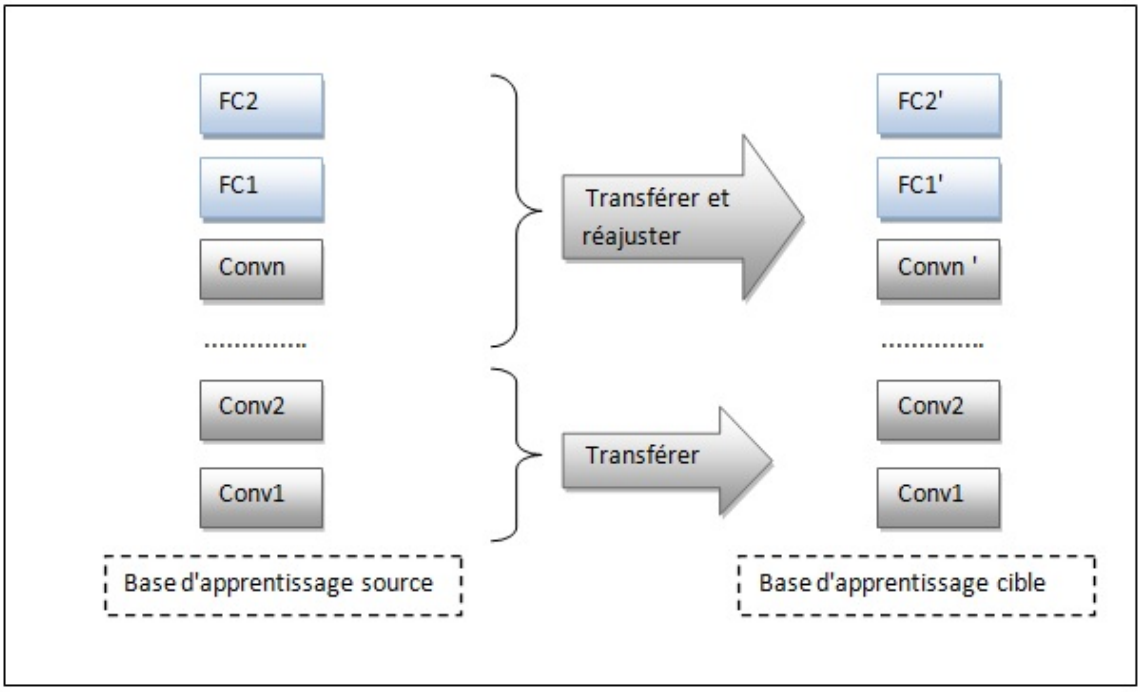

FIGURE 2.7 – *Le processus de fine-tuning dans un réseau de neurones convolutif.*

l'évolution de la capacité de stockage et de calcule (GPU) a encouragé la communauté de vision par ordinateur à proposer d'autres architectures de type CNN plus profondes. Ces architectures optimisent les couches de convolution classiques. Le but principal de cette variation est de réduire le nombre des paramètres et d'ajouter des couches supplémentaires qui permettent d'améliorer la non-linéarité. Cette non-linéarité est assurée par les fonctions d'activations qui développent la capacité du réseau dans la résolution des problèmes complexes.

En 2012, le réseau AlexNet [Krizhevsky *et al.* 2012] a été proposé. Ce réseau est composé de 5 couches de convolution et 3 couches entièrement connectées. Malgré sa simplicité, il a accompli le meilleur taux d'erreur dans l'état de l'art sur la base d'apprentissage ImageNet. Ce résultat a encouragé la communauté de vision par ordinateur à proposer d'autres versions optimisées de type CNN par l'analyse de la structure. En 2013, le réseau ZFNet a été développé [Zeiler & Fergus 2014]. Ce réseau imite la structure de AlexNet avec une légère réduction dans la taille du premier filtre. Le but de cette modification est de conserver plus d'information dans la première couche de convolution.

Augmenter la profondeur d'un réseau de neurones permet d'améliorer sa non-linéarité et donc sa capacité dans la reconnaissance des objets complexes. D'autre part, cette augmentation accroit le nombre de paramètres, et cela augmente le risque de sur-apprentissage et les exigences en termes de stockage. Afin d'éviter ces problèmes, d'autres stratégies de réduction de dimensionnalité ont été développées dans les couches de convolution.

Le réseau VGGNet [Simonyan & Zisserman 2014b] propose des configurations dont la profondeur varie de 11 à 19 couches. Ce réseau suggère

de réduire la taille des filtres à F=3 pour éviter l'augmentation exponentielle des paramètres par l'ajout des couches supplémentaires.

En 2015, le réseau Inception [Szegedy *et al.* 2015] a été développé. Ce réseau propose une stratégie de réduction de paramètres à travers les modules d'Inception. L'efficacité des blocs d'Inception dans la réduction de la dimensionnalité a encouragé les chercheurs en vision par ordinateur à proposer d'autres versions optimisées des modules d'Inceptions [Szegedy *et al.* 2017, Chollet 2017]. Inception-ResNet [Szegedy *et al.* 2017] intègre des liens résiduels dans les blocs d'inception et Xception [Chollet 2017] utilise des blocs d'inception extrêmes qui ont une architecture similaire aux convolutions séparables en profondeur. Ces deux versions ont prouvé leur efficacité par rapport à l'architecture Inception initiale [Szegedy *et al.* 2016].

Les réseaux CNN profonds sont caractérisés par leur efficacité dans la classification des objets complexes. Malgré cette spécificité, les réseaux très profonds risquent la dégradation du gradient. Afin de résoudre ce problème, le réseau ResNet [He *et al.* 2016] propose des structures basées sur les blocs résiduels dans des configurations composées de 18 à 152 couches. En 2017, le réseau DenseNet [Huang *et al.* 2017] a été développé afin de proposer des configurations plus profondes par rapport à ResNet en réduisant le nombre des paramètres. L'étude expérimentale sur la base d'apprentissage ImageNet a montré qu'un réseau DenseNet composé de 201 couches et 20 millions de paramètres a la même performance qu'un réseau de type ResNet composé de plus de 40 millions de paramètres.

Afin d'adapter les réseaux CNN aux appareils mobiles, les architectures MobileNetV1 [Howard *et al.* 2017], MobileNetV2 [Sandler *et al.* 2018], et ShuffleNet [Zhang *et al.* 2018] ont été développés. Dans ces architectures, la taille du réseau en termes de profondeur et nombre de paramètres est contrôlée par deux hyperparamètres : les multiplicateurs de largeur et de résolution.

La détection des objets est une technique en vision par ordinateur qui permet de classifier et détecter plusieurs objets dans une image. Cette tâche est caractérisée par sa complexité élevée par rapport aux méthodes de classifications, car elle exige une étape de localisation en plus. La localisation propose les régions d'intérêts candidates, ensuite, ces régions sont classifiées. Plusieurs méthodes de détection d'objets ont été proposées dans l'état de l'art. Le réseau R-CNN [Girshick *et al.* 2014] combine entre la méthode recherche sélective pour la détection des régions et les réseaux CNN pour la classification. Malgré son efficacité, il n'est pas adapté aux applications en temps réel à cause de sa complexité temporelle élevée. Afin de réduire cette complexité, d'autres structures ont été proposées : Fast R-CNN [Girshick 2015], Faster R-CNN [Ren *et al.* 2015], et YOLO [Redmon *et al.* 2016].

La segmentation sémantique est une technique qui permet de classifier chaque pixel dans l'image en entrée. Cette méthode est caractérisée par sa complexité élevée par rapport à la classification et à la détection des objets. Pour optimiser sa complexité, plusieurs architectures à base de CNN ont été proposées : FCN [Long *et al.* 2015], DeepLab [Chen *et al.* 2015, Chen *et al.* 2017b], SegNet [Badrinarayanan *et al.* 2017], U-net [Ronneberger *et al.* 2015], et Mask R-CNN [He *et al.* 2017].

Dans ce qui suit, nous détaillerons la structure des architectures connues de type CNN en classification, en détection des objets, et en segmentation.

2.5.1 Classification

LeNet

LeNet [LeCun *et al.* 1998] est la première architecture de type CNN proposée pour la classification supervisée. La figure 2.8 illustre la structure du réseau LeNet. Ce réseau est composé de 7 couches au total : 3 couches de convolution (C_x), 2 couches de Avg-pooling (S_x) et 2 couches entièrement connectées (F_x).

Les couches C_x et S_x sont composées d'un certain nombre de cartes de caractéristiques d'une taille définie (nombre@largeur $\times$ hauteur). Cette figure montre que la taille des couches internes est réduite par rapport aux premières couches, tandis qu'elles sont plus profondes par rapport aux couches en entrée. Dans cette architecture, la taille des filtres dans les couches de convolution a été fixée à 5x5 et les entrées sont des images de taille 32×32.

L'apprentissage à base de rétropropagation sur la base d'apprentissage MNIST a montré l'efficacité de l'algorithme d'apprentissage profond LeNet par rapport aux algorithmes d'apprentissage automatique classiques SVM et KNN. La base d'apprentissage MNIST a été conçue pour la classification des chiffres manuscrits et contient 60 000 instances pour l'apprentissage et 10 000 instances pour le test. Ces instances sont des images en noir et blanc normalisées et centrées.

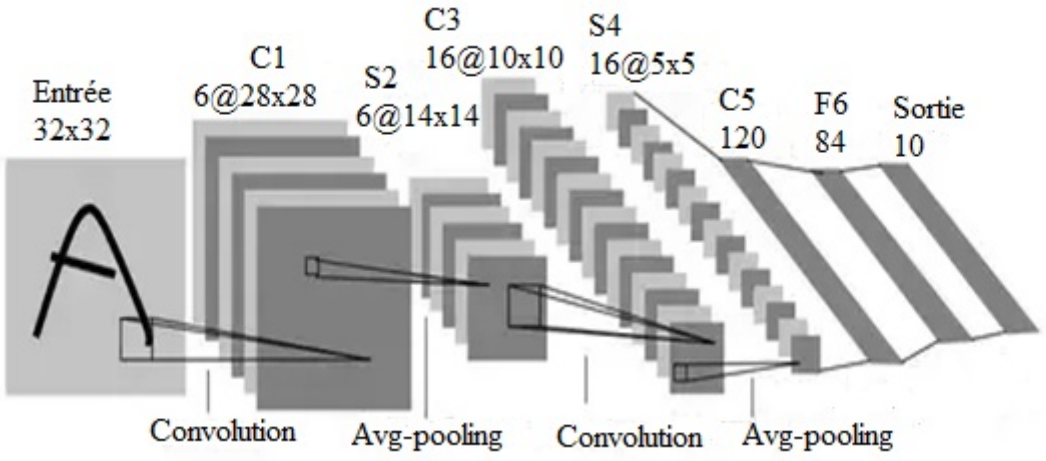

FIGURE 2.8 – *La structure du réseau LeNet [LeCun et al. 1998].*

AlexNet

En 2012, le réseau AlexNet [Krizhevsky *et al.* 2012] a été proposé dans la compétition ImageNet de reconnaissance visuelle à grande échelle (ILSVRC). Cette compétition utilise un sous ensemble de la base d'apprentissage ImageNet composé de 1000 catégories, 1.2 million instances d'ap-

prentissage, 50 000 instances de validation, et 150 000 instances de test. Dans cette compétition, le réseau AlexNet a atteint le meilleur taux d'erreur par rapport aux autres méthodes d'apprentissage automatique classiques. Ce résultat était un point important dans l'historique des réseaux d'apprentissage profond, car l'intérêt s'est attiré vers les méthodes DL plus que les méthodes ML classiques en vision par ordinateur.

Le réseau AlexNet a une architecture similaire et plus profonde par rapport à LeNet (figure 2.9). Ce réseau est composé de 5 couches de convolution et 3 couches entièrement connectées. Selon la figure 2.9, la première couche de convolution utilise 96 filtres de taille 11×11, tandis que les autres couches convolutives sont basées sur des filtres de taille 5×5 et 3×3. La première, la deuxième et la cinquième couche de convolution sont suivies par des couches de max-pooling et les deux premières couches sont suivies par une opération de normalisation (local response normalisation (LRN)).

La méthode LRN permet d'améliorer la généralisation et la non-linéarité du réseau. Cette opération est appliquée sur les résultats de la fonction d'activation ReLu.

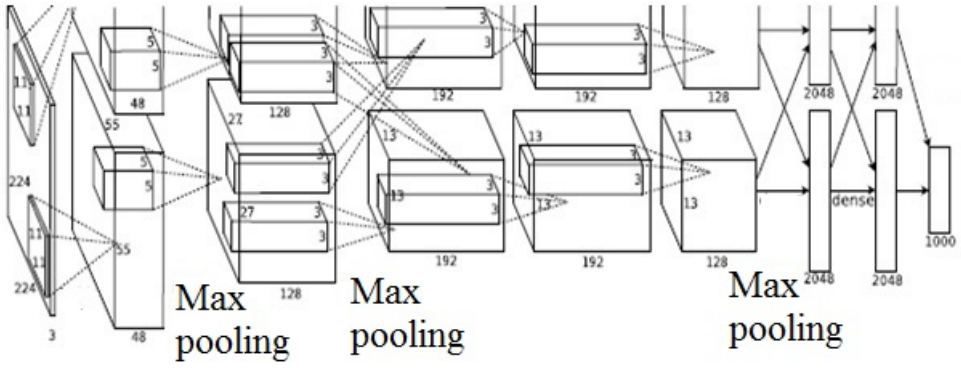

FIGURE 2.9 – *La structure du réseau AlexNet [Krizhevsky et al. 2012].*

Contrairement à LeNet, AlexNet propose l'exploitation de la fonction d'activation ReLu, car elle permet d'accélérer l'apprentissage par rapport à la fonction tanh. Cela conduit à une réduction remarquable dans la complexité temporelle dans le cas des grands modèles qui sont entrainés sur des données de taille considérable.

Malgré la taille considérable de la base d'apprentissage utilisée, le réseau AlexNet risque le problème de sur-apprentissage à cause du nombre élevé de paramètres (60 millions). Afin d'éviter ce problème, les techniques d'augmentation de données et de régularisation par abandon (Dropout) ont été exploitées. La méthode d'augmentation de données utilisée extrait des patchs aléatoires de taille 224×224. Ensuite, ces patchs sont augmentés par des réflexions horizontales. Dans la phase de test, la décision présente la moyenne des prédictions de l'ensemble des patchs. La méthode Dropout a été utilisée dans les deux premières couches entièrement connectées avec un taux d'abandon r = 0.5.

En apprentissage, la méthode descente avec inertie a été utilisée avec un lot de donnée de taille 128 et un taux d'apprentissage initialisé à 0.01. Ce taux est réduit manuellement 6 fois durant l'apprentissage.

Le processus d'apprentissage a pris de cinq à six jours de temps pour terminer l'exécution dans 90 époques sur deux GPUs de types NVIDIA GTX 580 3GB. L'utilisation des GPUs parallèles permet d'accélérer le temps d'exécution et d'offrir la possibilité de charger le modèle entier en mémoire, à cause de la mémoire limitée des GPUs (3 GB).

ZFNet

Malgré la performance de AlexNet dans ILSVRC, [Krizhevsky *et al.* 2012] n'ont pas justifié le choix des hyperparamètres (taille des filtres, nombre de couches) et comment les ajuster pour améliorer la performance du CNN. En plus, le comportement du réseau et son fonctionnement interne restent ambigus d'un point de vue scientifique.

Afin de comprendre le comportement des CNN et améliorer leur performance, [Zeiler & Fergus 2014] ont proposé une nouvelle méthode de visualisation qui permet de déchiffrer le contenu des couches intermédiaires. Le but de cette visualisation et d'étudier le comportement de AlexNet et de proposer la version optimisée ZFNet. La technique de visualisation est basée sur le réseau déconvolutionnel multicouche (Deconvnet) [Zeiler *et al.* 2011]. Ce réseau permet de projeter les cartes des caractéristiques internes aux entrées afin de visualiser leur contenu.

Le réseau Deconvnet est basé sur trois opérations : unpooling, rectification et filtrage. La méthode unpooling est l'inverse de l'opération pooling d'un CNN. Elle permet de restaurer le contenu des cartes des caractéristiques avant l'opération de pooling. La rectification est basée sur la fonction d'activation ReLu. Elle permet d'éliminer les valeurs négatives des cartes des caractéristiques. Le filtrage est l'opération inverse d'une convolution. Il est basée sur la version transposée des filtres utilisés dans l'opération de convolution. Cette opération applique les filtres transposés sur les cartes des caractéristiques pour obtenir la couche de convolution précédente. Ces trois opérations sont répétées successivement sur les cartes des caractéristiques internes jusqu'à atteindre l'espace des pixels en entrée.

Afin de visualiser le contenu des cartes des caractéristiques, le réseau Deconvnet est lié à chaque couche du réseau CNN. La figure 2.10 illustre le résultat obtenu par Deconvnet sur les 9 meilleures activations des couches 2 et 5. Ce résultat montre la nature hiérarchique du réseau CNN, où la 2 éme couche représente des caractéristiques simples comme les coins et les bords. Ensuite, ces caractéristiques deviennent plus complexes dans les couches plus profondes jusqu'à la dernière couche convolutionnelle (couche 5), où les objets sont complètement visibles.

Cette technique de visualisation a permis de détecter quelques problèmes. Les premiers filtres présentent un mélange d'informations d'une fréquence variable. En plus, la 2 éme couche illustre la présence d'un bruit résultant de la valeur large du pas $S = 4$ dans l'opération de convolution. Afin de résoudre ces problèmes, la taille du premier filtre a été réduite de 11×11 à 7×7 et la valeur du pas S de 4 à 2. La figure 2.11 illustre la nouvelle architecture du réseau ZFNET proposée.

Les modifications proposées ont permis d'améliorer la performance du réseau AlexNet par 1.7 %. Ce résultat illustre l'utilité des techniques de visualisation dans l'ajustement des hyperparamètres.

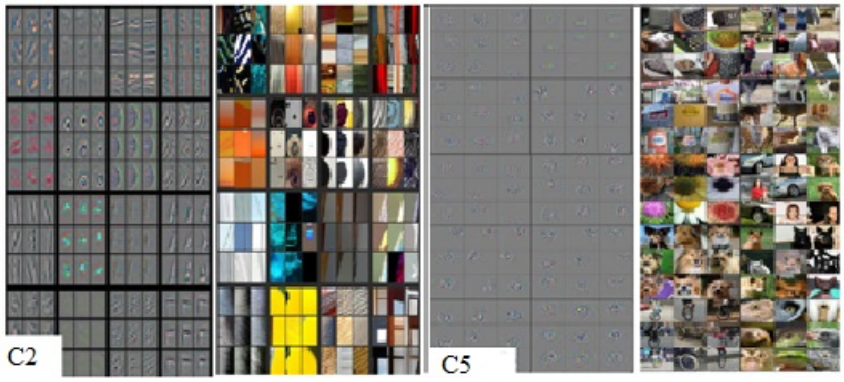

FIGURE 2.10 – *Le résultat de l'application du réseau Deconvnet sur les couches 2 et 5 [Zeiler & Fergus 2014].*

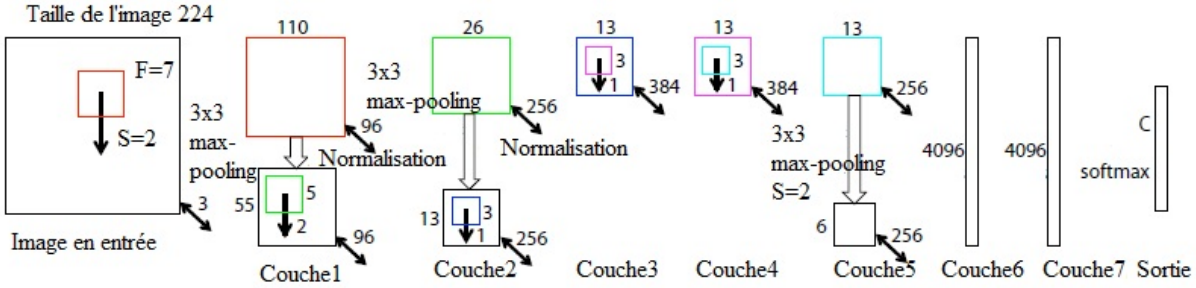

FIGURE 2.11 – *La structure du réseau ZFNET [Zeiler & Fergus 2014].*

VGGNet

VGGNet [Simonyan & Zisserman 2014b] est un réseau de neurones convolutif qui est basé sur les mêmes principes du réseau AlexNet [Krizhevsky *et al.* 2012]. Le but de cette version est de proposer des configurations profondes (16 à 19 couches) en se basant sur la technique de stabilisation structurelle. Cette technique permet de contrôler le nombre des paramètres dans les réseaux profonds afin de diminuer les risques de sur-apprentissage.

Pour réduire le nombre des paramètres, le réseau VGGNet propose de diminuer la taille des filtres de 7×7 et 5×5 à 3×3. Ce changement permet d'ajouter plus de couches intermédiaires sans risquer d'une augmentation exponentielle dans le nombre des paramètres.

L'étude comparative entre le nombre des paramètres dans trois couches de convolution empilées associées à des filtres de taille 3×3 et une seule couche de convolution associée à un filtre de taille 7×7 a montré que les petits filtres réduisent le nombre des paramètres.

Si chaque couche de convolution a une profondeur C, le nombre de paramètres dans 3 couches de convolution empilées (3×3) est $3(3^2C^2)$. Tandis que le nombre de paramètres dans une seule couche associée à des filtres de taille 7×7 est 7^2C^2. En résumé, l'empilement des couches de convolution associées à des petits champs réceptifs permet de réduire le nombre des paramètres et d'améliorer la non-linéarité du réseau à travers les fonctions d'activation (ReLu) supplémentaires.

La figure 2.12 illustre les configurations du réseau VGGNet. Ces configurations varient dans le nombre de couches de convolution et la taille des filtres. L'annotation convF-P exprime une couche de convolution associée à des filtres de taille FxF et d'une profondeur P.

La configuration A est composée de 11 couches (8 couches de convolution et 3 couches entièrement connectées). La deuxième configuration A-LRN intègre dans A une opération de normalisation (LRN) après la première couche de convolution. La configuration B ajoute à A deux couches de convolution. La configuration C intègre dans B trois couches de convolution supplémentaires associées à des filtres de taille 1x1. Ces filtres permettent d'améliorer la non-linéarité du réseau à travers les fonctions ReLu, car ils représentent une projection dans un espace de même dimensionnalité, où les couches d'entée ont la même dimensionnalité que les couches en sortie. Enfin, les configurations D et E intègrent dans B des couches de convolution supplémentaires de taille 3×3.

A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

FIGURE 2.12 – *Les configurations du réseau VGGNET [Simonyan & Zisserman 2014b].*

L'étude expérimentale sur la base d'apprentissage ImageNet a montré l'effet positif de la profondeur sur la performance, où les réseaux les plus profonds sont les plus performants.L'étude comparative entre les configurations A et A-LRN indique que l'opération de normalisation (LRN) n'améliore pas la performance du modèle A. En plus, contrairement aux

couches de convolution associées à des filtres de taille 3x3, les couches de convolution associées à des filtres de taille 1x1 ont un effet négatif sur la performance du réseau.

Inception

La méthode la plus simple pour améliorer la performance d'un réseau est d'augmenter sa taille en termes de largeur (nombre de paramètres dans chaque couche) et de profondeur (nombre de couches).

Malgré l'efficacité des réseaux profonds, ils ont plusieurs inconvénients liés au risque de sur-apprentissage sur les volumes limités de données. En plus, les réseaux profonds sont plus exigeants en termes de stockage et capacité de calcul. Afin de résoudre ces problèmes et d'adapter l'utilisation des réseaux profonds sur les applications en temps réel, la recherche s'est focalisée sur architectures partiellement connectées plus que les architectures entièrement connectées.

Le réseau Inception [Szegedy *et al.* 2015] est un réseau de neurones convolutif qui propose l'exploitation des modules d'Inception. Ces modules présentent des variantes optimisées des couches de convolution classiques. Les modules d'Inception introduisent des connexions partielles à l'intérieur d'une couche de convolution afin de réduire sa dimensionnalité.

La figure 2.13 illustre la structure d'un module d'Inception. Ce module utilise des filtres de taille variable ($1x1$, 3×3, et 5×5) qui sont appliqués sur la même couche de convolution. Ensuite, les cartes des caractéristiques résultantes sont empilées pour former la couche de convolution suivante.

La variation dans la taille des filtres permet d'éviter les problèmes d'alignement des patchs. Malgré ces connexions partielles, l'encapsulation des cartes des caractéristiques augmente rapidement la profondeur des couches de convolution. Afin de résoudre ce problème, des filtres de taille 1x1 ont été introduits avant les filtres de taille $3x3$ et $5x5$. Ces filtres réduisent la profondeur de la couche de convolution avant l'application des autres filtres, et améliorent la non-linéarité par les fonctions d'activation ReLu.

En résumé, les modules d'Inception augmentent la profondeur du réseau en contrôlant en parallèle la complexité de calcul par les techniques de réduction de dimensionnalité. En plus, la variation dans la taille des filtres permet de traiter l'information en entrée dans différentes échelles.

La figure 2.14 illustre la structure du réseau Inception qui est composé de 22 couches au total. Ce réseau est formé par une concaténation des couches de convolution classiques, modules d'Inception et des couches d'Avg-pooling.

Contrairement aux architectures proposées précédemment, Inception propose l'intégration des classificateurs auxiliaires qui sont connectés aux couches intermédiaires. Cette technique introduit la force discriminative des réseaux moins profonds à travers les couches intermédiaires, où le taux d'erreur est calculé en fonction d'une moyenne pondérée des résultats des 3 couches de prédictions. En résumé, le réseau Inception a montré son efficacité en réduction de dimensionnalité à travers les modules d'Inception. Ces modules proposent un réseau profond et

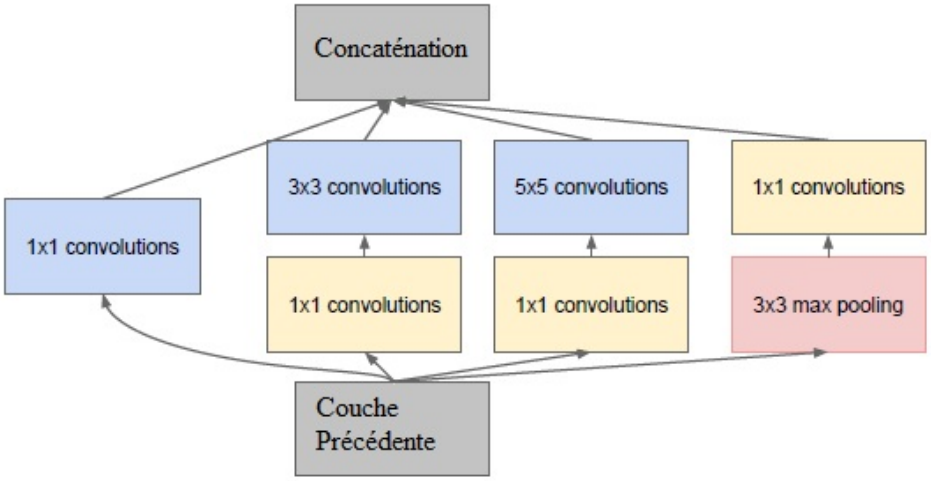

FIGURE 2.13 – *La structure d'un module d'Inception [Szegedy et al. 2015].*

plus performant en réduisant 12× le nombre de paramètres de AlexNet [Krizhevsky *et al.* 2012].

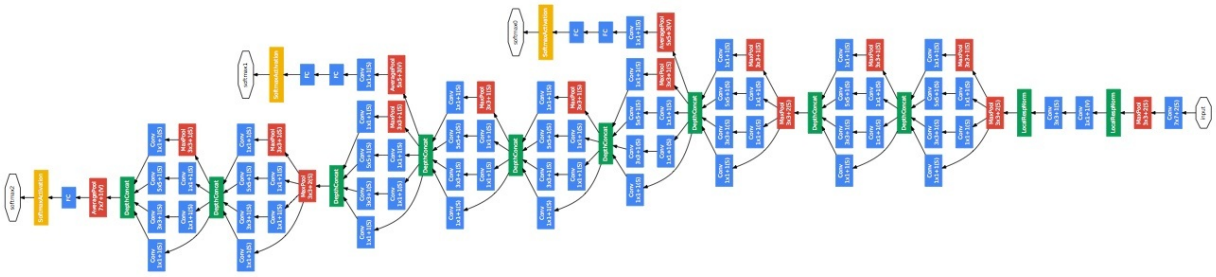

FIGURE 2.14 – *La structure du réseau Inception [Szegedy et al. 2015].*

InceptionV2 et InceptionV3

InceptionV2 et InceptionV3 [Szegedy *et al.* 2016] sont des versions optimisées du réseau Inception [Szegedy *et al.* 2015]. Ces architectures proposent des variantes de blocs d'Inception pour réduire le nombre des multiplications dans une convolution et donc optimiser la complexité de calcul. Ces variantes sont basées sur deux techniques de factorisation : factorisation des convolutions associées à des filtres de grande taille et la factorisation spatiale en convolutions asymétriques.

La première technique propose de réduire la taille des filtres de 5×5 à 3×3, car les filtres de taille 5x5 sont 2.78 fois plus couteux par rapport aux filtres de taille 3×3. Malgré l'efficacité de cette réduction dans la diminution du nombre des paramètres, elle peut engendrer une perte d'information. Pour éviter ce problème, la technique de factorisation propose de remplacer une couche convolution associée à un filtre de taille 5x5 par deux couches de convolution associées à des filtres de taille 3×3. Cela

réduit le nombre de paramètres de 25%. Les figures 2.15 et 2.16 illustrent la structure des blocs d'Inception dans InceptionV2 et InceptionV3.

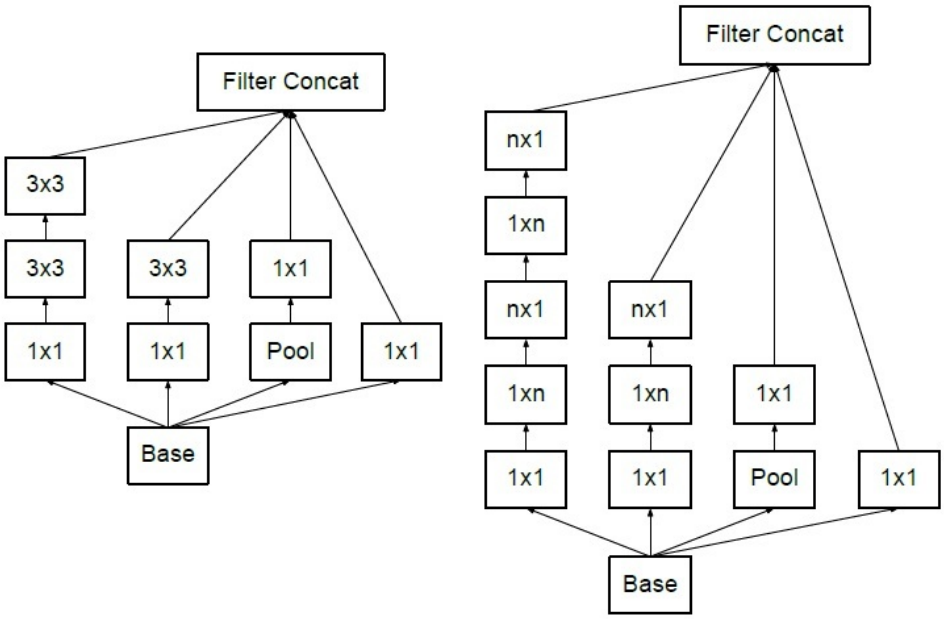

FIGURE 2.15 – *La structure des blocs d'Inception dans InceptionV2 et InceptionV3 [Szegedy et al. 2016].*

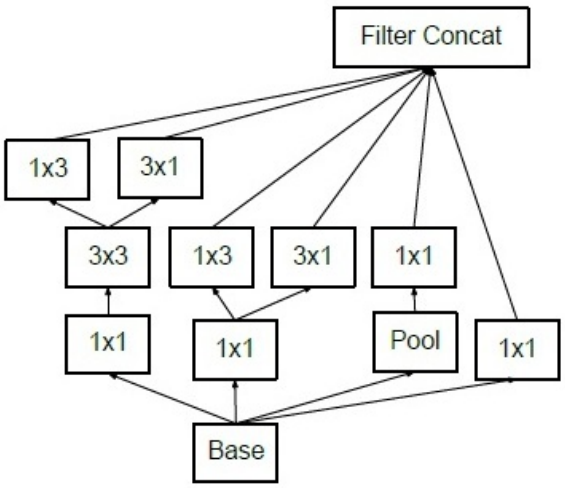

FIGURE 2.16 – *La structure des blocs d'Inception dans InceptionV2 et InceptionV3 [Szegedy et al. 2016].*

La deuxième technique propose de remplacer les convolutions classiques (nxn) par des convolutions asymétriques (nx1 et 1xn). Cette méthode réduit la complexité de calcul par 33% pour n=3 (Figure 2.15(bloc droit)).

En plus des techniques citées précédemment et exploitées dans InceptionV2, inceptionV3 propose l'utilisation de : (a) la méthode de régularisation par lissage (label smoothing), (b) les classificateurs auxiliaires où la couche entièrement connectée est normalisée par la méthode de normali-

sation par lot, et (c) la factorisation des filtres de taille 7×7 à des filtres asymétriques (1×7 et 7×1). La figure 2.17 illustre la structure du réseau InceptionV3 qui est composé de 42 couches au total.

Type	Taille du patch/Pas ou remarque	Taille d'entrée
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×73×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
3×Inception	figure (A)	35×35×288
5×Inception	figure (C)	17×17×768
2×Inception	figure (B)	8×8×1280
pool	8×8	$8 \times 8 \times 2048$
linéaire	logits	$1 \times 1 \times 2048$
softmax	Classificateur	$1 \times 1 \times 1000$

FIGURE 2.17 – *La structure du réseau InceptionV3 [Szegedy et al. 2016].*

L'étude comparative entre les résultats des réseaux Inception, InceptionV3 et VGGNet a montré l'efficacité de InceptionV3 sur la base d'apprentissage ImageNet.

ResNet

ResNet [He *et al.* 2016] est un réseau de neurones convolutif basé sur des blocs résiduels. Le but principal de cette architecture est de résoudre le problème de dégradation de gradient. Ce problème apparait dans les réseaux très profonds, où la précision commence à être saturée ensuite dégrade rapidement à cause de la diminution des valeurs des gradients. Afin de résoudre ce problème, les blocs résiduels ont été introduits.

Les blocs résiduels (figure 2.18) représentent des connexions résiduelles entre la sortie de la couche précédente et la sortie de la couche courante. Ces connexions sont formulées par l'équation 2.2. Afin de réaliser l'addition, une projection linéaire W_s est effectuée pour obtenir des dimensions équivalentes (x et F(x)).

$$y = F(x) + x \tag{2.2}$$

Pour démontrer l'efficacité des blocs résiduels dans la résolution du problème de dégradation de gradient, [He *et al.* 2016] ont effectué une comparaison entre les performances des CNN avec et sans blocs résiduels (Figure 2.19). Ils ont utilisé deux réseaux inspirés de VGGNet et composés de 18 et 34 couches (figure 2.20) au total. L'étude comparative montre que dans la version classique des réseaux CNN, les réseaux moins profonds (18 couches) sont les plus performants. Tandis que dans les réseaux ResNet, les réseaux plus profonds sont les plus performants. Ces résultats

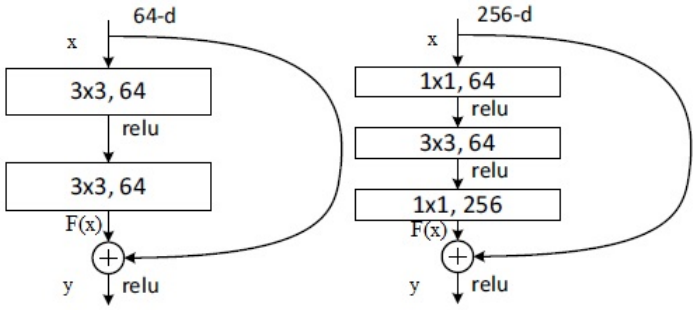

FIGURE 2.18 – *La structure des blocs résiduels [He et al. 2016].*

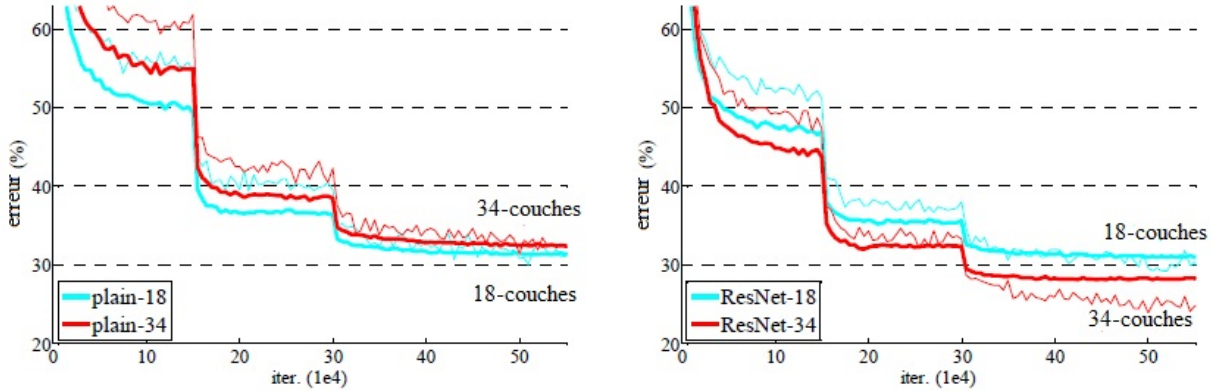

FIGURE 2.19 – *Comparaison entre les résultats des réseaux de neurones convolutif avec (ResNet) et sans blocs résiduels (plain) [He et al. 2016].*

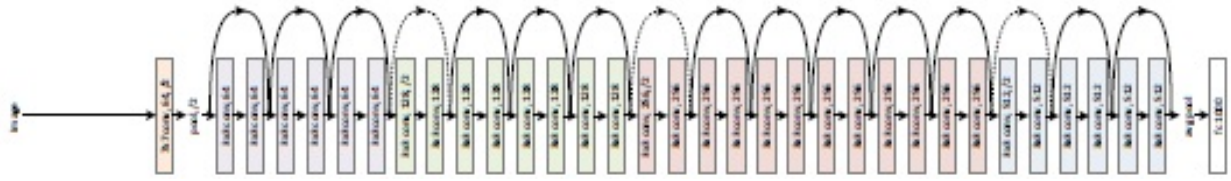

FIGURE 2.20 – *La structure du réseau ResNet (34 couches) [He et al. 2016].*

ont prouvé l'efficacité des blocs résiduels dans la résolution du problème de dégradation de gradient.

Pour analyser l'effet de la profondeur sur les réseaux ResNet, une étude comparative a été effectuée entre les réseaux ResNet-50, ResNet-101 et ResNet-152. Ces réseaux contiennent des connexions résiduelles entre 3 couches de convolution (Figure 2.18(B)) au lieu de 2 couches afin d'accélérer le temps d'apprentissage. Les résultats obtenus illustrent les avantages de la profondeur sur les performances et l'efficacité des blocs résiduels dans la résolution des problèmes de dégradation du gradient dans les réseaux très profonds.

Afin d'optimiser la structure des réseaux ResNet, plusieurs variantes ont été proposées récemment : ResNet−CutMix [Yun *et al.* 2019], SRM−ResNet [Lee *et al.* 2019], SGE−ResNet [Li *et al.* 2019], ResNet+SWA [Izmailov *et al.* 2018], AA−ResNet [Bello *et al.* 2019], Res2Net−DLA [Gao *et al.* 2019], et ResNeXt [Mahajan *et al.* 2018].

Inception-v4 et Inception-ResNet

[Szegedy *et al.* 2017] ont proposé deux variantes du réseau inceptionV3 [Szegedy *et al.* 2016] : inceptionV4 et Inception-ResNet.

Le réseau InceptionV4 est une version plus profonde, uniforme et simplifiée du réseau InceptionV3. Les figures 2.21, 2.22, et 2.23 illustrent les structures des modules d'Inception exploités (Inception-A, Inception-B, et Inception-C), respectivement, et la figure 2.24 présente l'architecture du réseau InceptionV4.

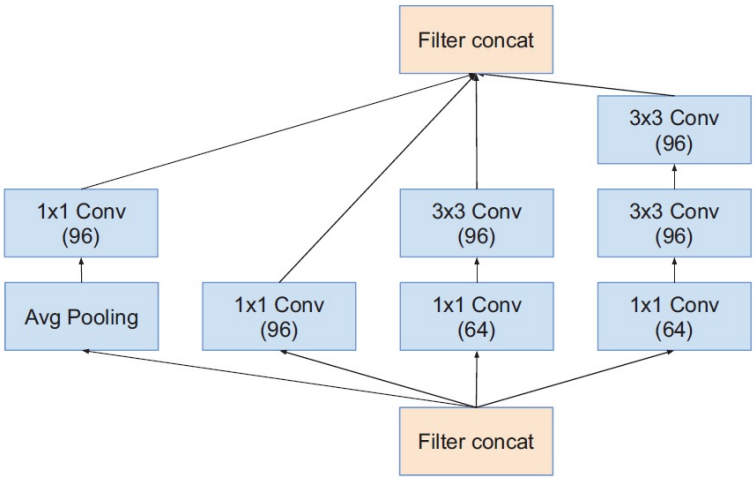

FIGURE 2.21 – *La structure des modules d'Inception (A) du réseau InceptionV4 [Szegedy et al. 2017].*

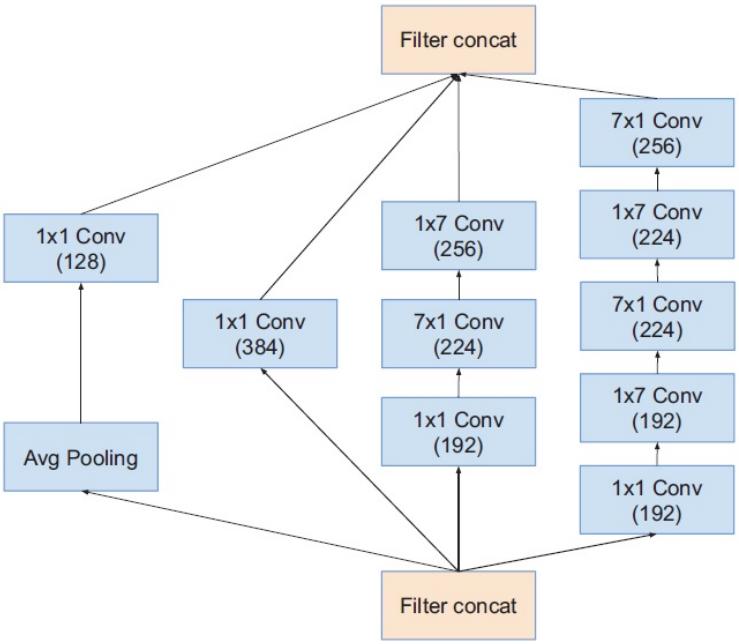

FIGURE 2.22 – *La structure des modules d'Inception (B) du réseau InceptionV4 [Szegedy et al. 2017].*

Le réseau Inception-ResNet propose l'hybridation entre les modules introduits dans le réseau InceptionV3 [Szegedy *et al.* 2016] et ResNet

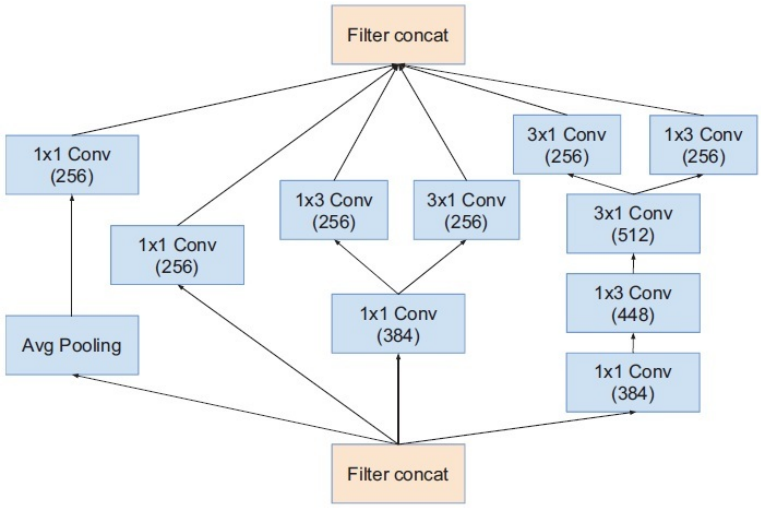

FIGURE 2.23 – *La structure des modules d'Inception (C) du réseau InceptionV4 [Szegedy et al. 2017].*

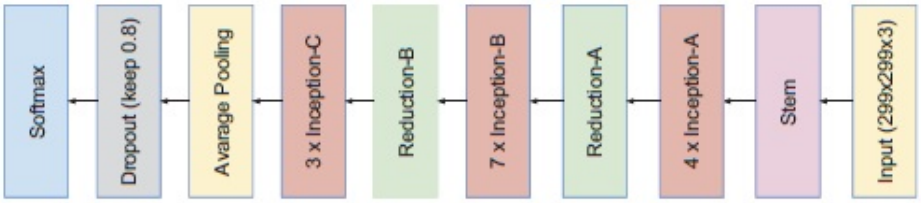

FIGURE 2.24 – *La structure du réseau InceptionV4 [Szegedy et al. 2017].*

[He *et al.* 2016]. Le but de cette combinaison est d'accélérer le temps d'apprentissage du réseau Inception et d'éviter le problème de dégradation de gradient des réseaux très profonds. Dans les couches d'inception, des filtres de taille 1×1 sont utilisés avant le lien résiduel afin d'obtenir la même dimensionnalité des données en entrée et de réaliser l'addition.

L'étude expérimentale sur la base d'apprentissage ImageNet a montré que les blocs résiduels ne sont pas nécessaires dans certains cas. Par exemple, le réseau InceptionV4 est plus performant par rapport à Inception-ResNetV1. Tandis que le réseau Inception-ResNetV2 était le plus performant en comparant avec tous les autres réseaux (IncetionV3, IncetionV4, et Inception-ResNetV1).

Xception

Xception [Chollet 2017] est un réseau de neurones convolutif basé sur les blocs extrêmes d'Inception. Ces blocs sont semblables aux convolutions séparables en profondeur (Depthwise separable convolutions (DSC)) que nous détaillerons par la suite dans le réseau MobileNet. La figure 2.25 illustre la structure d'un bloc extrême d'Inception qui est caractérisé par des convolutions associées à des filtres de taille 1×1. Les convolutions associées à des filtres de taille 3×3 sont appliquées par la suite sur des segments non superposés des cartes des caractéristiques résultantes.

Contrairement à DSC, dans un bloc extrême d'Inception, les convolu-

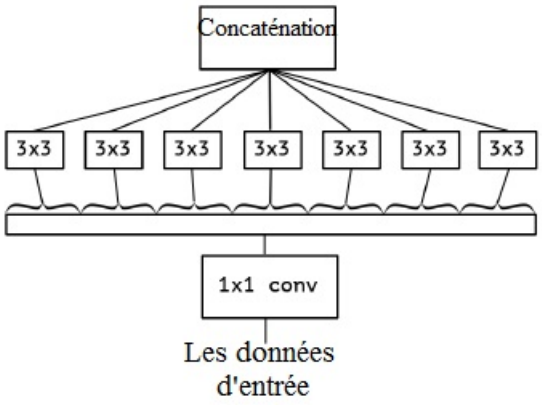

FIGURE 2.25 – *La structure d'un bloc extrême d'Inception [Chollet 2017].*

tions associées à des filtres de taille 1×1 sont appliquées en premier. En plus, ces blocs sont caractérisés par des fonction d'activation ReLu.

La figure 2.26 illustre la structure du réseau Xception qui est représenté par une concaténation des blocs extrêmes d'Inception liés par des connexions résiduelles. Ce réseau est composé de 36 couches de convolution pour l'extraction des caractéristiques et une couche de régression logistique pour la classification.

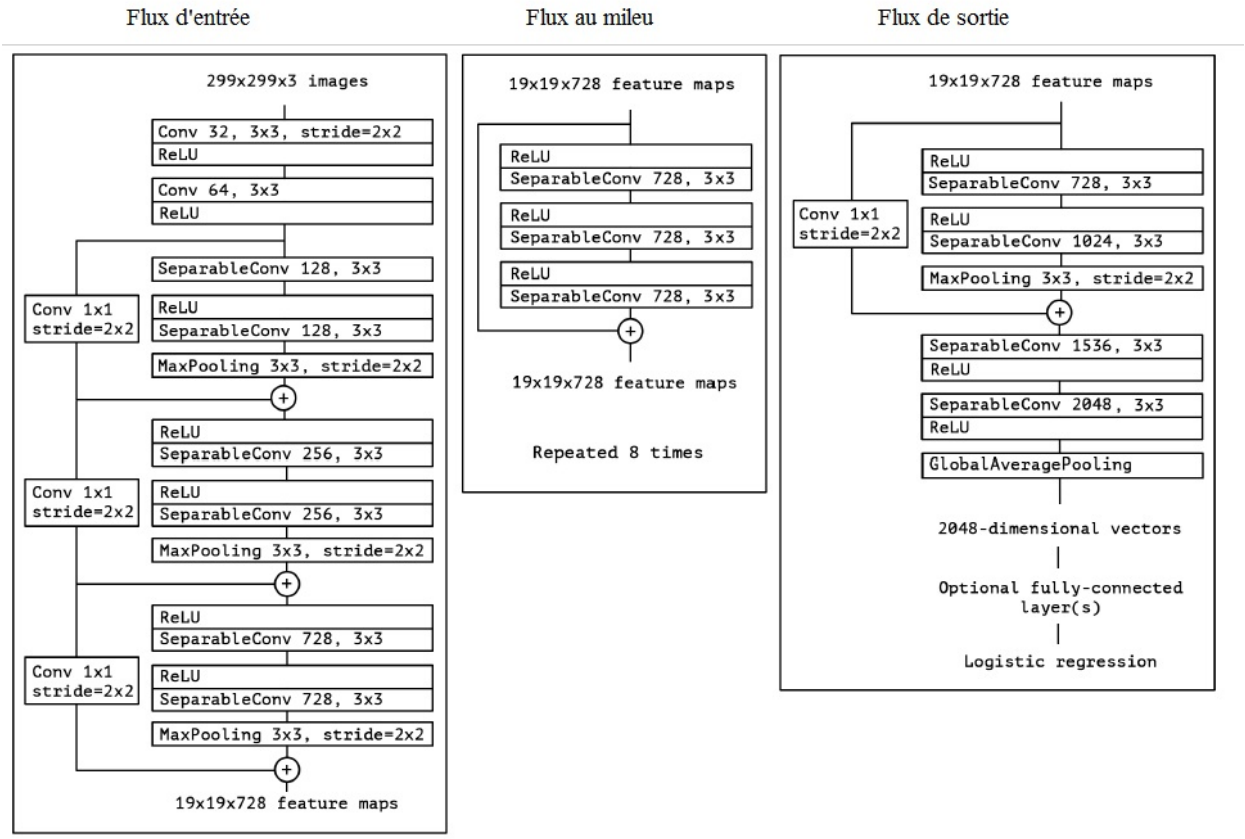

FIGURE 2.26 – *La structure du réseau Xception [Chollet 2017].*

L'étude comparative entre les réseaux Xception et InceptionV3 sur la

base d'apprentissage ImageNet a montré l'efficacité du réseau Xception. Cela indique les avantages des modules extrêmes d'Inception sur la performance, car ces deux architectures ont le même nombre de paramètres, et donc ces résultats ne sont pas liés à l'augmentation du nombre de paramètres, mais plutôt à la structure optimisée des blocs. En plus, ce réseau était plus performant par rapport aux différentes versions du réseau ResNet : ResNet-50, ResNet-101, et ResNet-152. Enfin, l'étude expérimentale sur les réseaux Xception avec et sans connexions résiduelles a démontré l'efficacité de ces liens dans l'accélération de l'apprentissage et l'amélioration de la performance.

DenseNet

DenseNet [Huang *et al.* 2017] est un réseau de neurones convolutif basé sur des connexions denses entre les couches de convolution. Selon la figure 2.27, ce réseau est composé d'un ensemble de blocs denses qui sont liés par des couches de transition. Chaque bloc contient un ensemble de couches de convolution, où chaque couche est connectée à toutes les autres couches suivantes appartenant au même bloc. Cela introduit $\frac{L(L+1)}{2}$ connexions dans un bloc contenant L couches au total.

Contrairement aux CNN classiques, chaque couche reçoit L entrées qui représentent les cartes des caractéristiques des $L-1$ couches précédentes. Ces connexions établissent des liens directs entre le gradient de la fonction du coût et les entrées originales. En plus, elles permettent d'améliorer la régularisation et donc réduire le problème de sur-apprentissage et de dégradation de gradient.

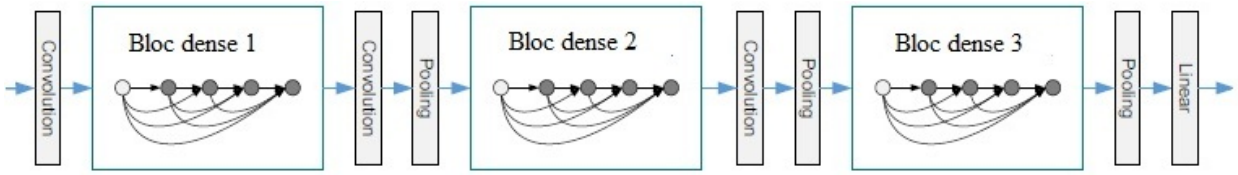

FIGURE 2.27 – *La structure du réseau DenseNet [Huang et al. 2017].*

Le nombre de couches dans un bloc dense dépend du taux de croissance k. Ce taux précise le nombre des cartes de caractéristiques en entrée et régularise la quantité d'informations ajoutée à chaque couche dans le réseau.

Afin de réduire le nombre total des paramètres, DenseNet utilise des couches de sous-échantillonnage et le taux de compression. Ces couches sont représentées par des couches de transitions et composées d'une couche de normalisation par lot, une convolution associée à un filtre de taille 1×1, et une couche de $\text{Avg}-\text{pooling}$. Le taux de transition $\theta \in \{0, 1\}$ permet de réduire le nombre des cartes de caractéristiques résultantes d'un bloc dense.

La figure 2.28 illustre la structure des différentes configurations du réseau DenseNet. Ces configurations varient dans la profondeur du réseau (de 121 à 201). Chaque bloc est composé d'un ensemble de couches de convolution, où chacune est associée à des filtres de taille 1×1 et 3×3.

Layers	Output Size	DenseNet-121($k = 32$)	DenseNet-169($k = 32$)	DenseNet-201($k = 32$)	DenseNet-161($k = 48$)
Convolution	112×112	7×7 conv, stride 2			
Pooling	56×56	3×3 max pool, stride 2			
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56×56	1×1 conv			
	28×28	2×2 average pool, stride 2			
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28×28	1×1 conv			
	14×14	2×2 average pool, stride 2			
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 36$
Transition Layer (3)	14×14	1×1 conv			
	7×7	2×2 average pool, stride 2			
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Classification Layer	1×1	7×7 global average pool			
		1000D fully-connected, softmax			

FIGURE 2.28 – *Les configurations du réseau DenseNet [Huang et al. 2017].*

L'étude comparative entre les résultats des réseaux DenseNet et ResNet a montré leur équivalence. Cela indique l'efficacité de DenseNet grâce à son nombre réduit de paramètres par rapport à ResNet. Par exemple, le réseau DenseNet-201 composé de 20 millions de paramètres a les mêmes performances qu'un réseau ResNet-101 composé de 40 millions de paramètres [Huang *et al.* 2017].

En résumé, le réseau DenseNet a plusieurs avantages liés à son efficacité dans l'extraction des caractéristiques, la réduction du nombre des paramètres, et la diminution des problèmes de dégradation de gradient.

MobileNetV1

MobileNetV1 [Howard *et al.* 2017] est un réseau de neurones convolutif conçu pour les appareils mobiles et les systèmes de vision intégrés. Ce réseau est basé sur les convolutions séparables en profondeur (DSC) [Sifre 2014]. Le but principal de ces modules est de réduire la dimensionnalité du réseau.

La figure 2.29 illustre la différence entre les filtres dans une convolution standard et dans un module DSC, où DSC factorise une convolution en depthwise convolution (DC) et pointwise convolution (PC). Contrairement à une convolution standard, dans un DSC, un seul filtre est appliqué sur chaque carte de caractéristiques en entrée. Ensuite, ces cartes sont combinées par la pointwise convolution qui est basée sur des filtres de taille 1×1. Cette factorisation réduit le nombre de paramètres par un facteur de $\frac{1}{N} + \frac{1}{D_k^2}$, où D_k et N sont la taille et le nombre des filtres, respectivement.

Afin d'adapter le réseau aux restrictions des systèmes de vision intégrés, le réseau MobileNet introduit deux hyperparamétres pour réaliser un compromis entre la période de réponse, l'espace de stockage, et la précision. Ces hyperparamétres sont : le multiplicateur de largeur et le multiplicateur de résolution.

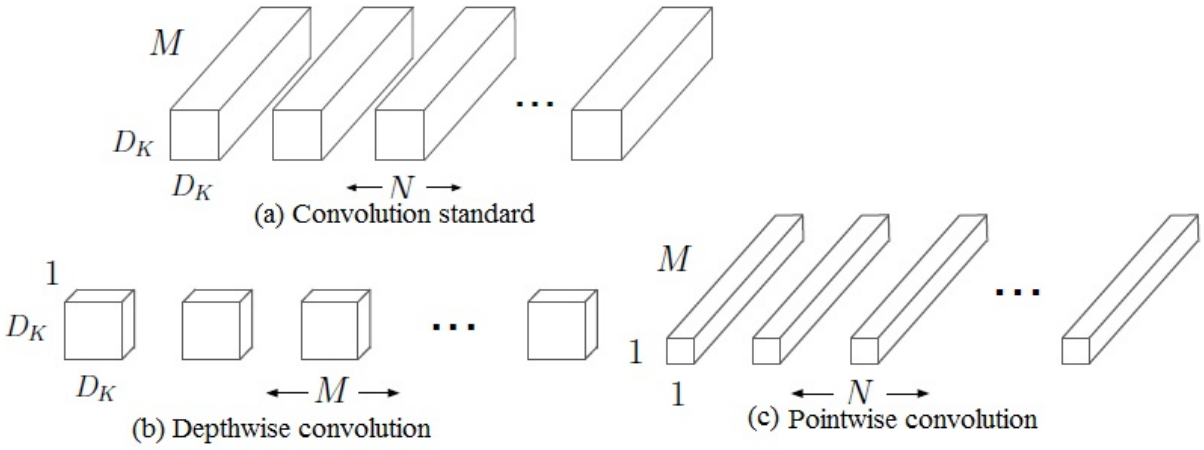

FIGURE 2.29 – *La différence entre les filtres d'une convolution standard et une convolution séparable en profondeur [Howard et al. 2017].*

Le rôle du multiplicateur de largeur ($\alpha \in [0,1]$) est de réduire le nombre des cartes des caractéristiques dans les couches de convolution de M à $\alpha \times M$. Tandis que le multiplicateur de résolution réduit le coût de calcule par la réduction de la résolution de l'image en entrée, et cela diminue automatiquement la taille des couches de convolution.

La figure 2.30 illustre la structure du réseau MobileNetV1. Ce réseau est composé de 28 couches au total représentées par : des couches de convolutions standard (Conv), des convolutions séparables en profondeur (Conv dw) et des couches entièrement connectées (FC). Tous les couches sont suivies par une opération de normalisation par lots et la fonction d'activation ReLu.

L'étude expérimentale sur la base d'apprentissage ImageNet a montré que les DSC et les multiplicateurs de largeur et de résolution réduisent la précision du réseau MobileNetV1 en diminuant en parallèle les complexités temporelles et spatiales.

L'étude comparative entre les résultats de MobileNetV1 et des réseaux Inception, VGG16, et AlexNet a montré que MobileNetV1 est plus performant par rapport au réseau Inception et AlexNet en réduisant la complexité de calcul par 2.5 et 9.4 fois, respectivement. Tandis que VGG16 est plus performant par rapport à MobileNet, mais plus exigeant 27 fois en capacité de calcul.

MobileNetV2

MobileNetV2 [Sandler *et al.* 2018] est une version modifiée du réseau MobileNetV1 [Howard *et al.* 2017] qui a été proposée pour améliorer les systèmes de vision intégrés. Ce réseau est basé sur des modules de type inverted residual with linear bottleneck (IRLB). Ces modules sont construits à base des DSC introduits dans MobileNetV1 [Howard *et al.* 2017] et les blocs résiduels utilisés dans RestNet [He *et al.* 2016].

La figure 2.31 illustre la différence entre les modules DSC et IRLB. Un module IRLB est constitué de trois opérations : expansion, DC, et projection. Premièrement, un facteur d'expansion ($F > 1$) est utilisé pour étendre

Type / Pas	Taille du filtre	Taille d'entrée
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classificateur	$1 \times 1 \times 1000$

FIGURE 2.30 – *La structure du réseau MobileNetV1 [Howard et al. 2017].*

la profondeur de la couche de convolution de N à FN. Ensuite, le résultat obtenu est passé à DC, et enfin la profondeur initiale est restaurée par l'opération de projection.

Les opérations d'expansion et de projection sont basées sur des convolutions associées a des filtres de taille 1×1. Dans les réseaux MobileNet, les DC sont des outils de réduction de dimensionnalité, mais malgré leurs efficacités dans l'optimisation de la complexité de calcul, ils peuvent causer une perte d'informations. Pour résoudre ce problème, le réseau MobileNetV2 propose l'utilisation des modules d'expansion pour restaurer cette information, ensuite l'information résultante est compressée par les modules de projection pour restaurer la taille initiale.

Le réseau MobileNetV2 est composé d'une couche de convolution standard suivie par 19 blocs de type IRLB, et il est basé sur les multiplicateurs de largeur et de résolution pour ajuster sa taille selon les restrictions des systèmes exploités.

L'étude comparative entre les performances des réseaux MobileNetV1 [Howard *et al.* 2017], ShuffleNet [Zhang *et al.* 2018], et NASNet-A [Zoph *et al.* 2018] a montré son efficacité en classification, segmentation et détection.

ShuffleNet

ShuffleNet [Zhang *et al.* 2018] est un réseau de neurones convolutif conçu pour les appareils mobiles qui ont une puissance de calcul limi-

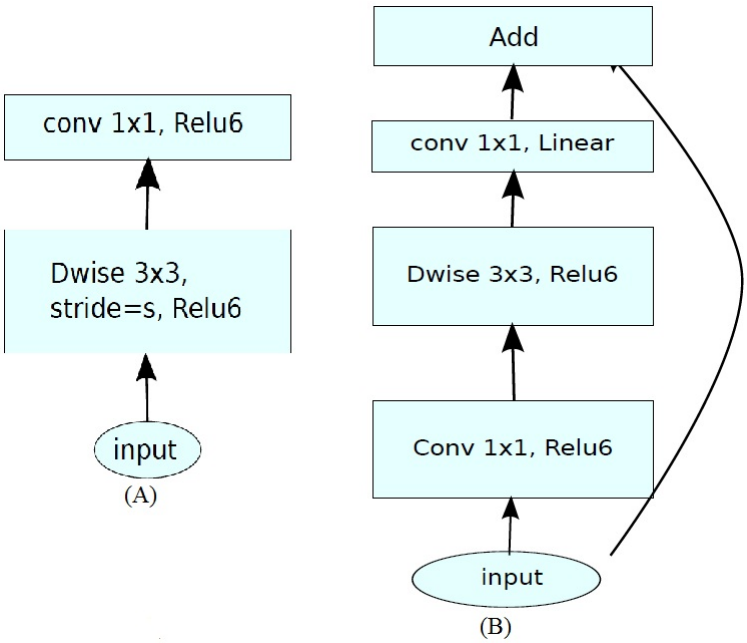

FIGURE 2.31 – *La différence entre (A) depthwise separable convolutions et (B) inverted residual with linear bottleneck [Sandler et al. 2018].*

tée. Ce réseau est basé sur les convolutions groupées et l'opération de mélange des canaux.

Dans une couche de convolution, une convolution groupée est définie par le processus de séparation des canaux en un certain nombre g de groupes. Le but de cette séparation est de réduire le nombre des paramètres et d'optimiser la complexité de calcul. Cette notion a été introduite précédemment dans les réseaux AlexNet [Krizhevsky *et al.* 2012] et MobileNet [Howard *et al.* 2017, Sandler *et al.* 2018].

En raison des restrictions de mémoire, AlexNet divise chaque couche de convolution en deux groupes afin de distribuer le calcul sur deux GPUs. Tandis que MobileNet divise chaque couche en groupes composés d'un seul canal, ensuite les canaux résultants sont regroupés par la pointwise convolution. Ce processus permet de réduire le nombre des paramètres par rapport à une convolution standard associée à des filtres de taille 3x3.

La figure 2.32 illustre la différence entre les modules de base dans les réseaux MobileNet et ShuffleNet respectivement. Une unité ShuffleNet remplace les 1x1 convolutions dans MobileNet par des convolutions groupées (GConv) et ajoute une opération de mélange des canaux après chaque 3x3 Depthwise convolution (DWConv). Le mélange des canaux permet de recevoir des informations en entrées de différents groupes, et cela génère et encode plus d'informations. Ce mélange peut être effectué avant ou après l'opération de convolution. Dans le premier cas, les données en entrées sont obtenues de différents groupes. Tandis que dans le deuxième cas, les canaux dans chaque groupe sont divisés en sous groupes, ensuite chaque groupe dans la couche suivante reçoit différents sous groupes de la couche précédente.

L'étude expérimentale sur les bases d'apprentissage ImageNet et

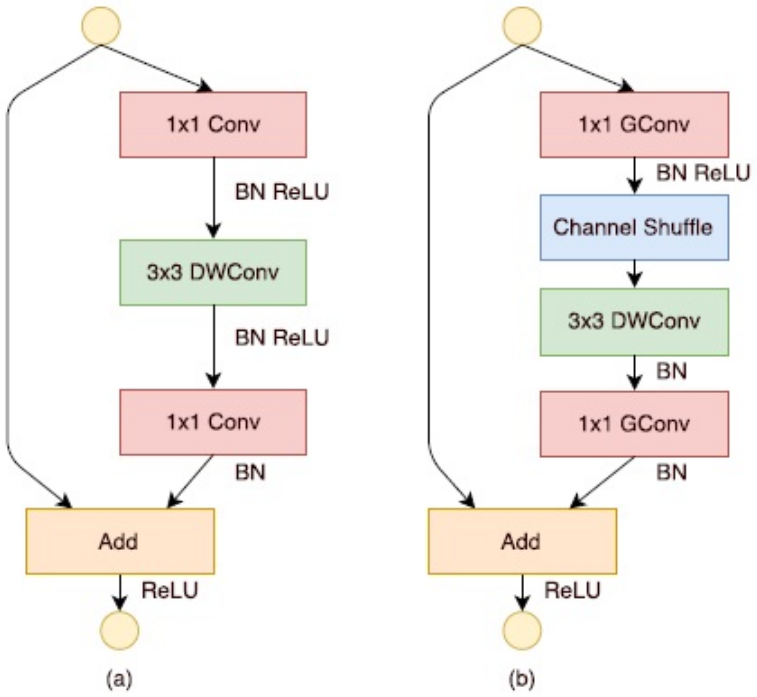

FIGURE 2.32 – *La différence entre les modules de base des réseaux (a) MobileNet et (b) ShuffleNet [Zhang et al. 2018].*

MSCOCO a montré l'efficacité de ShuffleNet par rapport à MobileNet. En plus, ce réseau est 13 fois plus rapide par rapport à AlexNet tout en conservant des résultats compétitives.

SqueezeNet

SqueezeNet [Iandola *et al.* 2016] est un réseau de neurones convolutif qui propose des techniques de réduction de dimensionnalité tout en maintenant une précision compétitive. Son architecture est basée sur un ensemble d'unités définies par des fire-modules (figure 2.33). Ces modules sont composés de deux couches : squeeze et expand.

La couche squeeze utilise des convolutions associées à des filtres de taille 1×1. Le but de cette couche est de réduire la dimensionnalité, car les filtres de taille 1×1 réduisent 9 fois le nombre de paramètres par rapport aux filtres de taille 3×3. En plus, cette étape permet de diminuer la profondeur des cartes des caractéristiques en entrée, et donc réduire le nombre des filtres de taille 3×3 de la couche expand. La couche squeeze génère des cartes de caractéristiques qui ont une taille identique aux cartes en entrée. Afin de réduire cette taille et d'augmenter leurs profondeurs, la couche expand utilise des filtres de taille 3×3 pour capturer l'information spatiale.

La figure 2.34 illustre l'architecture de SqueezeNet. Ce réseau est composé de 2 convolutions standards (conv1, conv10), 8 fire-modules, et des couches de Max-pooling après les couches conv1, fire4, fire8, et conv10.

L'étude expérimentale sur la base d'apprentissage ImageNet a montré que SqueezeNet a atteint la précision de AlexNet en réduisant 50 fois le

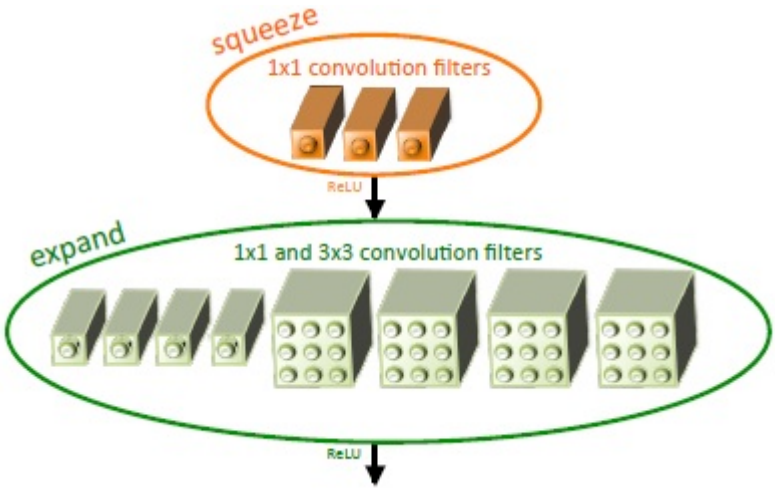

FIGURE 2.33 – *La structure d'un fire-module [Iandola et al. 2016].*

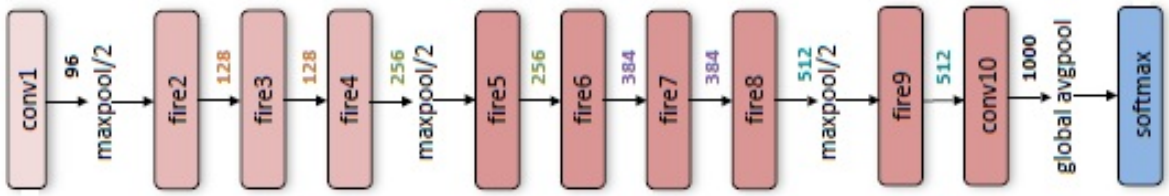

FIGURE 2.34 – *La structure du réseau SqueezeNet [Iandola et al. 2016].*

nombre de paramètres.

WideResNet

WideResNet [Zagoruyko & Komodakis 2016] est une version optimisée du réseau ResNet [He *et al.* 2016]. Les réseaux ResNet proposent l'utilisation des blocs boottleneck afin de concevoir des architectures profondes en réduisant en parallèle le nombre des paramètres. Cette stratégie diminue le partage des informations entre les blocs dans les réseaux profonds à cause de la taille réduite des filtres, et cela diminue la réutilisation des attributs dans le réseau ResNet [Srivastava *et al.* 2015]. Afin de résoudre ce problème, WideResNet suggère que l'élargissement des blocs ResNet est une méthode efficace pour améliorer les performances par rapport à l'augmentation de la profondeur.

La figure 2.35 illustre la différence entre les blocs ResNet (a, b) et WideResNet (c, d). Le bloc wide-dropout propose l'exploitation de la régularisation par abondon (Dropout) entre les couches de convolutions pour réduire le risque de sur-apprentissage.

WideResNet est basé sur deux hyper-paramétres : facteur d'approfondissement (l) et d'élargissement (k), où l est le nombre de convolutions dans un bloc et le paramètre k augmente le nombre de paramètres dans une couche de convolution. Le bloc basic-wide correspond à l = 2 et k = 1.

Le réseau WideResNet est composé d'une couche de convolution stan-

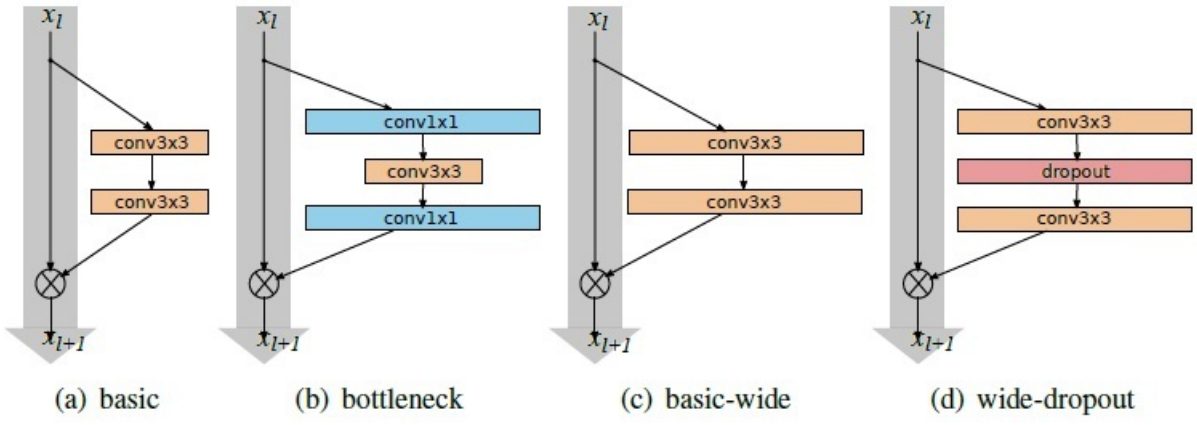

FIGURE 2.35 – *La différence entre les blocs ResNet (a, b) et WideResNet (c, d) [Zagoruyko & Komodakis 2016].*

dard suivie par 3 groupes de blocs résiduels, une couche Avg-pooling, et une couche de classification.

L'étude expérimentale sur les bases d'apprentissage CIFAR−10, CIFAR−100, SVHN et ImageNet a montré que deux blocs par convolution ($l = 2$) est le choix optimal. Tandis que l'élargissement ($k > 1$) améliore les performances des réseaux résiduels de différentes profondeurs.

Les résultats obtenus ont démontré qu'un réseau WideResNet composé de 50 couches est plus performant qu'un réseau ResNet composé de 152 couches sur la base ImageNet. Cela prouve l'efficacité des blocs Wide dans la réduction de dimensionnalité, l'amélioration des performances et l'accélération du temps de calcul par rapport aux blocs résiduels standard.

Discussion et comparaison

Le but de cette section est de comparer entre les résultats des réseaux profonds détaillés précédemment.

La majorité des architectures ont été soumises à la compétition ImageNet pour valider les résultats obtenus. AlexNet est le premier réseau de type DL qui a atteint un taux d'erreur intéressant (15.3%) par rapport aux résultats des méthodes ML classiques [Berg *et al.* 2010, Sánchez & Perronnin 2011].

En 2013, la variante ZFNet [Zeiler & Fergus 2014] du réseau AlexNet a atteint un taux d'erreur de 14.8%. En 2014, ce taux d'erreur a diminué à 6.8% en se basant sur l'architecture Inception [Szegedy *et al.* 2015]. Ce réseau propose une structure qui réduit de 12 fois le nombre de paramètres du réseau AlexNet. Dans la même année, le réseau VGGNET a atteint un taux d'erreur de 7.3%. En 2015, le réseau profond ResNet [He *et al.* 2016] composé de 152 couches a réduit le taux d'erreur à 3.57%. En 2016, le réseau Inception-ResNetV2 qui propose une hybridation entre les blocs d'Inception et les liens résiduels a atteint un taux de 3.7%.

Tableau 2.1 compare entre ces architectures en termes de profondeur, nombre de paramètres, et la précision.

Le but de ces architectures est de proposer des réseaux performants en réduisant la capacité de stockage et le temps de calcul. L'optimisation de la

Architecture	Profondeur	Paramètres (Million)	Top 1 (Accuracy)	Top 5 (Accuracy)
AlexNet [Krizhevsky *et al.* 2012]	8	60	63.3%	84.6%
ZFNet [Zeiler & Fergus 2014]	8	-	64%	85.3%
Inception [Szegedy *et al.* 2015]	22	6.8	69.8%	89.9%
MobileNet[Howard *et al.* 2017]	28	4.2	70.6%	89.5%
ShuffleNet[Zhang *et al.* 2018]	50	-	70.9%	89.8%
VGG-19 [Simonyan & Zisserman 2014b]	19	144	74.5%	92.0%
MobileNetV2 [Sandler *et al.* 2018]	20	3.4	74.7%	-
InceptionV2 [Szegedy *et al.* 2016]	-	11.2	74.8%	92.2%
ResNet-152 [He *et al.* 2016]	152	21.8 à 60.2	78.57%	94.29%
InceptionV3 [Szegedy *et al.* 2016]	-	23.8	78.8%	94.4%
Xception [Chollet 2017]	71	22.8	79%	94.5%
Inception-ResNetV2 [Szegedy *et al.* 2017]	-	55.8	80.1%	95.1%

TABLE 2.1 – *La comparaison entre les architectures CNN en termes de profondeur, nombre de paramètres, et précision.*

structure des couches de convolution classiques était l'une des stratégies principales pour réaliser ce traitement. Par exemple, le réseau profond Inception réduit le nombre des paramètres de AlexNet de 60 à 6.8 millions et améliore la performance de 63.3 % à 69.8 %. Cette réduction est réalisée par les blocs d'Inceptions à travers des techniques de réduction de dimensionnalité.

Les résultats obtenus indiquent aussi l'efficacité de MobileNetV2, où il a atteint un taux de 74.7 % avec seulement 3.4 millions de paramètres. Comme nous l'avons mentionné précédemment, les réseaux profonds améliorent la non linéarité à travers les fonctions d'activation (ReLu) supplémentaires. Cela permet d'améliorer la performance dans la majorité des cas sur les grands volumes de données. Cette stratégie a été exploitée par le réseau ResNet-152, où il augmente la profondeur de AlexNet de 8 à 152 tout en maintenant le même nombre de paramètres (60 millions). Les techniques utilisées dans ResNet-152 ont amélioré la performance à 78.57 %, cela renforce les hypothèses sur la puissance des réseaux profonds. Enfin, la performance des autres variantes de Inception (InceptionV3, Xception, et Inception-ResNetV2) varie de 78.8 % et 80.1 % avec un nombre de paramètres réduit par rapport à ResNet-152.

La figure 2.36 présente une vue plus informative sur les performances par rapport au tableau précédent. Elle indique que le réseau VGGNet est le plus exigeant en termes de stockage et capacité de calcul. Les architectures ResNet et Inception forment une ligne droite et s'organisent selon leur profondeur. Généralement dans la même catégorie, les réseaux les plus profonds sont les plus performants.

2.5.2 Détection des objets

Régions avec réseaux de neurones convolutif (R-CNN)

Régions avec réseaux de neurones convolutif (Regions with CNN features (R−CNN)) [Girshick *et al.* 2014] est une architecture d'apprentissage profond conçue pour la détection des objets. Cette architecture combine entre les méthodes de proposition des régions (regions proposal) et les CNN.

La figure 2.37 illustre sa structure qui est composée de 3 modules :

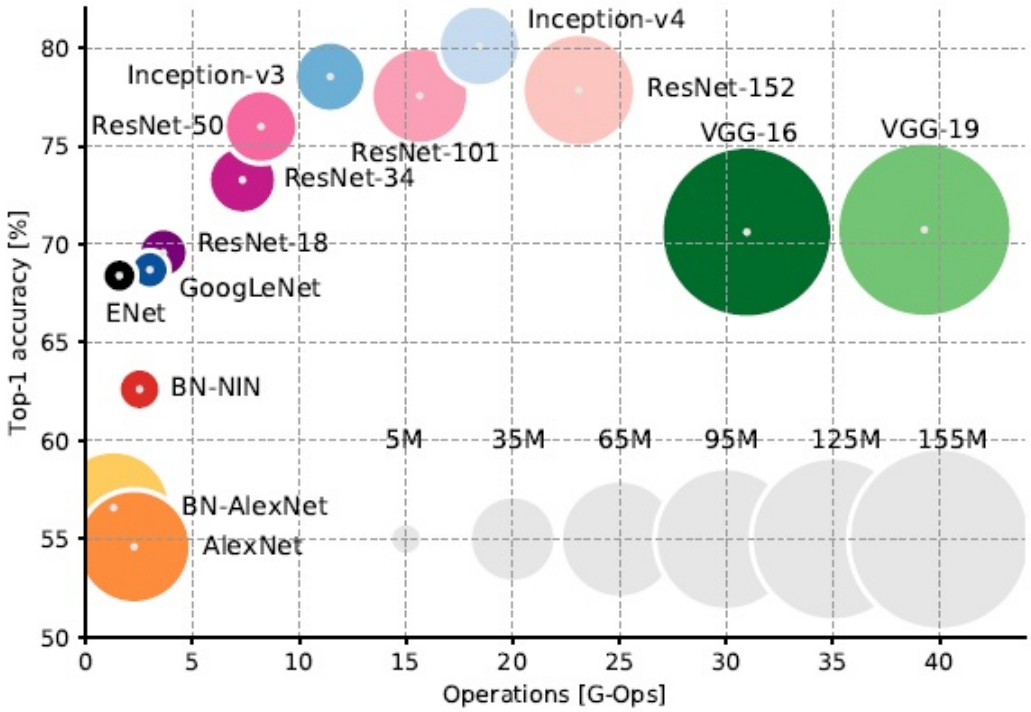

FIGURE 2.36 – *La représentation des architectures en termes de Top-1 accuracy, profondeur, nombre d'opérations, et nombre de paramètres [Canziani et al. 2016].*

extraction des régions d'intérêt, extraction des caractéristiques, et la classification. Le premier module est basé sur la méthode de recherche sélective (selective search) [Uijlings *et al.* 2013] qui propose 2000 régions indépendantes en entrée. Ensuite, ces régions sont ajustées pour obtenir une dimensionnalité conforme au réseau CNN.

R−CNN exploite le modèle Alexnet [Krizhevsky *et al.* 2012] préentraîné précédemment sur la base ImageNet, et suivi par un fine-tuning sur la base d'apprentissage cible PASCAL. Le but de ce transfert est d'éviter le problème de sur-apprentissage sur les volumes de données limitées de la base PASCAL. Cette étape permet d'extraire 4096 attributs à partir de chaque région. Enfin, ces vecteurs sont fournis à l'algorithme SVM pour la tâche de classification et à l'algorithme bounding-box regressors pour ajuster le cadre de sélection de la région proposée.

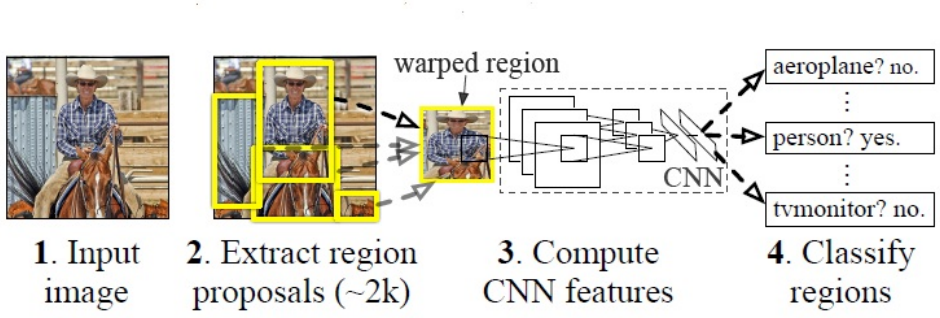

FIGURE 2.37 – *La structure de Régions avec réseaux de neurones convolutif (R-CNN) [Girshick et al. 2014].*

L'étude expérimentale sur la base d'apprentissage PASCAL VOC a montré que R-CNN a amélioré l'erreur moyenne (MAP) par 30% par rapport aux résultats obtenus précédemment. Malgré ces performances, R-CNN ne peut pas être exploité dans les applications en temps réel à cause du temps élevé pour le traitement des 2000 régions d'intérêts (47 s/image).

Fast R-CNN

Fast R-CNN [Girshick 2015] est une version optimisée de l'architecture R-CNN [Girshick *et al.* 2014]. Son but principal est d'accélérer le temps d'apprentissage et de test de R-CNN.

Comme nous l'avons détaillé précédemment, R-CNN effectue la tâche d'extraction des caractéristiques pour chaque région proposée. Cela exige de réaliser 2000 passes dans le réseau CNN, ce qui ralentit le temps de test. Pour résoudre ce problème, Fast R-CNN prend en entrée l'image entière et les cordonnées des régions d'intérêt, donc une seule passe est effectué pour chaque image au lieu de 2000.

La figure 2.38 illustre la structure du réseau Fast R-CNN. Ce réseau passe l'image en entrée au réseau CNN pour générer en sortie des cartes de caractéristiques. Ensuite, les régions proposées sont identifiées dans ces cartes et redimensionnées par la couche RoI-pooling. Les vecteurs en sortie sont passés ensuite aux couches entièrement connectées. Enfin, les sorties sont utilisées pour prédire la classe par le classificateur softmax et pour réajuster le cadre de sélection par le bounding-box regressor.

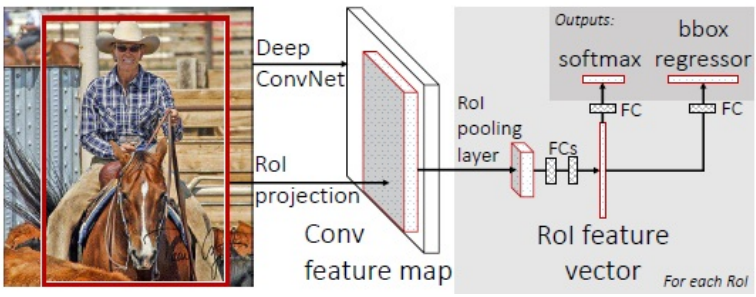

FIGURE 2.38 – *La structure du réseau Fast R-CNN [Girshick 2015].*

Comme R-CNN, Fast R-CNN utilise des modèles pré-entrainés sur la base d'apprentissage ImageNet et la technique de fine-tuning. Afin d'adapter ces modèles sur l'architecture Fast R-CNN, la dernière couche de max-pooling est remplacée par une couche de RoI-pooling et la dernière couche entièrement connectée et softmax sont remplacées par deux autres couches de type FC. Cette structure montre que l'apprentissage dans un Fast R-CNN est effectué dans une seule étape au lieu de trois étapes séparées (SVM, softmax et l'algorithme de régression). Tous ces facteurs ont permis d'accélérer le temps d'apprentissage et de test de R-CNN et d'améliorer sa performance.

L'étude expérimentale à base du modèle VGG16 sur la base d'apprentissage PASCAL VOC 2012 a montré que Fast R-CNN est 9 fois plus rapide que R-CNN en apprentissage et 213 fois en test.

Faster R-CNN

Faster R-CNN [Ren *et al.* 2015] est une version optimisée du réseau Fast R-CNN. Contrairement aux méthodes expliquées précédemment, cet algorithme élimine la recherche sélective et intègre le processus de sélection des régions d'intérêt à l'intérieur du réseau. Cette stratégie automatise la tâche de sélection des régions d'intérêt et accélère le temps de traitement.

Premièrement, Faster R-CNN passe l'image en entrée aux couches de convolution pour générer en sortie des cartes de caractéristiques. Ensuite, Region proposal network (RPN) génère les cadres de sélection des régions d'intérêt à partir de ces cartes de caractéristiques (Figure 2.39). Les régions proposées par RPN sont redimensionnées ensuite par la couche de RoI-pooling. Enfin, les vecteurs en sortie sont passés aux couches entièrement connectées pour classifier l'objet et optimiser les cadres de sélection.

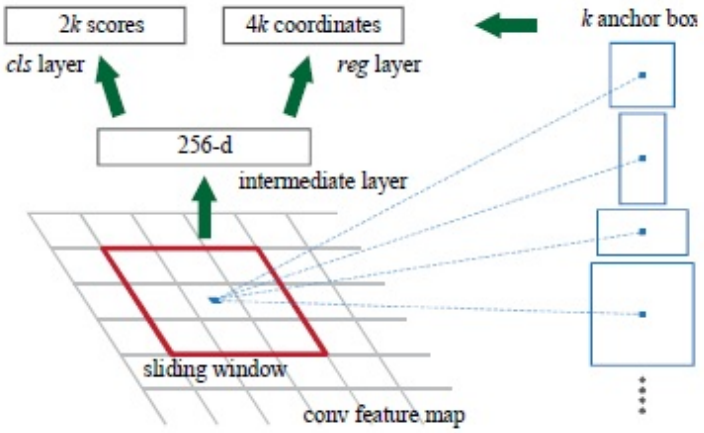

FIGURE 2.39 – *le processus d'un Region Proposal Network (RPN) [Ren et al. 2015].*

YOLO

You Only Look Once (YOLO) [Redmon *et al.* 2016] est un réseau de neurones convolutif conçu pour la détection des objets. Contrairement aux méthodes expliquées précédemment, cette architecture traite le problème de détection des objets comme un problème de régression pour prédire en parallèle les classes et les cadres de sélection des objets.

Le processus d'apprentissage et de prédiction à partir de l'image entière permet au réseau d'encoder l'information contextuelle, et donc réduire le taux des faux positifs. En plus, l'utilisation d'un seul réseau pour la détection a accéléré 1000 fois le traitement par rapport à R-CNN et 100 fois par rapport à Fast R-CNN, où il permet de traiter 25 cadres par second.

La figure 2.40 illustre le processus de détection effectué par YOLO. Premièrement, ce réseau divise l'image en entrée en $S \times S$ grilles. Chaque grille prédis B cadres de sélection et les probabilités d'appartenance de l'objet aux différentes classes C. Un cadre de sélection est caractérisé par 5 paramètres : 4 cordonnées (x, y, h, w) et un score de confiance P. Ce

score représente la probabilité d'appartenance d'un objet à ce cadre et sa précision. Pour résumer, chaque image en entrée est associée à une prédiction encodée sous forme d'un tenseur 3D de taille $S \times S \times (5B + C)$.

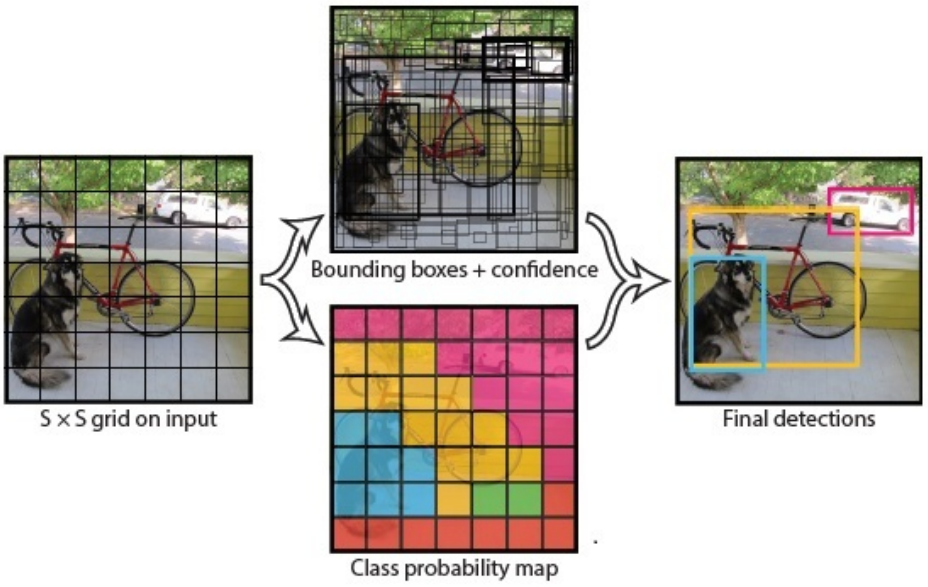

FIGURE 2.40 – *Le processus de détection des objets par YOLO [Redmon et al. 2016].*

L'architecture de YOLO est inspirée du réseau Inception [Szegedy *et al.* 2015]. Cette architecture est composée de 24 couches de convolution et 2 couches entièrement connectées. En apprentissage, les 20 premières couches sont initialisées par un prétraitement sur la base d'apprentissage ImageNet, tandis que les couches restantes sont initialisées aléatoirement.

Malgré l'efficacité de YOLO et son traitement accéléré par rapport aux systèmes introduits précédemment, ce réseau a quelques inconvénients liés à des contraintes spatiales comme : le nombre des cadres de sélection dans une grille. Cela limite la capacité de YOLO dans la détection des petits objets organisés en groupes. En plus, ce réseau a une erreur de localisation élevée par rapport aux autres systèmes basés sur les régions d'intérêts. Afin de résoudre ces problèmes, d'autres versions de YOLO ont été proposées : YOLOv2, YOLOv3 [Redmon & Farhadi 2018], et YOLO9000 [Redmon & Farhadi 2017].

Comparison et discussion

Les structures détaillées précédemment ont été évaluées sur les benchmarks de détection d'objets : PASCAL VOC 2007 [Everingham *et al.*] et PASCAL VOC 2012. La base d'apprentissage PASCAL Visual Object Classification (PASCAL VOC) est composée de 20 classes, y compris les humains, animaux, véhicules et les objets d'intérieur. Cette base d'apprentissage contient environ 10 000 images pour l'apprentissage et la validation, et elle a été exploitée dans 8 défis dans la période 2005-2012, où chaque défi avait ses propres spécificités.

Architecture	PASCAL VOC 2007	PASCAL VOC 2012
R-CNN	58.5%	53.3%
Fast R-CNN	70%	66%
Faster R-CNN	73.2%	70.4%
YOLO	63.4%	57.9%

TABLE 2.2 – *Comparaison entre les résultats des benchmarks VOC 2007 et VOC 2012 en terme de MAP.*

Le tableau 2.2 présente les résultats obtenus sur les benchmarks VOC 2007 et VOC 2012 en terme de AAP. Ces résultats valident l'efficacité de Faster R-CNN par rapport à Fast R-CNN, R-CNN, et YOLO. Tandis que la structure YOLO est la plus adaptée aux applications en temps réel grâce à son traitement rapide par rapport aux autres structures.

2.5.3 Segmentation sémantique

Les réseaux de neurones convolutif

La segmentation sémantique consiste à attribuer une classe à chaque pixel appartenant à l'image en entrée. Dans ce genre d'applications, les CNN sont utilisés comme des classificateurs de pixel. Le réseau prend en entrée des segments de l'image et classifie son centre. Cette opération est répétée pour tous les pixels de l'image, où chaque pixel est considéré comme un centre du segment proposé. L'inconvénient principal de cette méthode est son temps de traitement très élevé à cause des classifications denses de tous les pixels en entrée. Cela limite leur exploitation dans les applications en temps réel. En plus, les patchs en entrées des pixels voisins se chevauchent, et donc les mêmes convolutions sont calculées plusieurs fois.

Fully convolutional networks

Contrairement à un CNN, le réseau FCN [Long *et al.* 2015] effectue la phase de segmentation en un seul passage (figure 2.41). Ce réseau est composé de deux parties principales : sous-échantillonnage et un rééchantillonnage. Le sous-échantillonnage permet de capturer les informations sémantiques et contextuelles. Ensuite, le rééchantillonnage rétablit les informations spatiales.

Le réseau FCN peut gérer les entrées de taille variante. Cela est réalisé par l'élimination des couches entièrement connectées, car elles sont liées à la taille fixe en entrée. Ce réseau remplace les couches entièrement connectées par des couches de convolution pour produire des cartes spatiales. Ensuite, ces cartes sont passées aux couches de déconvolution [Zeiler *et al.* 2011] pour restaurer la taille en entrée et produire des sorties classifiées par pixel.

FCN est caractérisé par son processus d'apprentissage de bout en bout par rapport aux méthodes de segmentation sémantique par région. Ce réseau a atteint de bonnes performances en segmentation par rapport aux autres méthodes classiques sur la base d'apprentissage PASCAL VOC. Malgré son efficacité, FCN est caractérisé par certaines limitations liées

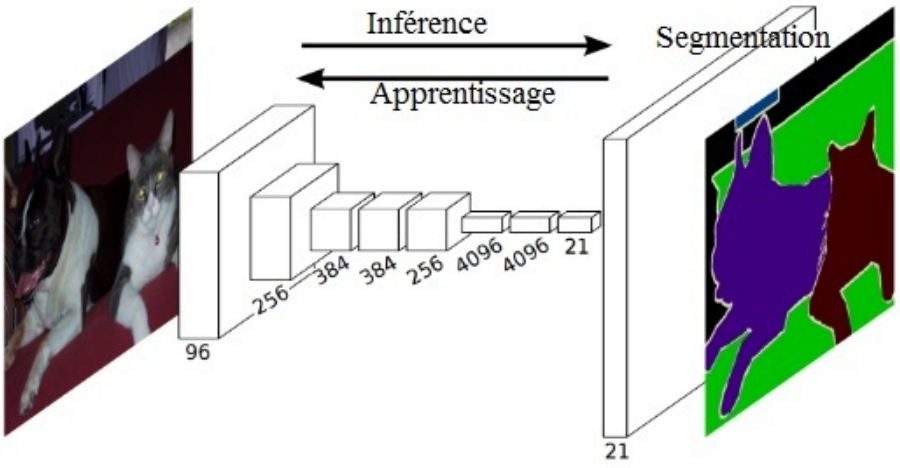

FIGURE 2.41 – *La structure du réseau fully convolutional network [Long et al. 2015].*

à l'invariance spatiale et le manque d'information contextuelle. En plus, la résolution de l'image en entrée diminue à cause de son passage par une succession de couches de convolution et de pooling.

En plus des architectures citées précédemment, il existe d'autres architectures connues en segmentation comme : DeepLab [Chen *et al.* 2015, Chen *et al.* 2017b], SegNet [He *et al.* 2017], U-net [Ronneberger *et al.* 2015], et Mask R-CNN [He *et al.* 2017].

2.6 CONCLUSION

Dans ce chapitre, nous avons détaillé la structure générale d'un réseau CNN et analysé quelques architectures connues en classification, en détection des objets, et en segmentation.

En classification, les architectures soumises à la compétition ImageNet ont connu un grand intérêt grâce à leurs performances remarquables. Le but principal de ces architectures était de proposer des variantes profondes en réduisant en parallèle le nombre total des paramètres. Ces variantes sont caractérisées par de nouveaux blocs de base comme les blocs d'Inception dans le réseau Inception, et les DSC dans le réseau MobileNet. En plus d'autres variantes proposent d'hybrider entre différents blocs, comme le réseau Inception-ResNet qui combine entre les blocs d'Inception et les liens résiduels. En détection et en segmentation, l'objectif était de proposer des structures qui permettent de détecter et segmenter les objets en temps réel comme les réseaux YOLO et FCN. YOLO accélère le temps de détection par l'intégration de la phase de génération des RoI dans le réseau, et FCN effectue la segmentation dans la phase de prédiction en un seul passage.

Le chapitre suivant présente les domaines d'application des architectures définies dans ce chapitre en classification, en détection, et en segmentation. Dans ce cadre, nous commencerons par une description générale de quelques domaines d'application. Ensuite, nous détaillerons le domaine de l'imagerie médicale qui est le domaine d'intérêt de cette thèse.

Les domaines d'application des réseaux de neurones convolutifs en vision par ordinateur et en imagerie médicale

3

Sommaire

Durant ces dernières années, l'apprentissage profond a été exploité dans plusieurs domaines. En vision par ordinateur, les CNN sont connus pour leur bonne précision dans la résolution des problèmes du monde réel. Le but de ce chapitre est de présenter un aperçu sur certains domaines d'application des CNN en vision par ordinateur, comme : la classification des images, la détection et la localisation des objets, la segmentation sémantique, la reconnaissance d'action et d'activité, l'estimation de la pose humaine, et la reconnaissance faciale. Nous avons aussi présenté quelques domaines d'application des CNN en traitement des images médicales à base des techniques de classification, détection, et segmentation. .

Mots clés : Vision par Ordinateur, Apprentissage profond, Réseau de neurones convolutif, Domaines d'application, Imagerie médicale.

3.1 Introduction

La vision par ordinateur est une branche de l'intelligence artificielle. Elle permet à un ordinateur d'analyser, de traiter, et de comprendre les images. Les systèmes de vision sont exploités pour extraire des informations pertinentes à partir des entrées visuelles (image ou vidéo) afin de les utiliser dans d'autres tâches de recommandation.

Les systèmes de prédiction en vision par ordinateur sont basés sur les algorithmes d'apprentissage automatique. Ils permettent d'analyser les entrées visuelles prises par un système d'acquisition. Ces algorithmes sont entrainés sur des données pour produire des modèles en sortie. Les modèles générés sont exploités ensuite dans la phase de prédiction.

Les méthodes d'apprentissage automatique classiques exigent une représentation formelle des données complexes (images, vidéo, ou texte). Cette représentation est réalisée dans la phase d'extraction des caractéristiques (les handcrafted features). L'inconvénient principal de ces approches est leur influence négative sur les résultats.

Contrairement aux méthodes ML classiques, les méthodes DL et particulièrement les CNN sont adaptés aux données complexes, car ils intègrent la phase d'extraction des caractéristiques dans le processus de l'apprentissage. Plusieurs facteurs, tel que la première implémentation GPU [Chellapilla *et al.* 2006] et la première application de Max-pooling [Karpathy *et al.* 2014] ont contribué à la popularité des CNN.

Les CNN sont composés d'un ensemble de couches de convolution et de pooling, qui sont regroupées en modules, et une ou plusieurs couches entièrement connectées. Ces modules sont empilés pour former un réseau d'apprentissage profond.

Ces dernières années, plusieurs architectures optimisées ont été proposées pour améliorer la précision de la classification et pour réduire le coût de calcul des CNN. Par conséquent, dans la catégorie des réseaux DL, les CNN sont devenus les algorithmes de base en vision par ordinateur. En raison de leur efficacité dans la gestion des grands volumes de données, les techniques d'apprentissage profond présentent des outils puissants dans le traitement et l'analyse des big data. Ces dernières années, des quantités massives de données ont été collectées dans divers domaines, y compris la cyber sécurité, l'informatique médicale, et les réseaux sociaux. Les algorithmes d'apprentissage profond sont utilisés pour extraire les caractéristiques de haut niveau de ces données afin obtenir des représentations hiérarchiques.

En raison du progrès important des méthodes de vision par ordinateur, ces techniques ont été exploitées dans plusieurs applications du monde réel comme les systèmes de surveillance [Muhammad *et al.* 2018], le domaine médical [Litjens *et al.* 2017], la robotique [Turan *et al.* 2018], et les voitures autonomes [Tian *et al.* 2018].

Le but de ce chapitre est de détailler certaines applications connues des réseaux de neurones convolutifs en vision par ordinateur comme : la classification des images, la détection et localisation des objets, la segmentation sémantique, la reconnaissance d'action et d'activité, l'estimation de la pose humaine, et la reconnaissance faciale (figure 3.1).

Les méthodes proposées dans cette thèse s'intéressent à la résolution

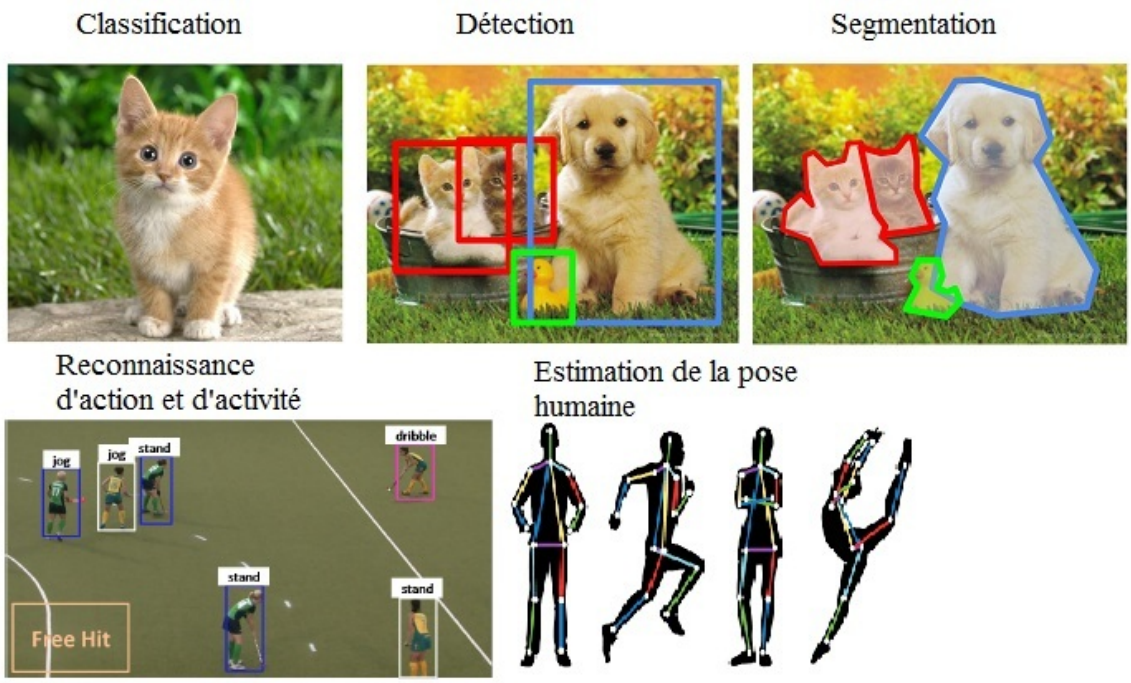

FIGURE 3.1 – *Les applications connues en vision par ordinateur.*

des problèmes liés aux images médicales et exactement aux images histologiques. Pour cela, nous avons présenté une section qui explique les différentes modularités d'images médicales et quelques méthodes proposées dans l'état de l'art pour le traitement de ces images.

3.2 Domaines d'application

3.2.1 Classification des images

La classification des images consiste à catégoriser une image dans une ou plusieurs classes. Ce problème est aussi défini par la classification des objets ou la reconnaissance des images. Il est considéré comme un problème basique en vision par ordinateur. Il constitue la base des autres tâches de vision par ordinateur telles que la localisation, la détection, ou la segmentation.

Au cours des dernières années, les techniques d'apprentissage profond ont considérablement progressé en vision par ordinateur, en particulier dans le domaine de la reconnaissance des objets.

Les méthodes d'apprentissage profond sont connues pour leur bonne performance sur les grands volumes de données et leur problème de surapprentissage sur les données limitées. Par conséquent, la base d'apprentissage ImageNet composée de 15 million d'images annotées a attiré beaucoup d'attention [Deng *et al.* 2009]. Comme nous l'avons mentionné dans le chapitre précédent, la performance des CNN peut être contrôlée par l'ajustement de la profondeur et de la largeur, et par le partage des poids. Cela implique un processus d'apprentissage court. Les CNN testés sur la base ImageNet ont atteint des performances satisfaisantes, cela a encou-

ragé la communauté de vision par ordinateur à les utiliser dans d'autres domaines (tableau 3.1).

Référence	Tâche
[Karpathy *et al.* 2014]	Classification des vidéos à grande échelle
[Lawrence *et al.* 1997]	Reconnaissance faciale
[Ciresan *et al.* 2011]	Classification des caractères manuscrits
[Tajbakhsh *et al.* 2016]	Analyse des images médicales
[Hu *et al.* 2015]	Classification des images hyperspectrales
[Spanhol *et al.* 2016]	Classification des images histologiques du cancer du sein
[Lee *et al.* 2015]	Identification des plantes
[Lakhani & Sundaram 2017]	Classification à partir des images radiologiques

TABLE 3.1 – *Les domaines d'application des CNN en classification des images.*

3.2.2 Détection et localisation des objets

La classification des images consiste à attribuer une classe à une image. Tandis que la détection des objets implique d'entourer un ou plusieurs objets dans une image par des cadres de sélection.

La détection des objets est une tâche plus difficile par rapport à la classification, car elle combine entre les notions de classification et de localisation. Elle permet de localiser et classifier avec précision les objets cibles dans une image. Par exemple, il est possible d'utiliser les méthodes de détection des objets pour identifier les cellules ou les tissus dans les images médicales [Xu *et al.* 2016].

La détection des objets est parmi les domaines connus en vision par ordinateur, et qui ont reçu beaucoup d'intérêt [Huang *et al.* 2014, Zhang *et al.* 2013]. Les méthodes de détection d'objets standard étaient basées sur les handcrafted features. Ces méthodes sont connues pour leurs manques de généralisation, car les attributs extraits dépendent du domaine de la tâche traitée. En plus, leurs évolution étaient très lentes entre 2010 et 2012 dans le défi PASCAL VOC, où les stratégies proposées se basent sur des méthodes ensemblistes et des algorithmes d'apprentissage classiques [Girshick *et al.* 2014]. Récemment, plusieurs efforts ont été faits pour résoudre ces problèmes en se basant sur les CNN.

Le réseau de neurones convolutif en tant que modèle DL a reçu un grand succès dans plusieurs domaines en vision par ordinateur. En 2012, [Krizhevsky *et al.* 2012] ont exploité ce réseau pour la classification des images et ils ont réussi à réduire le taux d'erreur des méthodes classiques de 26.2% à 15.3%. Ce progrès a encouragé la communauté de vision par ordinateur à utiliser les CNN en détection des objets.

En 2014, [Girshick *et al.* 2014] ont proposé R-CNN, qui est basé sur la recherche sélective et les algorithmes CNN et SVM. Cette méthode a atteint de bonnes performances et a réduit le temps de détection par rapport aux méthodes basées sur les fenêtres coulissantes pour la proposition des régions d'intérêts. Malgré l'efficacité de cette méthode en détection des objets, son temps de traitement n'est pas adapté aux applications en temps réel. Pour l'accélérer, plusieurs structures basées sur les CNN ont été proposées (Fast R-CNN[Girshick 2015] et Faster-RCNN [Ren *et al.* 2015]).

[Ren *et al.* 2015] ont développé un RPN qui permet presque de détecter les objets en temps réel. Ce réseau permet de prédire simultanément les

cadres de sélection et leurs précisions dans chaque position. La structure Faster R-CNN [Ren *et al.* 2015] combine entre les réseaux CNN et RPN pour effectuer une détection de bout en bout. Cependant, Faster R-CNN ne répond pas toujours aux exigences de la détection des objets en temps réel. La méthode YOLO [Redmon *et al.* 2016] est l'une des stratégies proposées pour adapter le temps de détection aux exigences d'applications en temps réel. Cette approche transforme le problème de détection des objets en un problème de régression. Le chapitre précédent explique en détail la structure des réseaux R-CNN, Fast R-CNN, Faster R-CNN, et YOLO.

Dans les travaux proposés en détection des objets, une variété d'architectures de type CNN ont été proposé : weakly supervised cascaded CNN [Diba *et al.* 2017], subcategory-aware CNN [Chen *et al.* 2017c], Alexnet [Girshick *et al.* 2014], et une architecture inspirée du réseau Inception [Redmon *et al.* 2016].

Les méthodes CNN en détection des objets ont été exploitées dans plusieurs domaines : télédétection [Long *et al.* 2017], diagnostic médical [Cireşan *et al.* 2013], et vidéo surveillance [Kang *et al.* 2017].

3.2.3 Segmentation sémantique

Au cours des dernières décennies, la segmentation sémantique a présenté l'un des grands défis en vision par ordinateur. Elle consiste à segmenter une image en différentes parties et objets. Son but est d'attribuer une classe à chaque pixel de l'image en entrée. Pour un ensemble de k classes $L = \{l_1, l_2, ..., l_k\}$ et N variables $X = \{x1, x2, ..., x_N\}$, chaque entrée x_i est associée à une classe l_j. L'espace de classes est composé de k états possibles, et qui sont généralement étendus à $k + 1$ pour traiter la classe fond de l'image. En général, X est une image 2D de $WxH = N$ pixels.

En segmentation, le traitement est plus compliqué par rapport à la reconnaissance et la détection des objets. La classification attribue une classe à chaque image et la détection classifie les objets et définit leurs cadres de sélection, tandis qu'un algorithme de segmentation peut également segmenter de nouveaux objets.

La segmentation des images a connu un grand intérêt pour la communauté de vision par ordinateur et d'apprentissage automatique. Les algorithmes de segmentation des images classiques sont généralement basés sur les méthodes de regroupement et des informations supplémentaires sur les contours et les bords [Weinland *et al.* 2011, Ilea & Whelan 2011]. Plusieurs approches ont été proposées pour améliorer la performance du regroupement. La modélisation à base du processus de Markov [Sacco 2005] et la combinaison de détection de contour dans une approche hiérarchique [Arbelaez *et al.* 2010] sont parmi les méthodes connues. Malgré la popularité des méthodes classiques, le nouveau succès des techniques d'apprentissage profond dans diverses tâches a rendu ces méthodes très populaires en vision par ordinateur y compris en segmentation.

Les techniques DL présentent l'alternative qui permet d'apprendre automatiquement les caractéristiques du problème traité au lieu de les extraire par les méthodes d'extraction, car ce processus nécessite une expertise dans le domaine, des efforts, et souvent trop d'ajus-

tement pour les adapter au problème traité. En apprentissage profond, les performances des CNN en classification [Krizhevsky *et al.* 2012] [Simonyan & Zisserman 2014b, Szegedy *et al.* 2015] et en détection des objets [Girshick *et al.* 2014, Girshick 2015, Ren *et al.* 2015] ont encouragé les chercheurs à les exploiter dans les problèmes de classification des pixels comme la segmentation sémantique. Ces réseaux ont été utilisés comme des composants dans plusieurs architectures de segmentation.

Les méthodes de segmentation des images en DL sont divisées en trois catégories : segmentation sémantique par région, segmentation sémantique basée sur les FCN, et la segmentation faiblement supervisée [Sinha *et al.* 2018].

Les méthodes de segmentation sémantique par région commencent par l'extraction des régions d'intérêts, ensuite, ces régions sont classifiées par des techniques de classification. R-CNN [Girshick *et al.* 2014] est l'une des architectures de type DL exploitée en détection des objets et en segmentation sémantique. Elle permet d'effectuer la phase de segmentation en se basant sur les résultats de la détection. Malgré l'efficacité de cette méthode, elle peut causer une perte d'informations liées au domaine, car les attributs utilisés proviennent des couches entièrement connectées, tandis que les couches intermédiaires contiennent plus d'informations spécifiques. En plus, la phase de génération des segments proposés a une complexité temporelle élevée, et cela peut affecter la performance finale.

L'idée des méthodes de segmentation sémantique à base de FCN est d'effectuer une transition pixels à pixels, sans avoir besoin de passer par l'étape de proposition des régions d'intérêts. FCN [Long *et al.* 2015] est parmi les réseaux les plus utilisés en segmentation sémantique. Il est considéré comme une extension des réseaux CNN, où les architectures connues (AlexNet [Krizhevsky *et al.* 2012], VGGNet [Simonyan & Zisserman 2014b], Inception [Riedmiller & Braun 1993], et ResNet [He *et al.* 2016]) sont transformées en FCN. Malgré son efficacité, FCN est caractérisé par certaines limitations liées à l'invariance spatiale, le manque d'information contextuelle, et la mauvaise résolution des images en sortie.

DeepLab [Chen *et al.* 2015, Chen *et al.* 2017b] présente l'une des solutions qui permettent d'améliorer la résolution en sortie. Cette méthode utilise un fully connected pairwise CRF [Krähenbühl & Koltun 2011] en tant que module séparé pour effectuer un post-traitement et affiner le résultat de la segmentation. D'autres travaux proposent d'améliorer la segmentation par l'exploitation des informations contextuelles. Par exemple, [Liu *et al.* 2015] ont utilisé la couche de Avg-pooling globale pour obtenir le contexte global. D'autres recherches ont résolu le problème de prédiction multi-échelle par la proposition d'un réseau composé de N FCN qui traitent différents échèles [Bian *et al.* 2016].

La segmentation en apprentissage faiblement supervisé est un autre domaine d'intérêt en segmentation sémantique [Papandreou *et al.*]. Le but de cette méthode est d'accélérer l'annotation des images dans la base d'apprentissage, car la génération des masques de segmentation pour l'apprentissage est une tâche difficile et couteuse en termes de temps. La segmentation en apprentissage faiblement supervisé propose l'utilisation des cadres de sélection au lieu des masques de segmentation pour réduire la

charge. Par exemple, [Dai *et al.* 2015] ont utilisé une annotation à base de cadre de sélection pour l'apprentissage, et ils ont obtenu d'une manière itérative les masques de segmentation.

La base d'apprentissage PASCAL VOC [Everingham *et al.* 2015] est parmi les bases connues en segmentation, et qui a été largement utilisée pour la validation des méthodes proposées en segmentation sémantique. Pour améliorer cette base, plusieurs extensions ont été développées : PASCAL Context [Mottaghi *et al.* 2014] et PASCAL Part [Chen *et al.* 2014]. Microsoft COCO [Lin *et al.* 2014] est une autre base de segmentation composée de plus de 80 classes.

Les méthodes DL en segmentation sémantique ont été exploitées dans plusieurs domaines d'application : voitures autonomes [Levi *et al.* 2015], imagerie médicale [Milletari *et al.* 2016], télédétection urbaine [Kampffmeyer *et al.* 2016], et segmentation des actions [Lea *et al.* 2016].

3.2.4 Reconnaissance d'action et d'activité

La reconnaissance d'action et d'activité est définie par la classification d'une activité à partir d'une séquence d'observations d'un objet. Son but est d'automatiser la reconnaissance des actions à partir des vidéos collectées par des systèmes d'acquisition. Cette technique est plus difficile par rapport à la reconnaissance des images à cause de sa dépendance des informations temporelles.

La reconnaissance des actions est un champ de recherche difficile, et qui a reçu un grand intérêt de la part de communauté de vision par ordinateur. Ces dernières décennies, un nombre important de méthodes a été proposé dans plusieurs domaines comme les vidéos de surveillance [Han *et al.* 2018]. En plus, différentes bases d'apprentissage spécialisées en reconnaissance d'actions ont été publiées [Karpathy *et al.* 2014].

Précédemment, les systèmes de reconnaissance des actions ont été basés sur les méthodes d'apprentissage automatique classiques. Ces méthodes utilisent des handcrafted features, où plusieurs méthodes d'extraction de caractéristiques ont été exploitées comme : HOG, HOF, et MBH [Yao *et al.* 2019]. Pour plus d'informations sur les méthodes d'extraction des caractéristiques en reconnaissance d'action, [Zhu *et al.* 2016] présentent une étude détaillée.

Durant ces dernières années, les réseaux d'apprentissage profond ont exploité les grands volumes de données et la quantité importante de vidéos disponibles sur Internet. En plus, le succès des CNN en classification des images a encouragé la communauté de vision par ordinateur à exploiter les architectures connues en classification des images (VGGNet, ResNet) dans la reconnaissance des actions. En classification, les réseaux CNN sont appliqués sur des espaces 2D, tandis qu'en reconnaissance d'actions, une vidéo est considérée comme un signal spatio-temporel 3D. Par conséquent, les CNN qui sont désignés pour l'extraction des informations spatiales ne conviennent pas au traitement des vidéos à cause de l'absence de l'information temporelle. Pour adapter les CNN à l'utilisation de cette information, plusieurs stratégies ont été proposées : convolution

3D, prendre en considération les informations relatives au mouvement en entrée du CNN, et fusion [Yao *et al.* 2019].

Les réseaux de neurones convolutifs tridimensionnels (3D CNN) appliquent des convolutions 3D sur les données temporelles et spatiales en entrée. [Ji *et al.* 2012] ont exploité le 3D CNN en reconnaissance des actions, où ils ont optimisé le réseau par la régularisation des sorties avec des caractéristiques de haut niveau. Dans une autre contribution, [Tran *et al.* 2015] ont développé un 3D CNN basé sur l'architecture VGGNet.

Les réseaux 3D CNN sont coûteux en termes de complexité de calcul et de capacité de stockage. Pour optimiser leur utilisation, la factorisation des 3D CNN à des 2D CNN est l'une des solutions proposées. Par exemple, [Sun *et al.* 2015] ont développé un CNN spatio-temporel factorisé, qui factorise l'apprentissage des filtres de convolutions 3D en un apprentissage 2D spatial suivi par un apprentissage 1D temporel.

Pour exploiter les informations temporelles, certains travaux ont pris en considération les informations relatives au mouvement, tels que le flux optique et le vecteur du mouvement. [Simonyan & Zisserman 2014a] ont proposé un modèle à deux flux, où le flux spatial effectue la reconnaissance d'action à partir des images fixes. Tandis que le flux temporel est utilisé pour reconnaître l'action à partir des informations de mouvement. Chaque flux est basé sur un modèle de type CNN, ensuite les scores softmax en sortie des deux flux sont combinés. Dans une autre contribution, [Christoph & Pinz 2016] ont introduit un réseau ResNet spatio-temporel par la combinaison du réseau ResNet et les techniques des modèles à deux flux.

3.2.5 Estimation de la pose humaine

L'estimation de la pose humaine est un problème connu en vision par ordinateur. Il permet de déterminer la position des articulations humaines à partir des images ou des séquences d'images. L'estimation de la pose humaine est une tâche très difficile à cause de la grande dimensionnalité des données en entrée et la variation élevée des poses humaines. La reconnaissance d'action et l'estimation de la pose humaine sont deux problèmes liés, mais généralement traités différemment dans la littérature. [Luvizon *et al.* 2018] étaient les premiers à proposer un réseau CNN multitâche qui permet de gérer les deux problèmes en même temps.

L'estimation de la pose humaine a présenté un rôle important dans différentes applications du monde réel motivé par les avancements technologiques actuelle. Parmi les applications connues, nous avons :

1. Les vidéos de surveillance [Kang *et al.* 2017] : la surveillance permet de suivre et de surveiller les mouvements dans des circonstances particulières, par exemple dans les aéroports ou les supermarchés.
2. L'interaction homme machine [Wang *et al.* 2018] : dans ces systèmes, les ordinateurs peuvent être contrôlés par exemple par des gestes humains ou le langage des signes.
3. L'interaction homme-robot [Liu & Wang 2018] : dans certaines situations de vie assistée, les robots doivent estimer les positions humaines pour assurer une bonne interaction.

4. L'imagerie médicale [Kügler *et al.* 2018] : l'estimation de la pose humaine peut être exploité pour assister les médecins à vérifier à distance les activités des patients.

Précédemment, le problème de l'estimation de la pose humaine a été traité à l'aide des structures picturales [Popa *et al.* 2017]. Dans les dernières années, l'apprentissage profond a prouvé son efficacité dans ce domaine, où différentes architectures ont été exploitées. Ces stratégies sont catégorisées en méthodes holistiques et à base de parties [Voulodimos *et al.* 2018]. Les méthodes de traitement holistiques réalisent la tâche d'une manière globale sans avoir besoin de définir explicitement un modèle pour chaque partie et leurs relations spatiales. En revanche, les méthodes à base de parties commencent par détecter individuellement les parties du corps humain, ensuite, un modèle graphique est utilisé pour intégrer l'information spatiale.

DeepPose [Toshev & Szegedy 2014] est le premier modèle DL proposé pour la détection de la pose humaine. Ce modèle appartient à la catégorie des méthodes holistiques.

Plusieurs travaux ont utilisé les CNN pour accomplir cette tâche. Par exemple, [Chen & Yuille 2014] ont proposé l'exploitation des patchs locaux et d'arrière plan pour l'entrainement d'un CNN afin de prédire les probabilités de présence des parties et leurs relations spatiales. Dans une autre contribution, [Jain *et al.* 2014] ont utilisé plusieurs CNN pour classifier indépendamment plusieurs parties du corps. [Tompson *et al.* 2014] ont proposé une hybridation entre un CNN et un champ aléatoire de Markov. Afin d'améliorer l'apprentissage du CNN, [Yang *et al.* 2016] ont combiné CNN avec un modèle de mélange de parties déformable (deformable mixture of parts model) pour effectuer un apprentissage de bout en bout.

3.2.6 Reconnaissance faciale

Le but des systèmes de reconnaissance faciale est de scanner, enregistrer, et reconnaitre les visages afin de les identifier. Ces systèmes étudient la correspondance entre l'image capturée et les images de la base d'apprentissage.

En comparant aux autres systèmes biométriques basés sur l'empreinte digitale et l'iris, la reconnaissance faciale est plus pratique dans les systèmes de surveillance, car le processus s'effectue sans contact direct, où les images peuvent être capturées à distance. La reconnaissance faciale est parmi les problèmes de vision par ordinateur qui a connu un grand intérêt par la communauté d'apprentissage automatique. Elle représente un domaine de recherche difficile à cause de plusieurs défis : la variation dans les expressions faciales, le changement de pose, et la différence dans l'éclairage.

Précédemment, les systèmes de reconnaissance faciale ont été basés sur les méthodes ML classiques, où une variété de méthodes d'extraction des caractéristiques ont été proposés [Tolba *et al.* 2006].

Ces dernières années, plusieurs méthodes à base de CNN ont été développées pour résoudre la tâche de reconnaissance faciale : light CNNs [Wu *et al.* 2018], Face Descriptor [Parkhi *et al.* 2015], Google's Fa-

ceNet [Schroff *et al.* 2015], Facebook's Deep-Face [Taigman *et al.* 2014], et OpenFace [Amos *et al.* 2016].

3.3 Les réseaux de neurones convolutifs pour l'analyse des images médicales

Les systèmes d'aide au diagnostic (CAD) sont exploités pour aider les spécialistes en médecine dans leurs décisions. Ces systèmes permettent de réduire l'inter-variabilité entre les décisions de différents experts et d'éviter la subjectivité. Ces dernières décennies, le développement du matériel et de logiciel, et la bonne qualité des images scannées ont encouragé la communauté de vision par ordinateur à améliorer l'efficacité des CAD.

Malgré les efforts faits, les CAD ont plusieurs limitations liées à la collecte et l'annotation des données nécessaires pour leur conception. La collecte exige un nombre important de patients et de tests pour assurer la quantité nécessaire de données. En plus, l'annotation est une tâche couteuse en termes de temps et d'effort, surtout la segmentation qui exige d'annoter chaque pixel dans l'image d'entrée. Généralement, l'annotation est assurée par un groupe d'experts pour garantir la validité des classes.

Les premiers CAD se basaient sur les systèmes à base de règles GOFAI [Litjens *et al.* 2017].

À la fin des années 1990, les techniques d'apprentissage classiques (ML) ont connu un grand intérêt par la communauté d'imagerie médicale [Shamir *et al.* 2008, Kather *et al.* 2016]. L'étape critique dans ces méthodes est la phase d'extraction des caractéristiques discriminantes à partir de l'image. Ce processus nécessite une étude approfondie par des experts en médecine et en vision par ordinateur. En plus, les caractéristiques extraites dépendent fortement de la tâche médicale traitée. Cela limite l'application des méthodes proposées sur d'autres types d'images médicales. Afin de résoudre ce problème, la solution logique est d'automatiser la tâche d'extraction des caractéristiques pour l'adapter à tout type d'application. Ce processus est réalisé par plusieurs types d'algorithmes d'apprentissage profond.

Les réseaux de neurones convolutifs sont parmi les réseaux les plus utilisés en traitement des images [Litjens *et al.* 2017]. Le développement dans la structure des CNN en vision par ordinateur a encouragé la communauté d'imagerie médicale d'exploiter ces architectures pour concevoir des CAD puissants.

En imagerie médicale, les images numérisées ont plusieurs types comme : ultrason (US), rayon-X, tomodensitométrie (CT) et imagerie par résonance (MRI), tomographie par émission de positrons (PET), et lames histopathologiques [Ker *et al.* 2017] (figure 3.2). Les algorithmes DL utilisent ces images pour résoudre différentes tâches en imagerie médicale : classification, localisation, détection, et segmentation.

D'un point de vue médical, une classification consiste par exemple à vérifier si une maladie existe ou pas ou pour distinguer entre différents types de cancers. La localisation permet de localiser et identifier des régions d'intérêts. Par exemple, localiser les zones tumorales sur une image pour les classifier. La détection permet de détecter plusieurs régions d'in-

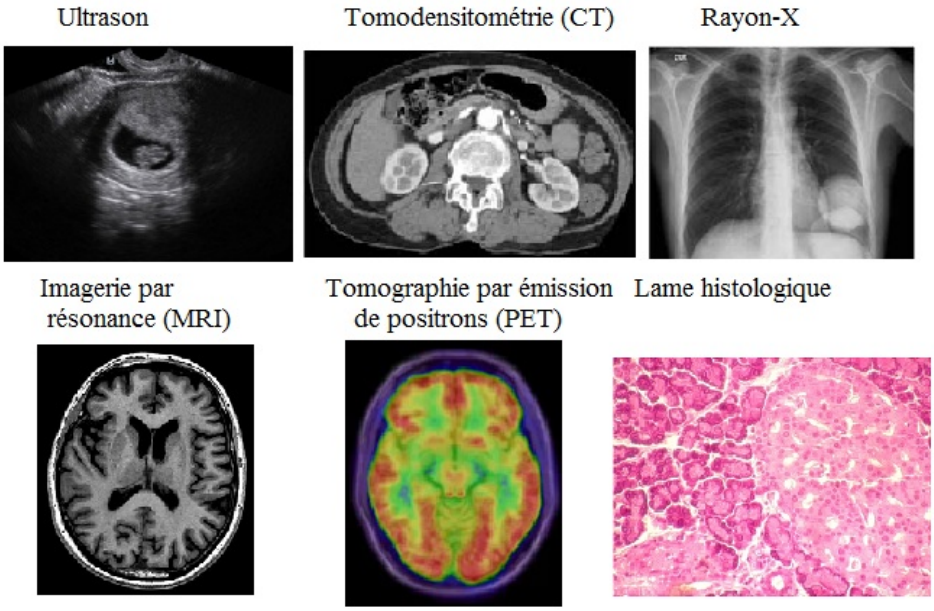

FIGURE 3.2 – *Les modularités des images médicales.*

térêt similaires sur la même image. Par exemple, il est possible de détecter les noyaux dans les images histopathologiques. Enfin, la segmentation permet d'identifier avec précision les régions d'intérêt, par exemple, la segmentation des tumeurs cérébrales sur les images MRI.

Nous détaillerons dans ce qui suit quelques méthodes (tableau 3.2) proposées dans l'état de l'art pour le traitement de différentes modularités d'images médicales. Pour plus d'informations, [Ker *et al.* 2017] et [Litjens *et al.* 2017] ont présenté des synthèses détaillées sur l'application des réseaux DL en imagerie médicale.

3.3.1 Classification

La classification des objets dans les images médicales consiste à classifier des parties précédemment extraites en deux ou plusieurs classes. Pour effectuer cette tâche, la localisation de la lésion est un prétraitement important pour assurer une bonne précision.

Plusieurs travaux ont été proposés dans l'état de l'art pour automatiser la classification de différentes modularités d'images médicales : rayon-X [Rajkomar *et al.* 2017, Rajpurkar *et al.* 2017], CT [Shen *et al.* 2015], MRI et PET [Li *et al.* 2014b], lames histologiques [Spanhol *et al.* 2016, Xu *et al.* 2016], et US [Byra *et al.* 2018], où l'exploitation des CNN a connu un grand intérêt. [Rajkomar *et al.* 2017] ont utilisé une version modifiée du réseau pré-entrainé GoogLeNet pour la classification des images de radiographie pulmonaire (rayon-X). Dans une autre contribution, [Rajpurkar *et al.* 2017] ont proposé une version modifiée de DenseNet composée de 121 couches de convolution (CheXNet). Le but de ce réseau est d'automatiser la classification de 14 maladies à partir des images de radiographie pulmonaire (rayon-X). Le système proposé a atteint les mêmes

Tâche	Référence	Modularité	Réseau	Application
Classification	[Rajkomar *et al.* 2017]	Rayon-X	Variante de GoogLeNet	La classification des images de radiographie pulmonaire
	[Rajpurkar *et al.* 2017]	Rayon-X	ChexNet	La classification des maladies à partir des radiographies pulmonaires
	[Li *et al.* 2014b]	MRI et PET	3D CNN	Assister dans le diagnostic d'Alzheimer
	[Spanhol *et al.* 2016]	Lames histologiques	Variante de AlexNet	La classification des tissus du cancer du sein
	[Xu *et al.* 2016]	Lames histologiques	CNN	La classification des tissus épithélial (EP) et stroma (ST)
	[Byra *et al.* 2018]	Ultrason	Inception−ResNet−v2 pré-entrainé	Evaluer la graisse dans la foie
Détection	[Lo *et al.* 1995]	Rayon-X	CNN	La détection des nodules
	[Ribli *et al.* 2018]	Mammographie	R-CNN	La détection et la classification des lésions malignes ou bénignes
	[Fan *et al.* 2018]	CT	Variante de Faster R-CNN	La détection des nodules pulmonaires
	[Yang *et al.* 2015]	MRI	3 réseaux CNN	La détection des points de repère sur la surface du fémur distal
	[Dou *et al.* 2016]	MRI	3D CNN	La détection des micro-saignements cérébraux
	[Sirinukunwattana *et al.* 2016]	Lames histologiques	CNN	La détection des noyaux
	[Cireşan *et al.* 2013]	Lames histologiques	CNN	La détection de mitose
Segmentation	[Pereira *et al.* 2016]	MRI	CNN	La segmentation du gliome dans le cerveau
	[Ronneberger *et al.* 2015]	Microscopie électronique	U-net	La segmentation des structures neuronales
	[Mercadier *et al.* 2019]	Lames histologiques	DCNN	La segmentation des noyaux
	[Peng *et al.* 2018]	Lames histologiques	FCN	La segmentations des tissus

TABLE 3.2 – *Résumé sur quelques travaux proposés pour le traitement des images médicales par les réseaux de neurones convolutif.*

performances des radiologistes. Dans la catégorie des scans MRI et PET, [Li *et al.* 2014b] ont utilisé un 3D CNN pour reconstruire les images Pet manquantes. Le système proposé permet d'assister les radiologistes dans le diagnostic de la maladie d'Alzheimer.

L'analyse des images histologiques est une étape non triviale dans le diagnostic de différents types de cancers. Pour faciliter cette tâche, plusieurs systèmes ont été proposés. Par exemple, [Spanhol *et al.* 2016] ont utilisé une variante du réseau AlexNet pour automatiser la classification des tissus du cancer du sein en tissu malin et normal. Dans une autre contribution, [Xu *et al.* 2016] ont utilisé un réseau CNN composé de deux couches de convolution pour la classification des tissues épithélial (EP) et stroma (ST) des cancers du sein et colorectal. Dans la catégorie des scans ultrason, [Byra *et al.* 2018] ont utilisé le réseau Inception-ResNet-v2 pré-entrainé pour l'extraction des caractéristiques et l'algorithme SVM pour la classification des images. Le système proposé permet d'évaluer la graisse dans le foie (stéatose).

3.3.2 Localisation et détection

La détection des objets d'intérêts dans une image est une étape importante dans le diagnostic qui exige beaucoup d'efforts par les cliniciens. Par exemple, une lame histologique peut contenir des centaines à des milliers de cellules à détecter [Ker *et al.* 2017].

Plusieurs travaux dans l'état de l'art ont automatisé la tâche de détection dans les images médicales. Le but principal des systèmes proposés est d'améliorer la précision et de réduire le temps de traitement manuel. La localisation des organes ou des points de repère a connu un grand intérêt, car elle présente une étape de prétraitement importante dans plusieurs tâches de segmentation.

Le premier système de détection à base de CNN a été proposé pour détecter les nodules dans des images à rayon-X [Lo *et al.* 1995]. Dans une autre étude, [Ribli *et al.* 2018] ont exploité le réseau R-CNN pour la détection et la classification des lésions malignes ou bénignes sur une mammographie. Dans la catégorie des images CT, [Fan *et al.* 2018] ont proposé une version optimisée du réseau Faster R-CNN pour la détection des nodules pulmonaires dans les images CT. Dans une autre contribution, [Yang *et al.* 2015] ont identifié sur les images MRI les points de repère sur la surface du fémur distal en se basant sur trois réseaux CNN. Dans une autre étude, [Dou *et al.* 2016] ont exploité un réseau 3D CNN pour détecter les micro-saignements dans des scans MRI cérébraux.

En histopathologie, la phase de recherche des marqueurs pour le diagnostic est une étape difficile qui nécessite un effort considérable à cause de plusieurs paramètres. Par exemple, le nombre élevé des noyaux et des mitotiques. En plus, l'architecture cellulaire atypique dans certains cancers provoque des confusions. Pour faciliter la tâche aux pathologistes, plusieurs systèmes de détection ont été développés. [Sirinukunwattana *et al.* 2016] ont exploité le réseau CNN pour détecter les noyaux dans les images histologiques d'adénocarcinome colorectal. [Cireşan *et al.* 2013] ont développé un système de détection de mitose basé sur un réseau CNN.

Dans cette thèse nous avons consacré une section qui détaille les travaux récents proposés pour la détection de la mitose en raison de la complexité de cette tâche et les défis qu'elle présentent.

3.3.3 Segmentation

La segmentation permet d'identifier l'ensemble de voxels ou de pixels qui constituent le contour ou l'intérieur des objets d'intérêts. Cette tâche est une étape importante dans les CAD, car elle permet d'identifier avec précision les régions d'intérêts dans les images médicales. L'automatisation de la segmentation des tumeurs dans le cerveau a connu un grand intérêt [Akkus *et al.* 2017], car elle permet de réduire l'effort manuel effectué par les spécialistes sur les volumes MRI et CT. Ce traitement est nécessaire pour diriger avec précision la résection chirurgicale.

Par exemple, [Pereira *et al.* 2016] ont utilisé un CNN composé de 11 couches de convolution pour la segmentation du gliome dans le cerveau à partir des scans MRI.

Dans la catégorie des images de microscopie électronique, [Ronneberger *et al.* 2015] ont proposé une nouvelle architecture U-net basée sur le réseau CNN pour la segmentation des structures neuronales. Cette architecture est composée du même nombre de couches de rééchantillonnage et de sous-échantillonnage. Elle permet de réaliser la tâche de segmentation en un seul passage et de prendre en considération l'information contextuelle. Dans une autre contribution, [Ciresan *et al.* 2012] ont appliqué une segmentation à base de pixel en utilisant la stratégie des fenêtres coulissantes et le réseau CNN.

Dans la catégorie des images histopathologiques, [Mercadier *et al.* 2019] ont exploité un DCNN composé des structures de codage-décodage pour la segmentation des noyaux. Pour accélérer le temps de traitement des images histopathologiques, [Peng *et al.* 2018] ont utilisé le réseau FCN pour la segmentation des tissus.

3.4 Conclusion

Les systèmes de vision par ordinateur sont basés sur les algorithmes d'apprentissage automatique. Ces systèmes sont utilisés pour analyser les entrées visuelles comme les images et les vidéos.

Durant ces dernières années, l'exploitation des réseaux DL en vision par ordinateur et spécialement des CNN a connu un grand intérêt dans plusieurs applications du monde réel, comme : la classification des images, la détection et la localisation des objets, la segmentation sémantique, la reconnaissance d'action et d'activité, l'estimation de la pose humaine, et la reconnaissance faciale.

Les CNN sont caractérisées par leur efficacité dans l'extraction des caractéristiques et ils sont aussi stables aux transformations. Ces caractéristiques ont fait des CNN un bon cas d'utilisation dans certaines applications en vision par ordinateur. Plusieurs variantes de type CNN ont été proposées dans la littérature. Le but principale était d'adapter l'architecture classique à la nature de l'application. En plus, la nature et la complexité de l'architecture varient selon la complexité du problème traité,

par exemple, la segmentation des objets est une tâche plus difficile par rapport à la détection et à la classification.

Le traitement des images médicales est parmi les problèmes connus en vision par ordinateur. Plusieurs architectures de type CNN proposées précédemment ont été exploitées et adaptées pour le traitement de différentes modularités d'images médicales.

En conclusion, malgré les résultats prometteurs des architectures proposées dans l'état de l'art, des défis importants subsistent. Ces défis sont liés au choix de l'architecture optimale pour résoudre une tâche définie. En plus, certaines architectures sont caractérisées par leur grande complexité, et cela limite leurs exécutions en temps réel dans les applications du monde réel.

Le chapitre suivant détaille les étapes de préparation et d'acquisition des images histopathologiques qui présententent le sujet d'intérêt de cette thèse. Dans ce cadre nous commencerons par l'explication de la routine de prédiction à base des images histopathologiques. Ensuite, nous détaillerons les techniques de prétraitement de ces images. Enfin, nous expliquerons la structure de quelques bases d'apprentissage histopathologiques publiques.

La préparation des images histopathologiques

4

Sommaire

L'analyse des images histolopatologiques colorées à l'hématoxyline et à l'éosine (H & E) est une tâche non triviale dans le diagnostic de plusieurs types de cancers. L'examen manuel de ces images a plusieurs enjeux liés à la subjectivité des décisions des pathologistes et la variance de ces images entre différents laboratoires. Le but de ce chapitre est de présenter les étapes d'acquisition et de prétraitement de ces images, et enfin nous décrivons quelques bases d'apprentissage histopathologiques accessibles au public. Ces bases sont exploitées par la communauté de l'imagerie médicale pour concevoir des CAD qui permettent d'automatiser quelques tâches en histopathologie et de réduire la charge des pathologistes.

Mots clés : Images histopathologiques, Acquisition, Prétraitement, Normalisation de couleurs, Microscope, Numérisation, Augmentation de données, Bases d'apprentissage histopathlogiques.

4.1 Introduction

La pathologie est une branche en médecine qui désigne la science de l'étude des maladies. L'histopathologie est une branche de l'histologie, où les tissus et les cellules biologiques malades sont examinés sous le microscope. L'étude de la structure et de la fonction des cellules et des tissus permet de révéler leur état de fonctionnement. Les structures non régulières et les déformations indiquent la présence d'une maladie.

Généralement, les pathologistes observent les biopsies colorées à H & E pour le pronostic, le classement et l'identification de différents types de cancers. Cette analyse est une étape non triviale dans le diagnostic des cancers [Araújo *et al.* 2017].

Récemment, les biopsies sont numérisées sous forme d'images en champ large (Whole slide images(WSI)) par les scanners de lame entière (Whole slide digital scanners (WSD)). Ces scanners sont des outils puissants pour la numérisation, l'acquisition et le partage des WSI, et cela facilite la phase de collecte des données pour la conception des CAD.

Le diagnostic des biopsies sous microscope est une tâche difficile et nécessite des années d'expérience, et cela engendre une grande variance entre les décisions de différents pathologistes. Comme solution, les CAD sont exploités pour réduire cette variabilité et diminuer la charge de travail des pathologistes.

Dans la plupart des applications, les CAD sont conçus par une succession d'étapes d'analyse. La première étape consiste à détecter ou à segmenter les régions d'intérêt à partir des WSI [Paeng *et al.* 2017]. Ensuite, ces régions sont traitées selon le type d'application, comme : la détection de la mitose [Cireşan *et al.* 2013], la segmentation des glandes [Kainz *et al.* 2015], et la classification des sous types de lymphome [Janowczyk & Madabhushi 2016].

Malgré les efforts faits, l'analyse des images histopathologiques a plusieurs défis liés au manque de données d'apprentissage et l'intervariabilité entre les images issues de différents laboratoires. Pour résoudre ces problèmes, des méthodes de normalisation et d'augmentation de données ont été suggérées. En plus, plusieurs bases d'apprentissage ont été proposées dans l'état de l'art pour encourager la communauté de vision par ordinateur à concevoir des applications pour le traitement des images histopathologiques.

Les travaux présentés dans cette thèse s'intéressent à l'analyse des images histopathologiques par les méthodes d'apprentissage profond.

Dans ce qui suit, nous détaillerons le processus de prédiction et les différents prétraitements effectués pour l'analyse des images histopathologiques. Ensuite, nous présenterons quelques bases d'apprentissage conçues pour le traitement de ces images.

4.2 Acquisition des tissus et numérisation des lames

Le cancer est défini par une masse de tissu composée de cellules génétiquement modifiées. Il présente l'une des principales causes de décès à cause des diagnostics tardifs [Torre *et al.* 2015].

Pour confirmer le diagnostic du cancer, l'analyse microscopique des cellules et des autres composants du tissu doit être effectuée par un pathologiste. Le pathologiste commence par l'acquisition des biopsies de tissu en suivant la technique d'extraction appropriée. L'étape d'extraction est une tâche très délicate, car le tissu peut être facilement endommagé lors du retrait. Ensuite, des sections de tissu sont extraites à partir des biopsies, et ces échantillons sont prétraités par différentes techniques : fixation, enrobage, se ctionnement et coloration [Mescher 2013] (figure 4.1).

FIGURE 4.1 – *Les étapes de prétraitement des échantillons de tissu en histopathologie.*

— **Fixation** : les échantillons extraits par le pathologiste sont délicats, ce qui rend la préparation des sections de tissu impossible. Afin de faciliter cette tâche, la biopsie est fixée par un fixateur chimique comme le formol. Ce fixateur empêche les changements physiques et chimiques afin de maintenir l'état actuel du tissu et de le protéger dans les étapes de prétraitement suivantes.
— **Enrobage** : cette étape permet de préparer les échantillons fixés pour l'étape de sectionnement. La paraffine est utilisée pour transformer le tissu d'une forme liquide à une forme solide et robuste au sectionnement.
— **Sectionnement** : dans cette étape, les blocs de paraffine sont coupés par microtome rotatif. Cette procédure permet de réduire l'épaisseur des échantillons pour les adapter aux lames microscopiques.
— **Coloration** : cette étape permet d'identifier les composants du tissu. Il existe différents protocoles de coloration en histopathologie, où H & E est le protocole de coloration le plus utilisé. H colore les noyaux des cellules en bleu noir et E colore les autres structures par différents degrés de rose [Fischer *et al.* 2008].

Après le prétraitement des biopsies extraites, le pathologiste observe les échantillons colorés à H & E sous le microscope. Il existe plusieurs techniques pour visualiser les lames histologiques, comme : le microscope électronique [Goodhew & Humphreys 2000], le microscope optique [Courjon 1990], et les scanners de lames (WSD) [Pantanowitz *et al.* 2011] (figure 4.2).

Les microscopes électroniques sont caractérisés par leur bonne résolution et une capacité de grossissement élevée par rapport aux microscopes

optiques. Ces deux types de microscopes peuvent être équipés d'une caméra de microscopie électronique pour capturer les images histopathologiques. Le pathologiste capture plusieurs régions d'intérêts à partir de la lame de verre sous différentes magnifications. De cette façon, plusieurs images sont générées à partir d'une seule lame de verre.

FIGURE 4.2 – *Les types des microscopes.*

Contrairement aux microscopes optique et électronique, un scanner de lames (ou microscope virtuel) peut capturer et numériser la lame de verre complète. Cette lame histologique est numérisée sous forme d'une WSI qui est caractérisée par une haute résolution. Ensuite, le pathologiste analyse l'échantillon capturé directement sur l'écran et effectue le diagnostic à l'aide d'un logiciel dédié [Farahani *et al.* 2015]. Cette technologie a facilité la numérisation et le partage de données, car il devient possible d'analyser à distance une lame de verre histologique. En plus, les images histologiques sont numérisées automatiquement, ce qui facilite la collecte des grands volumes de données pour la conception des systèmes d'analyse des images médicales. Enfin, l'acquisition de la lame de verre complète permet de conserver plus d'informations contextuelles.

4.3 Les techniques de prétraitement des images histologiques

L'analyse des échantillons histologiques est l'une des méthodes importantes pour le diagnostic, le pronostic, et le traitement de différents types de cancers [Thomas *et al.* 2015].

Le pathologiste évalue les caractéristiques histologiques sur les échantillons extraits comme : l'architecture du tissu, la densité des cellules, et l'activité de la mitose... etc. Par exemple dans le cancer du sein, le grade histologique est un facteur pronostic qui est déterminé par le système de gradation de Nottingham (NGS) [Frierson Jr *et al.* 1995]. Ce système dépend de trois caractéristiques morphologiques : la formation des tubules, le pléomorphisme nucléaire, et l'index mitotique.

Les informations microscopiques présentes sur les lames histologiques

reflètent le comportement des tissus et des cellules cancéreuses. Elles fournissent donc plus de détails sur l'agressivité de la maladie.

Le diagnostic précoce et l'évaluation précise du stade du cancer sont parmi les enjeux principaux pour la proposition du traitement approprié aux patients, où le suivi des patients varie considérablement et dépend de la nature des cellules tumorales.

Les WSI sont caractérisées par leur haute résolution et une structure biologique complexe. Par exemple, une image histologique est composée d'un nombre important de noyaux et de mitotiques. L'analyse manuelle de tous ces composants est laborieuse et exige un temps de traitement important. Ces facteurs rendent les décisions prises très subjectives, et cela augmente l'inter-variabilité entre les décisions des pathologistes.

L'utilisation des systèmes d'aide au diagnostic est parmi les solutions proposées pour réduire l'inter-variabilité entre les décisions des pathologistes. Ces systèmes permettent d'automatiser certaines tâches pour diminuer la charge de travail et d'optimiser les décisions prises. Le chapitre précédent détaillait le principe des CAD en traitement des images et expliquait la différence entre les méthodes ML et DL dans la conception de ces systèmes.

La disponibilité des WSD pour la numérisation des lames histologiques et la haute résolution des WSI ont fait des méthodes DL une bonne application en histopathologie en raison de leur adaptation aux grands volumes de données. Dans ce cadre, différentes tâches ont été automatisées, comme : la détection et la segmentation des noyaux [Sirinukunwattana *et al.* 2016], la détection de la mitose [Cireşan *et al.* 2013], et la classification des tissus [Spanhol *et al.* 2016, Xu *et al.* 2016].

Le but de cette thèse est d'automatiser et d'optimiser la classification des tissus histologiques en se basant sur différentes techniques d'apprentissages profond, et précisément sur les réseaux de neurones convolutifs. Les méthodes proposées visent à améliorer la performance des réseaux CNN et à éviter le problème de sur-apprentissage sur les volumes limités de données histopathologiques. Dans ce cadre, nous avons exploité plusieurs stratégies : apprentissage transféré, méthodes de régularisation, et méthodes ensemblistes.

4.3.1 Les méthodes de normalisation des images colorées à H&E

La préparation et la numérisation des échantillons histologiques peuvent conduire à une variation dans la couleur des images générées. Cette variation est causée par plusieurs facteurs, comme : l'utilisation de différents scanners et logiciels dans l'étape de numérisation et la différence entre les méthodes de prétraitement et de coloration entre laboratoires.

Les WSI provenant d'un seul ou de différents laboratoires risquent des problèmes d'intra-variabilité et l'inter-variabilité, respectivement. Cela peut influencer la performance des modèles DL dans l'analyse des images histologiques à cause de leur manque de généralisation sur les données non considérées durant l'apprentissage, par exemple, les images provenant des autres laboratoires et prétraitées sous différentes conditions. Dans ce cadre, les méthodes de normalisation de couleurs sont exploi-

tées pour ajuster la couleur des images. Cela réduit la variance entre les données d'apprentissage et de test et améliore la généralisation du modèle entrainé sur les images normalisées.

Dans ce qui suit, nous détaillerons quelques méthodes de normalisation de couleurs. Ces méthodes sont exploitées dans les systèmes d'analyse des images histologiques. La figure 4.3 illustre la différence entre les résultats de ces méthodes.

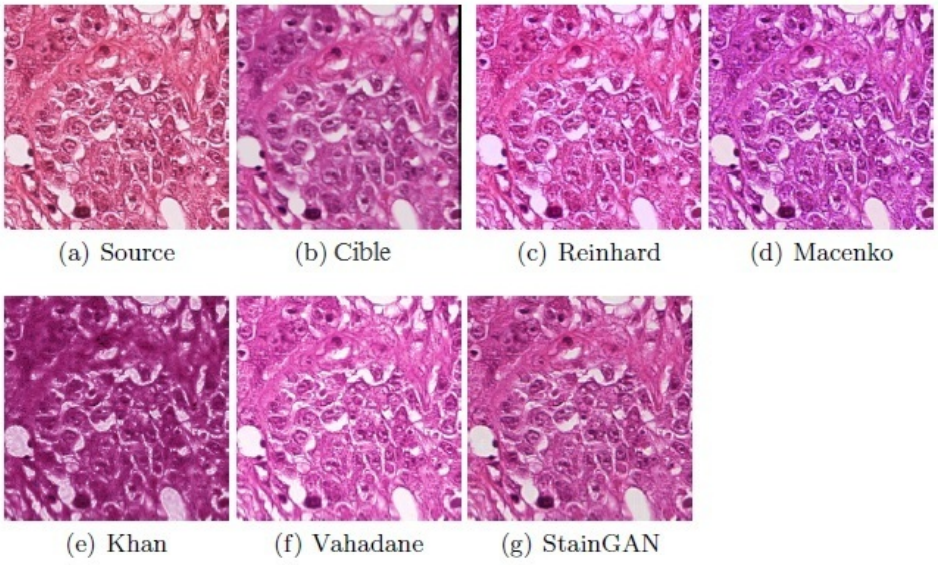

FIGURE 4.3 – *Le résultat des méthodes de normalisation de couleurs [Shaban et al. 2019].*

[Reinhard *et al.* 2001] ont proposé une méthode qui permet d'aligner les composants RGB de couleur pour réaliser une correspondance entre l'image source et cible. Cette méthode convertit l'ensemble des images à l'espace de couleurs $l\alpha\beta$, qui est défini par l'équation 4.1, où M représente la moyenne, σ est l'écart type, et m, o, t sont les images résultante, source, et cible respectivement. Cette méthode peut conduire à un mauvais alignement, car la même transformation est appliquée à toutes les images. En plus, elle ne prend pas en considération le degré de variation dans la coloration de ces images

$$\begin{cases} l_m = \frac{l_o - M_{l_o}}{\sigma_{l_o}} \sigma_{l_t} + M_{l_t} \\ \alpha_m = \frac{\alpha_o - M_{\alpha_o}}{\sigma_{\alpha_o}} \sigma_{\alpha_t} + M_{\alpha_t} \\ \beta_m = \frac{\beta_o - M_{\beta_o}}{\sigma_{\beta_o}} \sigma_{\beta_t} + M_{\beta_t} \end{cases} \tag{4.1}$$

D'autres méthodes proposent d'appliquer la normalisation sur chaque canal indépendamment [Macenko *et al.* 2009, Khan *et al.* 2014]. Par exemple, [Macenko *et al.* 2009] ont développé une méthode automatique basée sur le calcul des vecteurs de couleurs, et qui correspondent à chaque composant de coloration (H ou E) dans la WSI. Dans une autre contribution, [Khan *et al.* 2014] ont proposé une stratégie qui permet d'estimer la matrice de coloration à l'aide d'un classificateur de couleur. Ce classificateur attribue chaque pixel à la composante de coloration (H ou E) appropriée.

Contrairement aux méthodes citées précédemment, [Vahadane *et al.* 2016] ont développé une méthode qui effectue la normalisation en se basant sur

plusieurs modèles, où chaque modèle traite un composant de coloration (H ou E). [Shaban *et al.* 2019] ont proposé la méthode StainGAN basée sur Cycle GAN. Ce réseau permet d'aligner les images à un autre modèle de couleur en gardant la même structure du tissu. Dans une autre investigation, [Zanjani *et al.*] ont utilisé le réseau GAN dans le processus de normalisation des couleurs.

4.3.2 Les techniques d'augmentation des images histopathologiques

Les réseaux de neurones convolutifs ont été largement exploités ces dernières années grâce à leurs performances remarquables en vision par ordinateur. Ces réseaux sont généralement caractérisés par un nombre important de paramètres. Par exemple, l'architecture AlexNet contient environ 60 millions de paramètres [Krizhevsky *et al.* 2012]. Cette spécificité rend les réseaux CNN plus exigeants en termes de volumes de données par rapport aux méthodes d'apprentissage classiques, où il y a risque de problème de sur-apprentissage sur les volumes limités.

Le manque de données, la difficulté d'annotation, et le nombre déséquilibrés d'instances entre les classes présentent des défis majeurs dans le processus d'apprentissage des méthodes DL sur les données histologiques. Pour résoudre ces problèmes, plusieurs techniques ont été proposées dans l'état de l'art, par exemple : l'apprentissage transféré et l'augmentation de données.

En vision par ordinateur, les méthodes d'augmentation de données sont des techniques de régularisation qui permettent de générer plusieurs images cibles à partir d'une image source. Les techniques d'augmentation classiques sont basées sur des opérations basiques comme la rotation, la réflexion, le zoom, le recadrage... etc.

[Engstrom *et al.* 2017] ont montré que les réseaux DL peuvent facilement mal classer des images transformées par des rotations partielles. Par conséquent, l'augmentation de la base d'apprentissage par ce type de données permet d'améliorer la généralisation des réseaux DL et de générer des modèles robustes aux attaques contradictoires.

Malgré la quantité importante des images dans la base d'apprentissage ImageNet, les approches DL testées sur cette base ont exploité les techniques d'augmentation de données à cause du nombre élevé des paramètres des architectures proposées. Par exemple, Krizhevsky et al [Krizhevsky *et al.* 2012] ont commencé par l'extraction aléatoire des patchs de taille 224 x 224 à partir des images de taille 256 x 256. Ensuite, ils ont augmenté le nombre de patchs par des rotations et des réflexions. Ces opérations ont augmenté la taille de la base d'apprentissage par un facteur de 2048.

Les images histopathologiques sont caractérisées par leur grande dimensionnalité. L'exploitation directe de ces images dans l'apprentissage des réseaux CNN conduit à une augmentation dans le nombre de paramètres et cela accroit la complexité de calcul.

Les architectures CNN détaillées dans le chapitre 2 prennent en entrée des images de taille 224×224. Afin d'adapter la taille des images histologiques sur ce type d'architectures, plusieurs travaux ont proposé l'exploitation des fenêtres coulissantes sur ces images pour générer plusieurs

patchs de taille réduite. Cette stratégie est pratique dans le cas des données histologiques grâce aux structures microscopiques identiques distribuées dans l'image.

[Xu *et al.* 2017] ont extrait aléatoirement des patchs de taille 256×256 à partir des images de taille 3078×2752. Cette opération a généré environ 28000 patchs à partir de 4544 images. Dans une autre contribution, [Janowczyk & Madabhushi 2016] ont divisé chaque image de taille 1388×1040 à des patchs de taille 36×36 avec un pas de $S = 32$. Ensuite, ils ont rogné aléatoirement des patchs de taille 32×32 à partir des patchs extraits précédemment. Enfin, ils ont appliqué différents degrés de rotation à ces patchs.

En plus des méthodes d'augmentation de données classiques, les méthodes de transfert de style à base des réseaux GAN sont exploitées.

Le but principal des méthodes de transfert de style est de produire de nouvelles images qui imitent le style des images sources tout en préservant le contenu sémantique de la source.

[Roux *et al.* 2013] ont proposé un modèle stain-style transfer (SST) basé les réseaux GAN et qui permet d'apprendre la distribution des couleurs et aussi le schéma histopathologique correspondant. Dans une autre contribution plus récente, [Geessink *et al.* 2017] ont utilisé les méthodes de transfert de style dans la résolution du problème de segmentation des noyaux à base des méthodes DL. Ils ont commencé par le regroupement des images de style similaire en fonction de leur apparence. Ensuite, un réseau GAN est entrainé sur chaque groupe d'images pour générer des images synthétiques du style souhaité. Pour plus d'informations sur le réseau GAN, le chapitre 1 détaille son architecture.

4.4 La description des bases d'apprentissage hispatologiques

Les algorithmes DL sont caractérisés par leurs exigences en quantité de données. Cela présente l'un des grands enjeux dans la conception des CAD en histologie à cause du problème de la disponibilité de données médicales.

Plusieurs restrictions limitent la collecte des images histologiques : (a) la confidentialité, (b) le temps et les efforts considérables pour l'annotation de ces images, et (c) la différence entre les méthodes de coloration et de numérisation et des types des WSD dans les laboratoires. Afin d'encourager le développement des systèmes automatiques pour l'analyse des images histopathologiques, plusieurs défis (challenge) ont été organisés. Leur objectif était d'améliorer la performance sur des bases d'apprentissage annotées qui sont publiques et de haute qualité, où différentes tâches ont été traitées. En plus, dans la littérature, des efforts ont été consentis pour mettre des bases d'apprentissage histologiques à la disposition du public.

Après les étapes de prétraitement détaillées dans la section 2 de ce chapitre, les images générées sont sous forme de WSI ou sous forme de ROI extraites à partir de ces WSIs. L'exploitation des microscopes optiques ou électroniques liés à une caméra digitale ne permet pas la génération di-

recte des WSI. Dans ce cas, le pathologiste parcourt la WSI sous le microscope et extrait manuellement les ROI. Tandis que l'utilisation des WSD offre la possibilité de numérisation de la lame de verre complète. Ensuite, le pathologiste extrait manuellement les ROI à partir des WSI à l'aide d'un logiciel dédié.

Avec la disponibilité des WSD, il devient possible d'automatiser la tâche d'extraction des ROI à partir de la WSI à l'aide des techniques de vision par ordinateur, comme : la détection et la segmentation. Cela permet d'automatiser de bout en bout le traitement manuel effectué par le pathologiste.

Dans cette section nous détaillerons la structure de quelques bases d'apprentissage exploitées dans cette thèse (tableau 4.1).

Tâche	Base d'apprentissage	Type d'images
La détection de métastases dans les ganglions lymphatiques	CAMELYON16 [Bejnordi *et al.* 2017]	WSI
	CAMELYON17 [Geessink *et al.* 2017]	WSI
La classification et la détection des tissus du cancer du sein	BBHC-2015 [Araújo *et al.* 2017]	ROI
	Breakhis [Spanhol *et al.* 2015]	ROI
	ICIAR-2018 [Aresta *et al.* 2019]	WSI, ROI
La détection de la mitose	ICPR12 [Roux *et al.* 2013]	ROI
	MITOS-ATYPIA-14 [Roux *et al.* 2014]	ROI
Le score de l'atypie nucléaire		ROI
La classification des sous types de tissu dans le cancer colorectal	NCT-CRC-HE-100K-NONORM et CRC-VAL-HE-7K [Kather *et al.* 2019]	ROI
	CRC [Kather *et al.* 2016]	WSI, ROI
La classification des tissus métastatiques et non métastatiques	Pcam [Veeling *et al.* 2018]	ROI
La classification de 20 types histopathologiques de tissus	KIMIA-PATH960 [Kumar *et al.* 2017]	ROI
La classification des sous types de lymphomes non hodgkiniens	Lymphoma [Shamir *et al.* 2008]	ROI
Le score de la prolifération tumorale	TUPAC16 [Veta *et al.* 2019]	WSI

TABLE 4.1 – *La description de quelques bases d'apprentissage histopathologiques publiques.*

4.4.1 Bioimaging 2015 breast histology classification (BBHC-2015)

La base d'apprentissage BBHC−2015 [Araújo *et al.* 2017] est composée d'un ensemble d'images histologiques à haute résolution (2040×1536). Ces images ont été numérisées sous les mêmes conditions, sous un grossissement de $200\times$ et une taille de pixel de $0,42\mu m \times 0,42\mu m$. Chaque image était annotée par deux pathologistes par l'une des classes suivantes : normal, bénigne, cancer in situ, ou carcinome invasif.

4.4.2 Breakhis

La base d'apprentissage Breakhis [Spanhol *et al.* 2015] est composée d'un ensemble d'images histologiques des tumeurs malignes et bénignes du cancer du sein. Premièrement, des échantillons ont été prélevés par une biopsie chirurgicale (ouverte). Ensuite, après les étapes de prétraitement nécessaires, un microscope Olympus BX-50 lié à une caméra numérique Samsung a été utilisé pour numériser les lames de biopsie de tissu en images histologiques. Cette base d'apprentissage a été collectée à partir de 82 patients et numérisée sous différents niveaux de grossissement (40×, 100×, 200×, 400×) et une taille de pixel de 6.5 $\mu m \times 6.5 \mu m$. Le tableau 4.2 illustre le nombre et la taille des images sous chaque niveau de grossissement.

Grossissement	Taille	Nombre des images
40×	700 × 460	1995
100×		2081
200×		2013
400×		1820

TABLE 4.2 – *Le nombre et la taille des images dans la base d'apprentissage Breakhis.*

La base Breakhis est catégorisée en deux classes principales : (a) bénigne et (b) maligne, et chaque classe est sous catégorisée en 4 sous classes : (a) adénose (A), fibroadénome (F), tumeur phyllode (PT), adénome tubulaire (TA) et (b) carcinome canalaire (DC), carcinome lobulaire (LC), carcinome mucineux (MC), et carcinome papillaire (PC). La figure 4.4 présente quelques échantillons d'images appartenant à la base d'apprentissage Breakhis dans chaque classe.

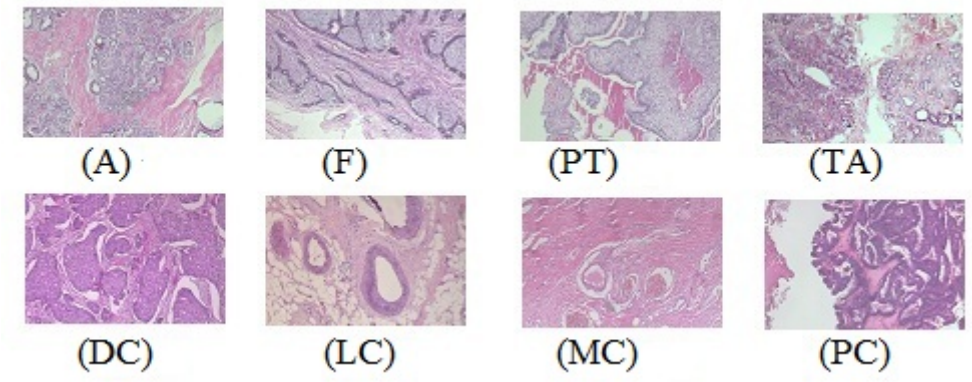

FIGURE 4.4 – *La structure des images de la base d'apprentissage Breakhis sous le grossissement* 40×.

Dans la phase d'acquisition des images, le pathologiste commence par l'identification de la tumeur et définit la région d'intérêt (ROI). Pour couvrir cette ROI, plusieurs images sont capturées sous le grossissement 40×. Ensuite, ce grossissement est manuellement augmenté à 100× et d'autres images sont capturées à partir de la ROI initial. Enfin, le même processus est répété pour les autres niveaux de grossissement.

4.4.3 ICIAR-2018

Dans la compétition ICIAR−2018 [Aresta *et al.* 2019], deux types de bases d'apprentissage ont été proposés. La première (ICIAR2018−A) est composée de 400 images microscopiques colorées à H & E d'une taille de 2048×1536 et une taille de pixel de $0.42\mu m \times 0.42\mu m$. Ces images ont été numérisées sous un grossissement de $200\times$ et annotées par deux pathologistes en : (A) tissu normal, (B) lésion bénigne, (C) carcinome in situ, ou (D) carcinome invasif, où chaque classe est composé de 100 images. La figure 4.5 présente des échantillons d'images appartenant à la base d'apprentissage ICIAR2018-A.

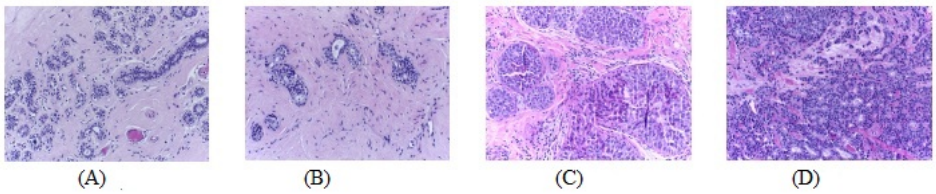

FIGURE 4.5 – *La structure des images de la base d'apprentissage ICIAR−2018−A [Aresta et al. 2019].*

La deuxième base d'apprentissage (ICIAR2018-B) est composée de 40 WSI au total. Ces images à haute résolution contiennent la lame de verre entièrement numérisée. Chaque WSI n'est pas associée seulement à une seule classe, mais à une liste de classes : (A) tissu normal, (B) lésion bénigne, (C) carcinome in situ, et (D) carcinome invasif (figure 4.6). La taille de ces images est variable, où la largeur $\in [39980, 62952]$ et la hauteur $\in [27972, 44889]$ pixels.

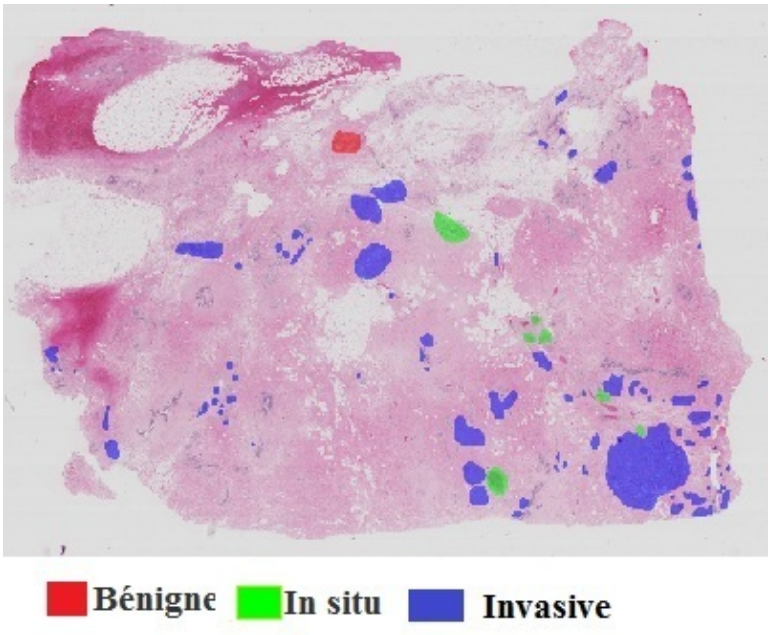

FIGURE 4.6 – *La structure d'une WSI annotée par pixel [Aresta et al. 2019].*

Ce type de bases d'apprentissage permet d'automatiser le processus en entier par l'automatisation de la phase de détection des régions d'in-

térêt. Tandis que les systèmes basés sur les bases d'apprentissage comme ICIAR2018-A et Breakhis nécessitent une phase d'extraction manuelle des ROI.

4.4.4 ICPR12, AMIDA13, MITOS-ATYPIA-14, et TUPAC16

Dans le cadre de la détection de la mitose, plusieurs bases d'apprentissage ont été proposées dans la littérature : ICPR12 [Roux *et al.* 2013], AMIDA13 [Veta *et al.* 2015], MITOS-ATYPIA-14 [Roux *et al.* 2014], et TUPAC16 [Veta *et al.* 2019]. Le tableau 4.3 présente une étude comparative entre ces bases d'apprentissage.

Base d'apprentissage		ICPR12	AMIDA13	MITOS-ATYPIA-14	TUPAC16	TUPAC16 auxiliaire
Scanners		Aperio (A) Hamamatsu (H) Microscope (M)	Aperio (A)	Aperio (A) Hamamatsu H)	Aperio (A)	Aperio (A) Leica SCN400 scanner (L)
WSI	Total	5	23	-	821	73
	Dimension	-	-	-	50000 × 50000	-
HPF	Total	50	596	136	-	656
	Dimension	A 2084 × 2084 H 2252 × 2250 M 2767 × 2767	2000 × 2000 -	1539 × 1376 1539 × 1376 -	-	2000 × 2000 L : 5657 × 5657 -
Mitoses	Apprentissage	226	550	-	-	1552
	Test	100	533		-	
Pathologistes		1	2	3	-	3
Gagnant		IDSIA [Cireşan *et al.* 2013]	IDSIA [Veta *et al.* 2015]	CUHK	LUNIT [Paeng *et al.* 2017]	LUNIT [Paeng *et al.* 2017]

TABLE 4.3 – *La description des bases d'apprentissage publiques proposées pour la détection de la mitose.*

ICPR 2012 est une base d'une taille réduite générée à partir de 5 WSI. Les images appartenant à cette base ont été collectées d'un seul laboratoire et annotées par un seul pathologiste. Par conséquent, cela réduit l'efficacité des modèles entrainés sur cette base en généralisation, car les problèmes d'intra-variabilité et d'inter-variabilité entre différents laboratoires ne sont pas traités lors de sa conception. Pour améliorer les systèmes proposés en détection de la mitose, d'autres bases d'apprentissage plus robustes ont été suggérées par la communauté de l'imagerie médicale : AMIDA13 et MITOS-ATYPIA-14.

La base d'apprentissage AMIDA13 contient un nombre important de champs de forte puissance (HPF) annotés par 2 pathologistes.

MITOS-ATYPIA-14 est une base d'apprentissage plus large par rapport à AMIDA13 composée de 1136 HPF et annotée par 3 pathologistes.

Malgré les efforts faits, les systèmes construits sur les bases d'apprentissage présentées précédemment n'automatisent pas la tâche complète, car les ROI sont sélectionnées manuellement par les pathologistes. Ces spécialistes calculent manuellement le score de prolifération en fonction du nombre des mitoses détectées par le système automatique.

Afin d'automatiser tous le flux de travail, la base d'apprentissage TUPAC16 offre la possibilité de prédire automatiquement le score de prolifération tumorale à partir des WSI, où deux bases d'apprentissage ont été proposées : TUPAC16-auxiliaire pour la détection de la mitose et une base des régions d'intérêt pour la sélection automatique des ROI. TUPAC16-auxiliaire est une extension de la base d'apprentissage AMIDA13 qui contient 50 WSI supplémentaires.

La base d'apprentissage MITOS-Atypia est conçue aussi pour automatiser la tâche du score de l'atypie nucléaire (NAS). Pour cela, des lames du cancer du sein invasif ont été numérisées par les scanners : Aperio Scanscope XT et Hamamatsu Nanozoomer 2.0-HT sous un grossissement de 20x. Ensuite, trois pathologistes ont annoté les ROI sélectionnées par l'une des classes suivantes : (1) atypie de bas grade, (2) atypie de grade modéré, et (3) atypie de haut grade.

4.4.5 CRC, NCT-CRC-HE-100K-NONORM et CRC-VAL-HE-7K

Les bases d'apprentissage CRC [Kather *et al.* 2016], NCT-CRC-HE-100K-NONORM et CRC-VAL-HE-7K [Kather *et al.* 2019] sont conçues pour automatiser la classification des sous types de tissu du cancer colorectal à partir des images colorées à H & E.

La base d'apprentissage CRC [Kather *et al.* 2016] est composée de 5000 images histologiques au total de taille 150 x 150. Ces images ont été extraites à partir de 10 WSI et annotées par l'une des 9 classes suivantes : (a) épithélium, (b) stroma simple, (c) stroma complexe, (d) lymphocytes, (e) débris, (f) muqueuse, (g) adipeux et (h) arrière-plan. La figure 4.7 présente quelques images représentatives pour chaque classe dans la base CRC.

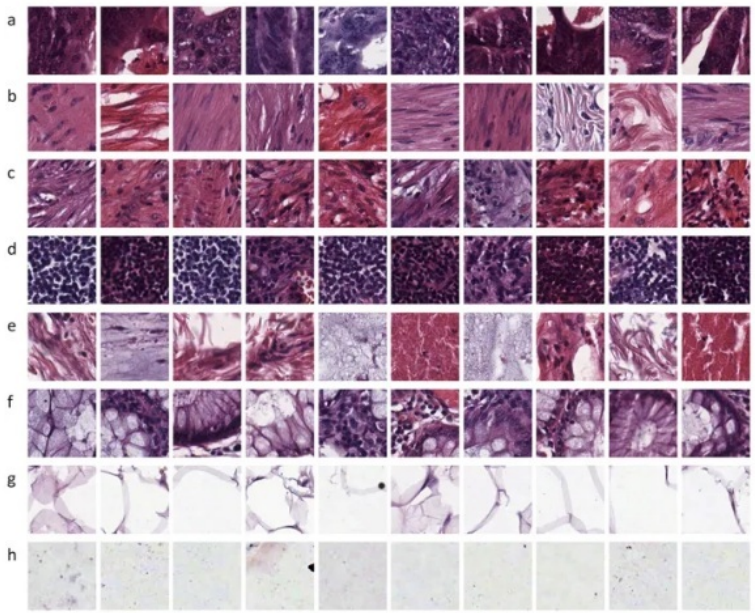

FIGURE 4.7 – *La structure des images de la base d'apprentissage CRC [Kather et al. 2016]*

La base d'apprentissage NCT-CRC-HE-100K-NONORM [Kather *et al.* 2019] est composée de 100000 images d'une taille de 224 x 224. Ces images ont été générées à partir de 86 lames de tissu du cancer colorectal et catégorisées en 9 classes : adipeux, fond, débris, lymphocytes, mucus, muscle lisse, muqueuse, stroma, et épithélium.

La base d'apprentissage CRC-VAL-HE-7K est composée de 7180 images de taille 224 x 224 et collectée de 50 patients. Les bases CRC-VAL-HE-7K et NCT-CRC-HE-100K-NONORM ont les mêmes types de classes.

4.4.6 Lymphoma

Le but de la base d'apprentissage Lymphoma [Shamir *et al.* 2008] est d'automatiser la classification de quelques sous types de lymphomes non hodgkiniens : CLL (leucémie lymphocytaire chronique), FL (lymphome folliculaire) et MCL (lymphome à cellules du manteau). Cette base a été collectée de plusieurs hôpitaux et annotée par différents pathologistes. Elle est composée de 375 images au total d'une taille de 1388 x 1040. Ces images ont été extraites à partir de 30 lames histologiques des ganglions lymphatiques et numérisées à l'aide du microscope Zeiss Axioscope et la caméra AXio Cam MR5.

4.4.7 Pcam

La base d'apprentissage Pcam [Veeling *et al.* 2018] a été conçue afin de mettre une base d'apprentissage histologique standard à la disposition de la communauté d'analyse des images médicales. Cette base a été générée à partir de la base Camelyon16 [Bejnordi *et al.* 2017], où 327 680 patches de taille 96 x 96 ont été extraits et catégorisés en métastatique ou non métastatique.

4.4.8 KIMIA-PATH960

La base d'apprentissage KIMIA-PATH960 [Kumar *et al.* 2017] a été collectée à partir des tissus musculaires, épithéliaux et conjonctifs. Premièrement, les régions d'intérêt sélectionnées par les pathologistes ont été numérisées, ensuite les 960 images résultantes ont été redimensionnées à 308 x 168 et annotées en 20 classes. La figure 4.8 illustre la structure de quelques images appartenant à cette base. Ces images sont caractérisées par une variation remarquable dans la texture.

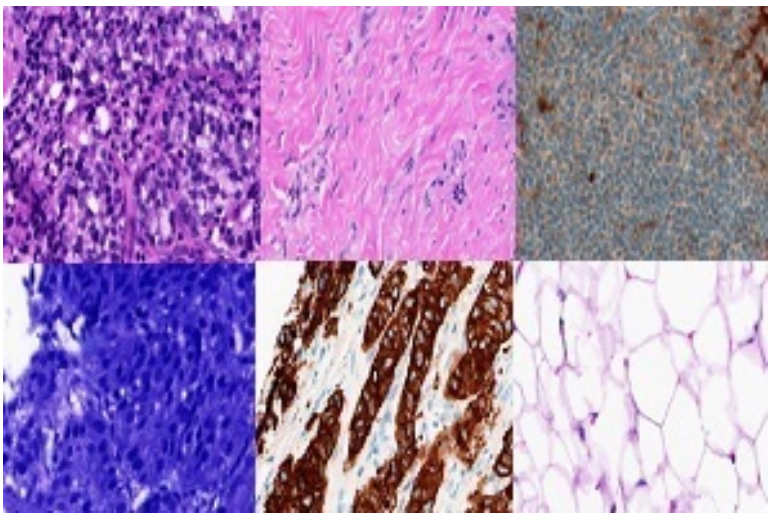

FIGURE 4.8 – *La structure des images de la base d'apprentissage KIMIA-PATH960.*

4.5 Conclusion

L'acquisition et la collecte des images histopathologiques ont connu un grand intérêt en vision par ordinateur. Les exigences des réseaux CNN en termes de volume de données ont encouragé les chercheurs en imagerie médicale à proposer des bases d'apprentissage publiques pour les mettre à la disposition de la communauté de vision par ordinateur. La collecte de ces données nécessite de passer par plusieurs étapes : extraction des échantillons de tissu, le prétraitement et la coloration, ensuite, la numérisation des lames de verre par les WSD.

Les données générées par différents laboratoires sont caractérisées par leur intra et inter-variabilité. Afin de créer des modèles robustes aux changements et d'améliorer la généralisation, des méthodes de normalisation de couleurs sont appliquées sur ces images.

Pour résoudre le problème de manque de données histologiques et la difficulté de leur annotation, plusieurs travaux ont proposé l'exploitation des méthodes d'augmentation de données afin de générer de grands volumes à partir des volumes limités de données.

Malgré les efforts faits, plusieurs enjeux subsistent dans ce domaine, car la collecte de ce type d'images pour la conception des CAD exige de prendre en considération plusieurs contraintes, comme : l'annotation des données par plusieurs pathologistes pour éviter les décisions subjectives, la prise en considération d'un maximum de cas variables de patients, et la collecte des données de plusieurs laboratoires et sous différentes conditions de numérisation afin d'améliorer la généralisation.

Le chapitre suivant présente l'état de l'art des contributions de cette thèse. Dans ce cadre, nous commencerons par un aperçu complet sur les méthodes DL proposées en détection de la mitose [Dif & Elberrichi 2020a]. Ensuite, nous détaillerons les travaux liés aux méthodes de régularisation : apprentissage transféré et méthodes ensemblistes en apprentissage profond.

5 Etat de l'art des méthode de régularisation en apprentissage profond et des méthodes DL en detection de la mitose

Sommaire

L'exploitation des méthodes DL pour l'analyse des images histopathologiques a connu un grand intérêt en raison de leur efficacité dans la résolution des problèmes complexes. Cependant, la collecte des données histopathologiques a plusieurs limitations liées à la sécurité, l'inter-variabilité entre laboratoires, et le temps élevé pour leur annotation. D'autre part, les méthodes DL exigent un grand volume de données afin d'ajuster équitablement leurs paramètres et d'éviter les problèmes de sur-apprentissage. Dans ce cadre, plusieurs travaux ont proposé une variété d'architectures et de méthodes de régularisation afin d'optimiser les méthodes DL pour la classification des images histopathologiques.

Mots clés : Détection de la mitose, Synthèse, Méthodes ensemblistes, Apprentissage transféré, Fine-tuning.

5.1 Introduction

Précédemment, les travaux publiés sur l'analyse des images histopathologiques étaient basés sur les méthodes d'apprentissage automatique, où le processus global dépend des modules d'extraction des caractéristiques et de classification.

Dans ce cadre, plusieurs descripteurs visuels ont été explorés, comme : LPB [Kather *et al.* 2016] et GLCM [Di Ruberto *et al.* 2015]. D'autre part, d'autres travaux ont suggéré la combinaison de plusieurs descripteurs [Di Ruberto *et al.* 2015]. Le but de cette combinaison est de définir une description standard d'une variété d'images. Cependant, les attributs extraits ont plusieurs inconvénients liés à leur dépendance du domaine traité.

Pour résoudre les différents problèmes, des études récentes suggèrent l'exploitation des réseaux DL en raison de leur processus d'extraction de caractéristiques supervisé à travers les couches intermédiaires. Cela est illustré dans l'étude effectuée par [Zeiler & Fergus 2014], où ils ont introduit une méthode de visualisation pour révéler le contenu des entités dans les couches intermédiaires. Leur visualisation illustre la nature hiérarchique des DNN. Par exemple, les premières couches présentent des structures simples : coin, bord et couleur. Tandis que les couches profondes identifient des structures plus complexes qui sont générées par la combinaison des structures précédentes.

Les développements récents en apprentissage profond ont encouragé l'exploitation de ces techniques dans différents domaines [Esteva *et al.* 2017, Cireşan *et al.* 2013, Nogueira *et al.* 2016, Lai *et al.* 2015] et particulièrement dans les CAD. Dans ce cadre, plusieurs algorithmes de type DL ont été développées : CNN [Li *et al.* 2014a], FCN [Roth *et al.* 2018], SSAE [Xu *et al.* 2015] pour résoudre différents problèmes en imagerie médicale : mammographie [Samala *et al.* 2016], histopathologie [Cireşan *et al.* 2013] et cardiovasculaire [Wolterink *et al.* 2016]. La plupart des travaux ont largement examiné l'efficacité des architectures CNN proposées dans le défi ILSVRC [Krizhevsky *et al.* 2012, Simonyan & Zisserman 2014b, Szegedy *et al.* 2015] en se basant sur trois techniques d'apprentissage : fine-tuning, apprentissage transféré et apprentissage à partir des initialisations aléatoires (Training from scratch). Pour plus d'informations sur les méthodes DL dans l'analyse des images histopathologiques, plusieurs synthèses détaillées ont été proposées dans l'état de l'art [Ching *et al.* 2018, Jimenez-del Toro *et al.* 2017, Anwar *et al.* 2018].

Le but de ce chapitre est de présenter les travaux liés à l'application des réseaux DL en histopathologie et l'optimisation de ces architectures par les méthodes ensemblistes et l'apprentissage transféré.

Premièrement, nous avons commencé par la présentation d'une synthèse sur les méthodes proposées pour la détection de la mitose [Dif & Elberrichi 2020a]. L'objectif de cette synthèse est de discuter un sujet qui a reçu un intérêt considérable par la communauté d'apprentissage profond. En plus, nous avons analysé, commenté, et comparé les approches proposées. Cette synthèse présente une référence importante pour les travaux futurs en détection de la mitose. Elle fournit une idée sur les techniques à éviter et elle présente quelques perspectives importantes

pour résoudre les problèmes et les enjeux non traités en traitement des images histopathologiques et précisément en détection de la mitose.

Ensuite, nous avons présenté les méthodes de régularisation liée aux méthodes ensemblistes et des techniques d'apprentissage transféré en apprentissage profond. L'objectif de cette partie est de discuter les approches précédemment proposées afin de justifier la contribution et les apports des approches proposées dans cette thèse. Le chapitre suivant détaille les méthodes proposées dans cette thèse et présente les résultats obtenus sur les bases d'apprentissage histopathologiques.

5.2 Les méthodes d'apprentissage profond pour la détection de la mitose à partir des images histopathologiques du cancer du sein : un aperçu complet

5.2.1 Introduction à la détection automatique de la mitose

Le cancer du sein (BC) est le cancer le plus fréquemment diagnostiqué chez les femmes [Bray *et al.* 2018]. En histopathologie, le pathologiste observe les biopsies du BC colorées à H & E sous le microscope pour déterminer le grade du cancer. Cette analyse est importante pour l'évaluation, le diagnostic et le traitement de la tumeur.

L'activité prolifératives est parmi les paramètres pronostiques dans le BC, où l'index mitotique est l'une des méthodes utilisées pour la mesurer [Beresford *et al.* 2006]. Généralement, le pathologiste détermine manuellement le nombre des mitoses sur les HPF sélectionnées. Cependant, cette tâche est fastidieuse, couteuse en temps de traitement, et risque de la subjectivité et de la variabilité entre les décisions des pathologistes. Pour réduire leur charge de travail, les CAD sont exploités. Ces systèmes sont basés sur les méthodes ML [Veta *et al.* 2014] et DL [Cireşan *et al.* 2013].

Dans ce cadre, [Kaman *et al.* 1984] ont proposé la première étude expérimentale automatisée pour le calcul des mitoses sur les images H & E. En 1993, [Ten Kate *et al.* 1993] ont automatisé cette tâche sur des spécimens colorés par Feulgen car il permet de révéler le contenu ADN. Dans la plupart des études récentes, un nombre important d'articles a été publié sur les techniques de détection de mitose à base des méthodes DL, où les réseaux CNN étaient les plus utilisés [Anwar *et al.* 2018, Tajbakhsh *et al.* 2016].

[Hamidinekoo *et al.* 2018] ont publié un article synthèse sur les méthodes DL en mammographie et histologie du cancer du sein. Ils ont détaillé 10 contributions pertinentes sur la détection de la mitose par les méthodes DL. Cependant, depuis 2018, les stratégies de détection de mitose à base des méthodes DL ont connu un grand intérêt. Malgré le nombre important des articles publiés, dans l'état actuel de nos connaissances, aucun papier synthèse n'a été publié.

Le but de cette section est de présenter une étude comparative entre les méthodes DL proposées pour la détection de la mitose [Dif & Elberrichi 2020a]. Cette tâche peut être effectuée sur les images de microscopie à contraste de phase [Su *et al.* 2017] et les images colorées à PHH3 [Tellez *et al.* 2018] ou à H & E [Chen *et al.* 2016a]. Dans ce cadre,

nous étions intéressé aux méthodes testées sur les images colorées à H & E en raison de leur disponibilité et vaste utilisation.

Dans cette section, nous avons collecté 28 publications à partir de Google Scholar et des articles synthèses sur les méthodes DL pour l'analyse des images médicales. Nous avons utilisé les mots-clés suivants pour la recherche des publications : « détection de la mitose », « apprentissage profond », « cancer du sein », « réseaux de neurones convolutifs ». Premièrement, nous avons sélectionné la période 2012-2019. Ensuite, la période 2018-2019 pour un maximum de publications récentes. Dans une deuxième stratégie, nous avons filtré les papiers concernés parmi tous les travaux qui ont cité les articles ICPR12 [Roux *et al.* 2013], AMIDA13 [Veta *et al.* 2015], MITOSIS-ATYPIA-14 [Roux *et al.* 2014] et TAUPAC16 [Veta *et al.* 2019].

5.2.2 Généralités sur le cancer du sein

Le cancer du sein est le cancer le plus fréquemment diagnostiqué chez les femmes avec 11.6% du nombre total des décès par le cancer [Bray *et al.* 2018]. Afin de détecter ce type de cancer, des examens de dépistage sont utilisés tels que la mammographie [Kallenberg *et al.* 2016, Samala *et al.* 2016], l'échographie, et l'IRM [Dalmış *et al.* 2017]. L'échographie a prouvé son efficacité par rapport à la mammographie dans le diagnostic des masses solides [Zhi *et al.* 2007]. Ces tests sont utilisés pour une détection précoce du cancer et donc ils améliorent les chances de survie.

Après un test anormal de dépistage, la biopsie est recommandée pour l'évaluation, le diagnostic, et le traitement de la tumeur. Il existe différents types de biopsies du sein : aspiration à l'aiguille fine (fine-needle aspiration : FNA), biopsie par forage (core needle biopsy : CNB) et biopsie par excision (excision biopsy : EB).

CNB est la technique privilégiée pour l'évaluation histologique et le traitement chirurgical [Willems *et al.* 2012]. Cette technique est moins coûteuse que EB. En plus, elle dévoile la structure histologique globale du tissu par rapport à FNA [Oyama *et al.* 2004]. Afin d'extraire avec précision le tissu à partir de la région d'intérêt, le processus d'extraction est guidé par l'ultrason [Fishman *et al.* 2003]. Ensuite, le tissu extrait est envoyé au pathologiste pour l'examen histologique. Le pathologiste prépare des échantillons à partir du tissu en suivant les techniques de prétraitement détaillées dans le chapitre précédent.

L'analyse des échantillons colorés à H & E permet de vérifier la présence du cancer. Si ce dernier est détecté, le pathologiste effectue une classification histologique et vérifie l'étendue du cancer (in situ ou invasif). Le BC peut être développé dans les tissus épithéliaux (carcinome) ou stromas (sarcomes), et les carcinomes peuvent être localisés dans les canaux galactophores ou les glandes, nommés carcinome canalaire (CD) et carcinome lobulaire (LC), respectivement [Makki 2015]. CD in situ présente 83 % des cas diagnostiqués chez les femmes [Ward *et al.* 2015].

Le pathologiste utilise les systèmes de classification comme facteurs pronostiques pour évaluer l'apparence de la cellule, la taille de la tumeur et son comportement prolifératif. Actuellement, le système de grade his-

tologique Nottingham [Frierson Jr *et al.* 1995] est utilisé pour déterminer le grade du BC.

Le système Nottingham [Frierson Jr *et al.* 1995] est basé sur trois caractéristiques morphologiques : la formation des tubules, le pléomorphisme nucléaire et l'index mitotique. Ces caractéristiques sont catégorisées à des index de 1 à 3. Le score de formation des tubules représente un indicateur sur le pourcentage des structures tubulaires dans la zone tumorale. Le pléomorphisme nucléaire indique le degré de variabilité des noyaux par rapport à leur état normal. L'index mitotique spécifie le nombre des structures mitotiques dans la tumeur et son comportement prolifératif [Beikman *et al.* 2013].

La partie suivante détaille la procédure de calcul de l'index mitotique qui est considéré comme un marqueur pronostique important pour l'analyse du BC invasif.

5.2.3 Le calcul de l'indexe mitotique

L'activité prolifération présente un paramètre pronostique important dans le cancer du sein. Elle est liée à l'agressivité du cancer, où une forte activité dépend d'une division cellulaire incontrôlée et révèle donc un risque élevé. Cette activité peut être mesurée par différentes méthodes comme : l'évaluation de cellule en phase S, les anticorps Ki-67 en immunohistochimie, et l'activité mitotique [Beresford *et al.* 2006].

En oncologie, l'indice mitotique informe sur le nombre de cellules subissant des divisions nucléaires (mitose). Dans le processus de la mitose, il existe quatre phases de base : prophase, métaphase, anaphase et télophase. Le noyau de la mitose apparaît plus dense au début de la mitose et se transforme en une cellule à deux noyaux en télophase.

Afin de calculer l'index mitotique, le pathologiste commence par l'identification des ROIs sous un faible grossissement, car la WSI peut contenir des dizaines de milliers de HPFs. Chaque ROI correspond à 2 mm^2 ou à 10 HPFs. Ensuite, les mitoses sont comptées manuellement sous un grossissement de $\times 40$. Cela permet d'indexer les ROIs par un score de 1 à 3 en fonction du nombre des mitoses par région. Ce processus est fastidieux, exige un temps de traitement important (5 à 10 minutes par ROI [Gal *et al.* 2005]), et peut causer des problèmes d'intra-variabilité et d'inter-variabilité entre pathologistes [Orchid & Puthanpurayil 2016].

La variabilité est liée à plusieurs facteurs : (a) la sélection subjective des ROI [Bonert & Tate 2017], la morphologie variable des mitoses durant le processus de transformation, (c) l'apparence similaire des mitoses à d'autres structures qui peut causer un taux élevé de faux positifs, et (d) le nombre réduit de mitoses par rapport aux noyaux des cellules normales.

Pour résoudre les problèmes cités auparavant et réduire la charge de travail du pathologiste, les CAD sont proposés pour automatiser la tâche de détection de la mitose. Dans ce cadre, plusieurs travaux ont été proposés dans l'état de l'art à base des méthodes ML et DL. Le but principal de ces contributions était de résoudre les différents obstacles liés à la détection automatique de la mitose. Par exemple, en télophase, la cellule contient deux noyaux séparés et désigne une seule mitose. En outre, la fréquence faible des mitoses et le nombre limité de cellules en cours de

mitose par rapport aux cellules normales engendrent des bases d'apprentissage déséquilibrées. En plus, les étapes de préparation, de coloration, et de numérisation des biopsies engendrent des images histologiques non uniformes.

La partie suivante discute les méthodes d'apprentissage profond proposées dans l'état de l'art pour la détection des mitoses.

5.2.4 Les méthodes d'apprentissage profond pour la détection de la mitose

La figure 5.1 présente la distribution des 28 articles sélectionnés dans cette synthèse, par an [Dif & Elberrichi 2020a]. Elle illustre le nombre important des travaux publiés en 2018 incluant janvier 2019. La première contribution a été publiée en 2008. Dans cette période, la plupart des travaux étaient plutôt intéressés par les approches d'apprentissage automatique classiques en raison de leur large exploitation en vision par ordinateur et le manque des ressources puissantes et des bases d'apprentissage publiques pour la détection de la mitose.

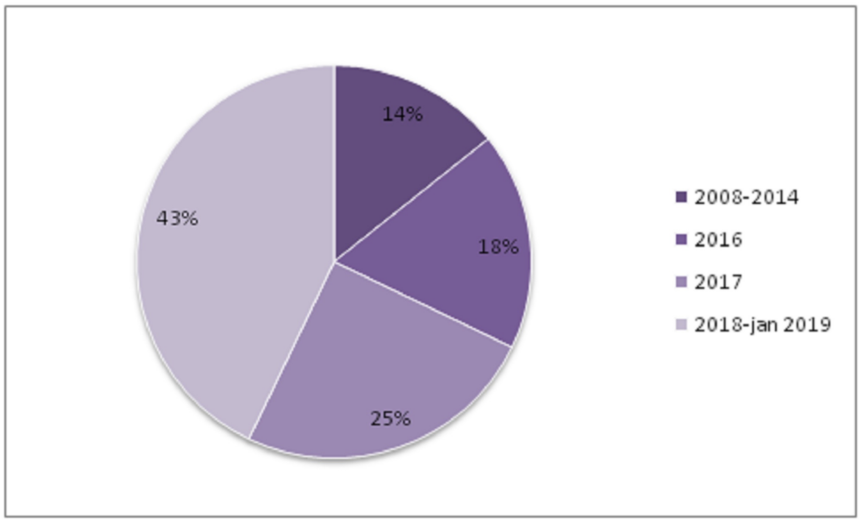

FIGURE 5.1 – *La distribution de travaux proposés en détection de la mitose par an.*

Le processus de la détection de la mitose en se basant les méthodes DL a plusieurs inconvénients liés à :

- Les volumes limités de données.
- Le nombre limité de figures mitotiques en raison de leur faible fréquence.
- Le taux élevé des faux positifs .
- La grande variance entre les images histopathologiques qui ont été numérisées dans différentes conditions.
- Problèmes de sur-apprentissage.
- Les exigences en termes de ressources matériel et le temps de traitement élevé.

Pour résoudre ces limitations, plusieurs travaux ont été proposés dans l'état de l'art, où différentes stratégies ont été exploitées, telles que :

— Les stratégies de régularisation pour réduire les problèmes de sur-apprentissage.
— Les techniques de l'apprentissage transféré, fine tuning, et l'exploitation des CNN en tant qu'extracteur de caractéristiques afin de réduire la complexité temporelle en apprentissage et de résoudre les problèmes de sur-apprentissage.
— L'exploitation du réseau FCN et les méthodes d'apprentissage profond en détection pour optimiser la complexité de détection en inférence.
— les réseaux de régression pour réduire le temps d'inférence.
— L'apprentissage multi-échelles pour améliorer le processus de détection.
— Les stratégies d'apprentissage en deux phases pour résoudre le problème du taux élevé des faux positifs

Les méthodes de régularisation

Malgré la disponibilité des bases d'apprentissage pour la détection de la mitose, le nombre d'instances dans ces bases reste limité pour les applications de type DL. En plus, ces bases sont déséquilibrées à cause du nombre réduit des figures mitotiques par rapport aux autres structures. Pour résoudre ces limitations, des études ont proposé l'exploitation des méthodes de régularisation, telles que : les réseaux moins profonds [Cireşan *et al.* 2013, Wang *et al.* 2014, Das & Dutta 2019], les techniques d'augmentation de données, l'apprentissage transféré [Xu *et al.* 2017], fine-tuning [Chen *et al.* 2016b, Wu *et al.* 2017], l'utilisation des CNN comme des modules d'extraction des caractéristiques [Albayrak & Bilgin 2016], les méthodes ensemblistes [Xu *et al.* 2017], et l'apprentissage basé sur les foules [Albarqouni *et al.* 2016]. Le but principal de ces techniques était d'améliorer la généralisation des modèles DL générés.

Les méthodes de normalisation des couleurs [Macenko *et al.* 2009, Khan *et al.* 2014, Reinhard *et al.* 2001] sont parmi les solutions importantes qui permettent de réduire la variabilité entre les laboratoires et d'améliorer la généralisation. Le but principal de ces stratégies était de convertir les lames traitées dans différentes conditions à un espace normalisé [Macenko *et al.* 2009]. Cette étape est importante pour l'exploitation des modèles entrainés sur les données des autres laboratoires. Le tableau 5.1 résume les techniques de normalisation exploitées comme prétraitement dans la tâche de détection de mitose.

À notre connaissance, la technique SVD-geodesic [Macenko *et al.* 2009] est la méthode la plus fréquemment utilisée dans les méthodes automatiques de détection de la mitose [Albarqouni *et al.* 2016, Das & Dutta 2019, Veta *et al.* 2016, Paeng *et al.* 2017, Zerhouni *et al.* 2017, Rao 2018]. Malgré l'importance et l'efficacité de ce prétraitement, plusieurs articles ont ignoré cette étape [Cireşan *et al.* 2013, Chen *et al.* 2016a]. Cependant, l'étape de normalisation des couleurs est un prétraitement trivial si les données d'apprentissage et de test sont générées sous les mêmes conditions. D'autre

Référence	Méthode de normalisation de couleurs
[Albarqouni *et al.* 2016] [Das & Dutta 2019] [Veta *et al.* 2016] [Paeng *et al.* 2017] [Zerhouni *et al.* 2017] [Rao 2018]	La technique SVD-geodesic [Macenko *et al.* 2009]
[Wu *et al.* 2017] [Kausar *et al.* 2018] [Akram *et al.* 2018]	Le transfert de couleur entre les images [Reinhard *et al.* 2001]
[Beevi *et al.* 2019]	Une approche de mappage non linéaire basée sur la technique de déconvolution de couleur [Khan *et al.* 2014]
[Shah *et al.* 2017]	Une méthode d'uniformisation de couleurs [Bejnordi *et al.* 2015]

TABLE 5.1 – *Les techniques de normalisation des couleurs utilisées dans les méthodes proposées pour la détection de la mitose.*

part, elle permet d'exploiter les modèles d'apprentissage dans d'autres laboratoires hétérogènes.

Les stratégies de classification par pixel et par patch

Plusieurs travaux récents ont exploité les méthodes DL pour automatiser la détection de la mitose sur les lames colorées à H & E. À notre connaissance, [Malon *et al.* 2008] étaient les premiers qui ont automatisé cette tâche à base des réseaux CNN. Ils ont utilisé un ensemble de 728 images numérisées sous un grossissement de 400×. Ensuite, ils ont appliqué l'algorithme SVM dans la phase d'apprentissage sur les données générées par CNN. Les tableaux 5.2, 5.3 et 5.5 résument les méthodes DL proposées pour la détection de la mitose.

Les méthodes DL proposées en détection de mitose sont classifiées en stratégie de classification par pixel [Cireşan *et al.* 2013, Albarqouni *et al.* 2016, Veta *et al.* 2016, Zerhouni *et al.* 2017] et de classification par patch [Janowczyk & Madabhushi 2016, Wu *et al.* 2017]. Les stratégies de classification par pixel sont considérées en tant qu'une segmentation sémantique, où chaque pixel appartenant à l'image est étiqueté séparément en mitose ou non-mitose.

[Cireşan *et al.* 2013] ont proposé un max-pooling CNN pour effectuer la classification par pixel en détection de la mitose. Dans ce cadre, ils ont combiné 3 CNN en se basant la méthode ensembliste par moyenne. Le but principal de cette combinaison était d'améliorer la généralisation des modèles générés. Les meilleurs résultats obtenus dans les défis ICPR12 [Roux *et al.* 2013] et AMIDA13 [Veta *et al.* 2015] prouvent l'efficacité de cette méthode. Cependant, l'inconvénient majeur de cette approche est son temps d'inférence élevé : 8 min par HPF. En outre, le pathologiste sé-

Méthode	Segmentation	Classification	Apprentissage	Base d'apprentissage
[Malon *et al.* 2008]	Color histogram	CNN SVM	Initialisation aléatoire	Un ensemble de 728 images sous un grossissement de 400X
IDSIA [Cireşan *et al.* 2013]	Max pooling CNN		Initialisation aléatoire	ICPR12
[Janowczyk & Madabhushi 2016]	Blue-ratio	CNN (cifar-10 AlexNet)	Initialisation aléatoire	-
[Das & Dutta 2019]	Blue-ratio + Otsu's thresholding	CNN	Initialisation aléatoire	ICPR12 MITOS-ATYPIA-14
FF-CNN [Wu *et al.* 2017]	Blue-ratio	FF-CNN	Fine tuning : le modèle AlexNet	MITOS-ATYPIA-14
[Albayrak & Bilgin 2016]	Regroupement par K-means	CNN pour l'extraction des caractéristiques SVM pour la classification	Initialisation aléatoire	MITOS-ATYPIA-14
[Beevi *et al.* 2019]	L'algorithme Krill Held (KHA)	CNN pour l'extraction des caractéristiques Softmax pour la classification	Fine tuning : Le modèle caffe VGGNet	MITOS-ATYPIA-14 Regional Cancer Centre (RCC)
HC + CNN [Wang *et al.* 2014]	Blue-ratio + laplacian of Gausian + globally fixed and local dynamic thresholdings	Régression logistique sur les attributs CNN Random forest sur les hadcrafted features	Initialisation aléatoire	ICPR12
[Malon & Cosatto 2013]	Color threshold + grid search	CNN (LeNet) pour l'extraction des caractéristiques SVM pour la classification	Initialisation aléatoire	ICPR12
[Saha *et al.* 2018]	Blue-ratio + érosion morphologique et opérations de dilatation	CNN	Initialisation aléatoire	ICPR12 MITOS-ATYPIA-14 AMIDA13
CasNN [Chen *et al.* 2016a]	FCN	CNN	Fine tuning (CNN)	ICPR12 MITOS-ATYPIA-14
DeepMitosis [Li *et al.* 2018a]	Segmentation : FCN Detection : faster R-CNN Verification : CNN (ResNet50)		-Fine tuning (FCN) à partir de VGGNet16 -Apprentissage transféré : (R-CNN) à partir de VGGNet_1024 Initialisation aléatoire (CNN)	ICPR12 MITOS-ATYPIA-14

TABLE 5.2 – *Les méthodes d'apprentissage profond proposées pour la détection de la mitose (1).*

lectionne plusieurs HPF à partir du WSI pour l'analyse. Cela rend cette méthode non pratique pour l'utilisation clinique en temps réel.

Afin de réduire la complexité temporelle d'inférence, les stratégies de classification par patch ont été largement exploitées. Premièrement, des patchs ou des candidats de mitose sont générés et catégorisés par la suite en mitose ou non mitose. Dans le processus de génération des patchs, les images sont converties en ratios de bleu. Cette étape permet de faire apparaitre les noyaux candidats en raison de leur intensité élevée de bleu dans les WSI. Ensuite, une méthode

Méthode	Segmentation	Classification	Apprentissage	Base d'apprentissage
MITOS-RCNN [Rao 2018]	MITOS-RCNN basé sur faster-RCNN		Fine tuning VGGNet-16	ICPR12 MITOS-ATYPIA-14 AMIDA13
[Li *et al.* 2018b]	Lightweight R-CNN		Initialisation aléatoire	ICPR12 MITOS-ATYPIA-14
[Chen *et al.* 2016b]	DRN		Fine Tuning a partir de [Chen *et al.* 2015]	ICPR12
[Wollmann & Rohr 2017b]	DRN + Hough voting		Initialisation aléatoire	AMIDA13
AggNet [Albarqouni *et al.* 2016]	-	Multi-scale CNN	Initialisation aléatoire	AMIDA13
MFF-CNN [Kausar *et al.* 2018]	Blue-ratio	MFF-CNN	Fine tuning à partir du modèle caffeNet	MITOS-ATYPIA-14
[Wahab *et al.* 2017]	Blue ratio + global binary threesholding	CNN	Initialisation aléatoire	ICPR12 TAUPAC16
MSSN [Ma *et al.* 2018]	-	CNN	Initialisation aléatoire	ICPR12 MITOS-ATYPIA-14
[Akram *et al.* 2018]	-	CNN	Initialisation aléatoire	MITOS-ATYPIA-14 TAUPAC16
Wide resNet [Zerhouni *et al.* 2017]	CNN (WideResNet)		Initialisation aléatoire	ICPR12 MITOS-ATYPIA-14 TAUPAC16
L-view [Paeng *et al.* 2017]	Otsu's method + binary dilatation	-CNN (L-view basé sur les blocs résiduels) -SVM pour le score de la tumeur	Initialisation aléatoire	TAUPAC16
[Wollmann & Rohr 2017a]	Blue-ration + thresholding	- DRN + Hough transform -Arbre de décision pour le score de la tumeur	Initialisation aléatoire	TAUPAC16

TABLE 5.3 – *Les méthodes d'apprentissage profond proposées pour la détection de la mitose (2).*

Method	Segmentation	Classification	Training	Dataset
[Veta *et al.* 2016]	Max pooling	CNN	Initialisation aléatoire	AMIDA13 Une base d'apprentissage de deux laboratoires de pathologie dans Netherlands [Al-Janabi *et al.* 2013]
[Pezzotti *et al.* 2017]	Max pooling	CNN	Initialisation aléatoire	AMIDA13
[Tellez *et al.* 2018]	Canaux marron et bleu	CNN	Initialisation aléatoire	TAUPAC16 Une base d'apprentissage de trois laboratoires de pathologie dans Netherlands
[Shah *et al.* 2017]	La méthode Otsu's + binary dilatation	MitosNet (variant de CNN)	Initialisation aléatoire	Une base d'apprentissage de trois centres de pathologie internationaux

TABLE 5.4 – *Les méthodes d'apprentissage profond proposées pour la détection de la mitose (3).*

de segmentation est appliquée selon différents mécanismes : seuillage dynamique global et fixe [Wang *et al.* 2014], l'algorithme de regroupement k-means [Albayrak & Bilgin 2016], seuil de couleur et grid search [Malon & Cosatto 2013], l'algorithme krill held (KHA) [Beevi *et al.* 2019], méthode de seuillage d'otsu [Das & Dutta 2019, Paeng *et al.* 2017], et

seuillage global binaire [Wahab *et al.* 2017]. Enfin, les réseaux DL sont exploités en classification selon différents cas d'utilisation : l'apprentissage à partir des initialisations aléatoires [Janowczyk & Madabhushi 2016, Das & Dutta 2019], l'apprentissage transféré [Chen *et al.* 2016a] ou fine tuning [Wu *et al.* 2017, Kausar *et al.* 2018], l'utilisation des CNN comme des modules d'extraction des caractéristiques [Albayrak & Bilgin 2016, Beevi *et al.* 2019], et la combinaison entre les attributs CNN et les handcrafted features [Wang *et al.* 2014, Malon & Cosatto 2013, Saha *et al.* 2018].

Apprentissage à partir des initialisations aléatoires et fine tuning

[Janowczyk & Madabhushi 2016] ont effectué un apprentissage sur le réseau AlexNet (version cifar-10) en se basant sur des patchs générés sous un grossissement de 20x. L'inconvénient majeur de cette approche est le faible grossissement qui peut être une source d'incertitude pour le réseau CNN. Dans une autre étude, [Das & Dutta 2019] ont évalué un CNN peu profond sur les sous-patchs décomposés par la méthode de décomposition Haar wavelet. Malgré l'efficacité de ces approches, l'apprentissage à partir des initialisations aléatoires est couteux en temps d'exécution. En plus, il peut causer des problèmes de sur-apprentissage à cause des volumes limités de données. Afin de résoudre ces limitations, les méthodes d'apprentissage transféré et de fine tuning ont été exploitées dans plusieurs travaux [Wu *et al.* 2017, Kausar *et al.* 2018]. Dans ce cadre, des modèles pré-entrainés sur la base d'apprentissage ImageNet sont adaptés à la nouvelle tâche de classification. Pour plus de détails sur les méthodes d'apprentissage transféré et d'ajustement, le chapitre 2 explique leurs principes.

[Wu *et al.* 2017] ont réajusté leur réseau FF-CNN à partir du modèle pré-entrainé AlexNet. FF-CNN fusionne entre plusieurs caractéristiques à multi-niveaux en reliant les sorties des couches de convolution Conv3 et Conv4 à la couche entièrement connectée. Leur approche a prouvé son efficacité par rapport à la meilleure méthode proposée dans le défi ICPR2014. Cela prouve l'efficacité des stratégies d'ajustement dans la résolution du problème de détection de la mitose. En plus, ces stratégies ont d'autres avantages liés à leurs exigences réduites en termes de calcul, où un processeur standard est suffisant pour compléter la tâche d'apprentissage.

Extraction des caractéristiques par le réseau CNN

Dans d'autres contributions, les CNN ont été utilisées comme des modules d'extraction de caractéristiques. [Albayrak & Bilgin 2016] ont exploité le réseau CNN pour l'extraction des caractéristiques, LDA et PCA pour la sélection des attributs et l'algorithme SVM pour la classification. Afin de réduire la complexité temporelle dans la phase d'extraction des caractéristiques, d'autres travaux suggèrent l'utilisation des modèles réajustés à partir des modèles pré-entrainé. Dans ce cadre, [Beevi *et al.* 2019] ont réajusté les 4 dernières couches de convolution du modèle Caffe VGGNet.

Les techniques ML et DL ont prouvé leur efficacité dans la résolution du problème de détection de la mitose. Les méthodes ML sont ba-

sées sur des attributs CNN ou des handcrafted features. Pour prendre avantage de ces deux techniques, plusieurs travaux ont combiné entre ces deux catégories d'attributs. [Wang *et al.* 2014] ont proposé une approche en cascade (HC + CNN), où ils ont généré séparément des modèles à base des attributs CNN et des handcrafted features. Ensuite, ils ont utilisé un troisième classificateur en cas de confusion entre la décision des deux classificateurs. La classe finale est calculée en fonction de la moyenne des décisions des différents modèles. Dans une autre contribution, [Malon & Cosatto 2013] ont combiné les caractéristiques nucléaires (texture, couleur et forme) et les caractéristiques extraites par le réseau CNN (LeNet 5 [LeCun *et al.* 1998]). [Saha *et al.* 2018] ont incorporé 24 attributs dans la première couche entièrement connectée du réseau CNN. Ces recherches ont prouvé l'efficacité de l'hybridation par rapport à l'utilisation séparée des attributs CNN ou des handcrafted features.

FCN et les méthodes d'apprentissage profond en détection

Afin de réduire le temps considérable d'inférence, d'autres travaux ont suggéré l'exploitation du réseau FCN [Long *et al.* 2015] pour la segmentation. [Chen *et al.* 2016a] ont proposé une méthode hybride basée sur le réseau FCN pour l'extraction des mitoses candidates et un modèle caffeNet réajusté pour la classification. Cette méthode a réduit le temps d'inférence de 8 minutes [Cireşan *et al.* 2013] à 0.5 s par HPF. D'autres travaux proposent de convertir les modèles DL obtenus à des FCN pour accélérer le processus de détection [Wu *et al.* 2017, Akram *et al.* 2018, Paeng *et al.* 2017].

Néanmoins, [Li *et al.* 2018a] ont critiqué l'utilisation du réseau FCN, car il ignore les informations régionales. Pour améliorer le processus de détection, ils ont exploité pour la première fois un réseau DL de détection dans la tâche de détection de la mitose. L'infrastructure hybride proposée (Deepmitosis) est composée d'un réseau de détection profond (DeepDet), de vérification (DeepVer) et de segmentation (DeepSeg). Le composant principal de cette infrastructure est le réseau DeepDet. Ce réseau localise les mitoses à base du réseau faster R-CNN [Ren *et al.* 2015]. Une autre étude réalisée par [Rao 2018] propose une nouvelle variante du réseau faster R-CNN (MITOS-RCNN) qui est désignée pour la détection des petits objets. Dans une autre contribution, [Li *et al.* 2018b] ont développé un lightweight region-based CNN inspiré du réseau R-CNN [Girshick *et al.* 2014], où l'objectif principal était de proposer un système de détections rapide.

Les réseaux de régression

La modélisation du problème de détection de mitose en tant qu'un problème de régression est une autre stratégie qui permet d'adapter le temps de détection à l'utilisation clinique [Chen *et al.* 2016b, Wollmann & Rohr 2017b, Wollmann & Rohr 2017a]. [Chen *et al.* 2016b] ont proposé une méthode basée sur le réseau de régression profond

(DRN) avec fully convolutional kernels. Ce réseau est composé des couches convolutives (CL) et déconvolutives (DL). Les CL effectuent la phase de sous-échantillonnage pour l'extraction des caractéristiques, tandis que les DL sont utilisées pour restaurer la taille originale en entrée. Pour éviter le problème de sur-apprentissage, ils ont réajusté le modèle pré-entrainé deepLab « off-the-self » [Everingham *et al.* 2010]. Dans une autre contribution, [Wollmann & Rohr 2017b] ont combiné entre le réseau résiduel profond et la méthode de vote Hough. Cette méthode permet de réduire le temps de calcul par rapport aux autres méthodes d'apprentissage ensemblistes car le processus d'apprentissage est effectué une seule fois.

Apprentissage multi-échelles

Les travaux cités précédemment ont suggéré d'entrainer les réseaux DL sur des images générées sous une seule échelle. D'autre part, les informations contextuelles sont importantes, car le pathologiste peut analyser les lames histopathologiques sous différents grossissements. Pour une détection plus précise, d'autres travaux ont exploité les techniques d'apprentissage multi-échelles. [Albarqouni *et al.* 2016] ont proposé l'architecture AggNet basée sur un réseau CNN à plusieurs échelles et une couche d'agrégation. Dans le domaine biomédical, cette étude était la première expérience qui utilise les réseaux CNN dans la génération des vraies étiquettes à partir des annotations des non experts dans la foule. Dans une autre contribution, [Kausar *et al.* 2018] ont développé un modèle FCNN à plusieurs échelles (MFF-CNN) qui est basé sur deux modules : FF-CNN [Wu *et al.* 2017] et une couche de fusion.

Les stratégies d'apprentissage en deux phases

Les mitoses sont caractérisées par leur faible fréquence, et cela peut biaiser la nature des bases d'apprentissage conçues pour la classification. Par exemple, [Cireşan *et al.* 2013] ont généré une base composée seulement de 6.6 % de pixels de mitoses. Par conséquent, cela peut entraîner une base déséquilibrée et un taux élevé des faux positifs (FP). Pour résoudre ces limitations, différentes approches ont été proposées [Wahab *et al.* 2017, Ma *et al.* 2018, Akram *et al.* 2018, Li *et al.* 2018a, Janowczyk & Madabhushi 2016]. Ces méthodes exploitent les stratégies d'apprentissage en deux phases.

[Wahab *et al.* 2017] ont proposé une méthode en deux phases basée sur les CNN. Dans la première étape, le CNN classifie les candidates non mitoses en facile, normale et dure. Ensuite, les candidates mitoses et non-mitoses dures sont augmentées par des rotations et des retournements, tandis que les non-mitoses faciles sont sous-échantillonnées par la méthode blue ratio histogram-based clustering. Ensuite, la base d'apprentissage générée est ré-entrainé par un deuxième CNN.

Dans une autre étude, [Ma *et al.* 2018] ont proposé une autre méthode en deux phases. Premièrement, un réseau d'apprentissage multi-scale and

similarity learning convnets (MSSN) a été utilisé pour traiter le problème des FN. Ensuite, ils ont entrainé un autre modèle de prédiction de similarité pour réduire le taux élevé de faux positifs (FP).

Dans une autre contribution, [Akram *et al.* 2018] ont proposé un algorithme DL auto-supervisé. Premièrement, ils ont entrainé un CNN sur les deux ensembles : BG-rand et FG-Lab qui contiennent des échantillons de fond et des patchs centrés sur la mitose. Ensuite, les échantillons faux positifs (FP) détectés (BG-hard) ont été exploités avec l'ensemble FG-WSI pour ré-entrainer le modèle CNN. Ce travail a analysé l'effet de l'apprentissage semi-supervisé par l'utilisation des patchs de mitose extraits à partir des données non étiquetées (FG-WSI).

Dans une autre étude, [Li *et al.* 2018a] ont développé un réseau DeepVer pour vérifier les faux positifs fournis par le réseau DeepDet. Malgré l'efficacité de ce système hybride, les résultats obtenus montrent que le modèle DeepVer n'a pas amélioré les performances de la base d'apprentissage ICPR12. Pour réduire le taux des FP, [Zerhouni *et al.* 2017] ont proposé l'exploitation des fonctions de fitness pondérées. Dans cette étude, ils ont utilisé le réseau wide residual dans une stratégie de classification par pixel. Pour enrichir leur base d'apprentissage, ils ont fusionné entre plusieurs bases hétérogènes : ICPR12, MITOS-ATYPIA-14 et TAUPAC16-auxiliaire.

La détection à partir des WSI

Les études citées précédemment se limitent à la détection des mitoses à partir des ROI. Cependant, pour l'exploitation des modèles générés, le pathologiste doit sélectionner manuellement les ROI des données de test à partir des WSI. Afin d'automatiser le processus complet de détection, la sélection manuelle des ROI doit être automatisée.

Contrairement aux défis : CIPR12, AMIDA13 et MITOS-ATYPIA-14, le TUPAC16 a exploré la prédiction directe du score de prolifération à partir des WSI. La disponibilité de cette base d'apprentissage a encouragé de nombreux chercheurs à proposer des méthodes automatiques pour la prédiction directe du score de prolifération tumorale [Paeng *et al.* 2017, Wollmann & Rohr 2017a]. Ces méthodes sont basées sur trois étapes principales : l'extraction des ROI, la détection de la mitose, et la prévision du score de prolifération tumorale.

[Paeng *et al.* 2017] ont utilisé la méthode d'Otsu et la dilatation binaire pour l'extraction des ROI, où les patchs extraits représentent un carré de 10 HPF consécutifs. Ensuite, le réseau L-view a été entrainé sur les régions caractérisées par une densité élevée de cellules. Enfin, ils ont prédit le score de prolifération tumorale en se basant sur le nombre des mitoses détectées, 21 attributs supplémentaires, et l'algorithme d'apprentissage SVM. Le meilleur résultat obtenu dans le défi TAUPAC16 prouve l'efficacité de cette méthode. Dans une autre contribution, [Wollmann & Rohr 2017a] ont exploité le mécanisme de threshold-based attention pour l'extraction des ROI. Ensuite, ils ont utilisé un réseau DNN associé à la méthode Hough transform pour la détection des mitoses. Enfin, ils ont entrainé l'algo-

rithme d'apprentissage arbre de décision sur les résultats obtenus afin de calculer le nombre des mitoses.

5.2.5 Discussion

Résultats

La tableau 5.5 compare entre les résultats obtenus par les méthodes proposées en détection de la mitose en termes de rappel (R), de précision (P), et de f-mesure ou taux des bien classés (F1 / Acc).

Base d'apprentissage	Méthode	précision	rappel	F-mesure/ Taux des biens classés
ICPR12	[Das & Dutta 2019]	0.8446	0.8365	0.8405
	DeepMitosis [Li *et al.* 2018a]	0.854	0.812	0.832
	[Wahab *et al.* 2017]	0.83	0.76	0.79
	[Chen *et al.* 2016b]	0.779	0.802	0.79
	[Li *et al.* 2018b]	0.78	0.79	0.784
	IDSIA [Cireşan *et al.* 2013]	0.88	0.70	0.782
	MSSN [Ma *et al.* 2018]	0.776	0.787	0.781
	HC+CNN [Wang *et al.* 2014]	0.84	0.65	0.7345
	[Malon & Cosatto 2013]	0.747	0.590	0.659
	CasNN [Chen *et al.* 2016a]	0.460	0.507	0.482
AMIDA13	[Cireşan *et al.* 2013]	0.610	0.612	0.611
	[Wollmann & Rohr 2017b]	0.547	0.686	0.609
	AggNet [Albarqouni *et al.* 2016]	0.441	0.424	0.433
MITOS-ATYPIA-14	[Das & Dutta 2019]	0.9964	0.987	0.9812
	[Albayrak & Bilgin 2016]	-	-	Acc 0.968
	[Beevi *et al.* 2019]	0.8739	0.9013	0.8860
	CasNN [Chen *et al.* 2016a]	0.804	0.772	0.788
	MSSN [Ma *et al.* 2018]	0.379	0.617	0.470
	DeepMitosis [Li *et al.* 2018a]	0.431	0.443	0.437
	MFF-CNN [Kausar *et al.* 2018]	0.405	0.453	0.428
	[Li *et al.* 2018b]	0.40	0.45	0.427
	FF-CNN [Wu *et al.* 2017]	-	-	0.393
TUPAC16-auxiliary	[Wahab *et al.* 2017]	0.57	0.53	0.55
	[Paeng *et al.* 2017]	-	-	0.652
	[Tellez *et al.* 2018]	-	-	0.480
ICPR12 + MITOS-ATYPIA-14 + AMIDA13	[Saha *et al.* 2018]	0.92	0.88	0.90
	MITOS-RCNN [Rao 2018]	-	-	0.955
ICPR12 + MITOS-ATYPIA-14 + TAUPAC16-auxiliaire	[Zerhouni *et al.* 2017]	-	-	0.648
MITOS-ATYPIA-14 + TAUPAC16-auxiliaire	[Akram *et al.* 2018]	0.613	0.671	0.640

TABLE 5.5 – *Les résultats obtenus par les méthodes d'apprentissage profond proposées pour la détection de la mitose.*

Les résultats obtenus sur la base d'apprentissage ICPR12 indiquent l'efficacité du réseau Faster RCNN en détection [Li *et al.* 2018a]. En plus, les CNN [Das & Dutta 2019] sont plus performant par rapport à leur hybridation avec les méthodes d'apprentissage classiques

[Wang *et al.* 2014] ou leur exploitation en tant qu'extracteurs de caractéristiques [Malon & Cosatto 2013]. Cependant, leur utilisation dans une stratégie de classification par pixel est trop coûteuse en termes de temps d'inférence. La solution optimale consiste à effectuer la sélection appropriée parmi plusieurs paramètres : le type d'architecture, la stratégie d'apprentissage (par pixel ou patch) et les hyper-paramètres du réseau.

Les résultats montrent aussi que malgré le temps d'inférence rapide la méthode CasNN [Chen *et al.* 2016a], elle est moins performante par rapport aux autres méthodes . Cela peut être justifié par les limites du réseau FCN dans l'inférence de l'emplacement des mitoses.

D'autre part, peu d'études ont examiné les méthodes DL sur la base d'apprentissage AMIDA13. Les meilleurs résultats ont été obtenus par le réseau max pooling CNN en termes de f-mesure [Cireşan *et al.* 2013] et par la méthode proposée par [Wollmann & Rohr 2017b] en termes de rappel. Tandis que AggNet [Albarqouni *et al.* 2016] a atteint un faible taux de f-mesure par rapport aux autres méthodes. Cela peut être justifié par les annotations bruyantes des non-experts dans la foule.

Contrairement aux résultats obtenus sur la base d'apprentissage ICPR12, les résultats obtenus sur MITOS-ATYPIA-14 montrent l'efficacité des méthodes DL exploitées en tant qu'extracteurs de caractéristiques [Beevi *et al.* 2019, Albayrak & Bilgin 2016] et la méthode CasNN [Chen *et al.* 2016a] par rapport aux autres approches [Ma *et al.* 2018, Li *et al.* 2018a, Kausar *et al.* 2018, Li *et al.* 2018b, Wu *et al.* 2017].

Les résultats présentés par [Albayrak & Bilgin 2016] prouvent l'efficacité de leur méthode, où ils ont réussi à améliorer la performance de 0.786 à 0.969 par leur stratégie de sélection de caractéristiques. Cependant, le nombre des caractéristiques sélectionnées (10) pour distinguer la morphologie complexe des mitoses doit être analysé.

Les résultats obtenus sur les bases d'apprentissage ICPR12, AMIDA13 et TUPAC16-auxiliaire valident le problème de sur apprentissage sur les données prévenant d'un seul laboratoire. Par exemple, l'annotation de AMIDA13 par divers pathologistes et la collecte de la base d'apprentissage TUPAC16-auxiliaire de différents laboratoires justifient leur faible précision par rapport à la base ICPR12.

Pour éviter le problème de sur-apprentissage, d'autres travaux [Rao 2018, Zerhouni *et al.* 2017, Akram *et al.* 2018] ont combiné entre plusieurs bases d'apprentissage. Des résultats remarquables ont été obtenus par [Saha *et al.* 2018] et [Rao 2018] en combinant les bases d'apprentissage ICPR12, MITOS-ATYPIA-14, et AMIDA13. [Saha *et al.* 2018] ont amélioré les performances de leur système DL par 14 % par l'inclusion de 24 attributs supplémentaires. Cependant, l'importance des handcrafted features n'est pas validée dans d'autres travaux. Cela peut être attribué à la nature des caractéristiques sélectionnées et des architectures DL exploitées.

Le tableau 5.6 compare entre les résultats obtenues sur la base d'apprentissage TAUPAC16. Les meilleurs résultats ont été obtenus par [Paeng *et al.* 2017] en termes du score quadratic weighted cohen's kappa.

Base d'apprentissage	Méthode	Le score Quadratic weighted Cohen's kappa (K)	Le coefficient Spearmans correlation
TAUPAC16	[Paeng *et al.* 2017]	0.567 [0.464, 0.671]	0.617 [0.581 0.651]
	[Tellez *et al.* 2018]	0.471 [0.340, 0.603]	0.519 [0.477, 0.559]
	[Wollmann & Rohr 2017a]	0.42	-

TABLE 5.6 – *Les résultats obtenus par les méthodes d'apprentissage profond sur la base d'apprentissage the TAUPAC16.*

Temps de calcul et matériel

Le tableau 5.7 résume la capacité des GPU exploitées et le temps de traitement des méthodes DL en détection de la mitose. Des GPU puissantes étaient utilisées [Wu *et al.* 2017, Rao 2018] et parallélisées pour accélérer le temps d'apprentissage et d'inférence.

Ce tableau illustre le temps de calcul élevé de la méthode de classification par pixel [Cireşan *et al.* 2013] par rapport aux autres approches [Chen *et al.* 2016a]. Pour une comparaison équitable, nous avons regroupé ces méthodes par type de GPU. Certains travaux utilisent seulement des CPU [Wang *et al.* 2014, Wahab *et al.* 2017] en raison de leurs exigences restreintes liées aux CNN peu profonds et à la taille limitée des bases d'apprentissage (ICPR12). Cependant, le temps d'inférence obtenu par [Wang *et al.* 2014] (1.5 min par HPF) n'est pas pratique pour une utilisation clinique en temps réel, et cela valide l'importance des GPU.

La version du réseau DRN proposée par [Chen *et al.* 2016b] est 6 fois plus lente par rapport à la version utilisée par [Wollmann & Rohr 2017b] sur une GPU moins puissante. Cela peut être justifié par l'influence des hyper-paramètres tels que la taille du patch sur le temps de traitement. Le temps de calcul optimal a été obtenu sur les GPU parallèles ($< 0.5s$) en raison de leur traitement distribué [Wu *et al.* 2017, Kausar *et al.* 2018, Rao 2018].

La complexité de la base d'apprentissage est un autre paramètre important qui influence le temps de traitement. Dans ce cadre, nous avons remarqué qu'il existe une différence remarquable entre le temps d'inférence [Wollmann & Rohr 2017b] sur les images de la base d'apprentissage ICPR12 et la base d'apprentissage TUPAC16. Le temps d'inférence considérable obtenu par [Wollmann & Rohr 2017a] est justifié par la stratégie de classification de bout en bout sur des WSI (50000×50000) au lieu des ROI.

5.2.6 Défis et perspectives

Plusieurs techniques de type DL ont été proposées dans l'état de l'art pour résoudre les problèmes liés à la tâche de détection de la mitose, où différentes bases d'apprentissage ont été publiées (ICPR12, AMIDA13, MITOS-ATYPIA-14, TAUPAC16) pour encourager la recherche dans ce domaine. Malgré les efforts faits, le nombre limité des figures mitotiques (1552 au maximum) présente l'un des problèmes principales, et cela limite l'efficacité des modèles DL générés à distinguer les mitoses des autres structures. En plus, ces bases d'apprentissage sont collectées à partir d'un maximum de trois centres de pathologie. Par conséquent, cela limite l'effi-

GPU	Référence	Temps d'apprentissage	Temps d'inférence
GPU	[Cireşan *et al.* 2013]	Un jour pour chaque réseau	8 minutes
	[Li *et al.* 2018b]	-	6.93 s
Sans GPU	[Wang *et al.* 2014]	11.4 h	1.5 min
	ICPR12 [Wahab *et al.* 2017]	15 h	48 s
Nvidia GeForce GTX 750M	[Albarqouni *et al.* 2016]	-	-
Nvidia Getforce GTX 970	[Wollmann & Rohr 2017b]	2.5 jours	2.5 s
Nvidia GeForce GTX titan X	[Chen *et al.* 2016b]	-	15 s
	[Li *et al.* 2018a]	-	0.4 s to 0.7 s
	[Chen *et al.* 2016a]	-	0.5 s 0.3 s
	TAUPAC16 [Wahab *et al.* 2017]	30 h	1 min
Nvidia Quadro K4200 graphics processor	[Das & Dutta 2019]	-	16 s
4 Nvidia Tesla M40 GPUs	[Wu *et al.* 2017]	-	0.375s
	[Kausar *et al.* 2018]	-	0.388 s
5 NVIDIA tesla K80 GPUs pour l'apprentissage et une seule GPU pour le test	[Rao 2018]		0.5 s
-	[Ronneberger *et al.* 2015]	-	5 min (WSI)
-	[Karpathy *et al.* 2014]	-	0.3 s

TABLE 5.7 – *La temps de calcul et le matériel utilisé dans les méthodes proposées pour la détection de la mitose.*

cacité de ces modèles en généralisation. La solution à ce problème consiste à apprendre tout type de variance. Certains travaux ont suggéré l'utilisation des techniques de normalisation des couleurs, tandis que plusieurs d'autres ont ignoré cette étape importante sur des données collectées à partir de plusieurs centres [Wahab *et al.* 2017, Wollmann & Rohr 2017a].

Le nombre restreint des échantillons d'images dans les bases d'apprentissage médicales n'est pas lié à leurs disponibilités, mais plutôt à la charge de travail considérable pour leur annotation par les spécialistes en pathologie. La technique de crowdsourcing est parmi les solutions proposées pour résoudre le problème de manque de données annotées. L'étude présentée par Albarqouni et al [Albarqouni *et al.* 2016] est parmi les premiers travaux sur l'exploitation du crowdsourcing dans la tâche de détection de la mitose. Malgré l'efficacité de ces techniques dans d'autres domaines, leur utilisation dans le domaine médical est délicate à cause des classes bruitées fournies par les non experts. Dans ce cadre, plus d'efforts doivent être effectués dans la phase de contrôle et de validation des annotations afin d'obtenir des résultats plus robustes.

Les techniques de fine tuning et d'apprentissage transféré ont été proposées pour résoudre le problème de sur-apprentissage qui est lié au manque de données annotées. Ces techniques représentent 21% des ar-

ticles sélectionnés. Dans ces travaux, des modèles entrainés sur la base d'apprentissage ImageNet ont été réutilisés et réajustés sur les bases de détection de la mitose. Le succès de ces méthodes est justifié par la similitude des caractéristiques de bas niveau (arêtes, angles... etc.). Cependant, dans ce cadre, il n'existe aucun principe théorique et beaucoup de questions se pose concernant la relation entre ces deux domaines hétérogènes. Par conséquent, le partage de modèles entrainés sur des bases du même domaine peut être plus utile en raison de la similarité entre les images histopathologiques par rapport aux autres domaines.

L'apprentissage semi-supervisé présente une autre solution au manque de données annotées. Cette technique permet d'entrainer des modèles sur des données étiquetées et d'autres non étiquetés. Cette spécificité encourage l'exploitation de cette technique comme alternative aux méthodes d'apprentissage supervisées dans le cas des quantités limitées de données classifiées. Jusqu'à présent, les travaux de détection de mitose proposés se focalise sur les techniques d'apprentissage supervisées plus que les méthodes d'apprentissage semi-supervisées [Akram *et al.* 2018]. Par conséquent, l'exploitation de ces méthodes dans les travaux futurs est parmi les perspectives suggérées.

D'autres travaux proposent l'exploitation des réseaux peu profonds pour éviter les problèmes de sur-apprentissage. Ces réseaux sont caractérisés par un nombre limité de couches. Tandis que peu de recherches ont exploité les architectures profondes basées sur les techniques de réduction de paramètres, comme : inception [Szegedy *et al.* 2015], MobileNet [Howard *et al.* 2017, Sandler *et al.* 2018],et suffleNet [Zhang *et al.* 2018].

Les résultats obtenus dans l'état de l'art prouvent l'efficacité des méthodes proposées, en particulier sur les bases d'apprentissage ICPR12 et MITOS-ATYPIA-14. Dans ces bases, plusieurs échantillons d'images appartenant aux bases d'apprentissage et de test ont été collectés à partir de la même source. Ainsi, l'efficacité des modèles générés sur les échantillons provenant de différentes sources n'est pas garantie. L'analyse effectuée par [Veta *et al.* 2016] prouve le degré élevé de désaccord entre les décisions des pathologistes et les prédictions des méthodes automatiques lorsqu'ils sont évalués sur une nouvelle base d'apprentissage hétérogène. Pour cela, ces modèles doivent être validés sur de nouveaux échantillons de données et testés par les pathologistes.

L'efficacité des résultats des pathologistes était justifiée par leur stratégie d'analyse descendante qui permet de préserver plus d'informations contextuelles. Cela est lié à la taille du patch dans les méthodes automatisées. Comme solution, l'apprentissage à multi-échelle a été exploité pour répondre au manque des informations contextuelles [Wu *et al.* 2017]. Dans ce cadre, nous avons remarqué que la majorité des travaux proposés en détection de la mitose sont basé sur un apprentissage à une seule échèle, et donc l'exploitation des réseaux multi-échelles comme multi-scale contextual networks [Wang *et al.* 2017b] peut être prometteuse.

Dans la majorité des méthodes DL proposées, les architectures étaient présentées sous forme d'une boîte noire, dans laquelle aucune stratégie claire n'a été définie concernant le choix du nombre des couches et des valeurs des hyper-paramètres. Le système analytique visuel progressif (Deepeyes) [Pezzotti *et al.* 2017] prouve l'importance de la visualisation dans

l'identification des filtres ou des couches inutiles. Par conséquent, cela peut constituer un bon outil d'analyse des futures architectures proposées dans le cadre de détection de mitose.

En conclusion, les travaux présentés dans cette synthèse identifient l'efficacité des méthodes DL dans la détection de la mitose. Néanmoins, ils sont toujours associés à quelques limitations, et cela encourage leur exploitation en tant qu'un second outil d'aide au pathologiste plutôt que leur exploitation directe dans des utilisations cliniques.

5.3 Les méthodes de régularisation en apprentissage profond

5.3.1 Les méthodes ensemblistes

La conception d'un bon modèle en généralisation est parmi les enjeux principaux en apprentissage profond. La complexité de ce processus est liée essentiellement au problème de sur-apprentissage.

Les réseaux de neurones sont caractérisés par leur processus d'apprentissage stochastique. Généralement, l'apprentissage commence par l'initialisation aléatoire des poids. Ensuite, ces poids sont ajustés afin de réduire le taux d'erreur. L'utilisation de différentes initialisations aléatoires permet de générer différents modèles sur la même base d'apprentissage. Le nombre élevé des poids et la complexité des ANN peuvent piéger la convergence des poids vers une combinaison optimale qui est adaptée aux données d'apprentissage et des poids initiaux. D'autre part, la bonne performance sur les données d'apprentissage n'est pas considérée comme un critère d'évaluation final à cause de la sensibilité des ANN aux changements des données. Ces réseaux sont caractérisés par une variance élevée entre les performances des données d'apprentissage et de test, et cela est défini par le problème de sur-apprentissage. Afin de réduire cette variance, les méthodes de régularisation sont recommandées, comme : la régularisation par abandon [Veta *et al.* 2016] et les méthodes ensemblistes [Ju *et al.* 2018].

Les méthodes ensemblistes sont des techniques basées sur la combinaison de plusieurs modèles. Elles permettent de réduire l'erreur de généralisation et la grande variance entre les résultats d'apprentissage et de test. En plus, elles améliorent la performance finale dans certaines situations. Le but principal de ces méthodes est de regrouper un ensemble de modèles faibles pour former un modèle plus fort.

Les ANN sont caractérisés par leur instabilité et dépendance de plusieurs conditions initiales, comme : les poids initialisés aléatoirement et le bruit dans la base d'apprentissage. L'idée des méthodes ensemblistes est d'exploiter les prédictions de différents modèles afin de réduire leur sensibilité et d'assurer la stabilité des prédictions faites.

Il existe plusieurs méthodes qui permettent de générer un ensemble de modèles, comme la variation dans la base d'apprentissage, la variation dans les conditions initiales, la variation dans l'architecture du réseau, et la variation dans la technique de combinaison.

La stratégie de variation dans la base d'apprentissage est basée sur plusieurs techniques comme le rééchantillonnage de la base d'apprentissage

avec (bootstrap [Reed *et al.* 2014]) ou sans remplacement. Dans cette thèse, nous avons proposé une technique de variation sans remplacement. Elle consiste à allouer des portions aléatoires (90 %) de la base d'apprentissage pour chaque modèle généré.

Dans le cadre de vision par ordinateur et exactement en traitement des images histopathologiques, il est possible d'exploiter plusieurs techniques pour générer différentes versions de la base d'apprentissage, comme : la variation dans les méthodes d'augmentation de données, de normalisation des couleurs, ou dans la résolution des images en entrée. Par exemple, [Xu *et al.* 2017] ont proposé un multi-resolution convolutional network (MR-CN-PV) pour le score le l'atypie nucléaire. L'approche proposée est basée sur un processus de vote entre 3 CNN entrainés sur des images à différentes résolutions.

La technique de variation dans les conditions initiales consiste à entrainer le même modèle sur le même espace de données en variant dans les valeurs des paramètres ou des hyper-paramètres, comme : la méthode d'initialisation aléatoire des poids, la méthode d'apprentissage, l'optimiseur, la valeur du taux d'apprentissage... etc. Ensuite, les modèles générés sont combinés par les techniques de moyenne ou de vote.

La technique de moyenne non pondérée a été largement exploitée dans les méthodes proposées dans la compétition ImageNet. Par exemple, [Krizhevsky *et al.* 2012] ont réussi à améliorer le taux d'erreur des top-5 de 18.2 % à 16.4 % par la combinaison de 5 CNN similaires. Dans une autre contribution, [Szegedy *et al.* 2015] ont amélioré le taux d'erreur de 7.89 % à 6.67 % par la combinaison de 7 réseaux de même configuration incluant une version plus large.

D'autre part, d'autres travaux ont proposé de combiner entre plusieurs modèles caractérisés par différentes configurations ou architectures. Par exemple, dans la compétition ImageNet, [Simonyan & Zisserman 2014b] ont combiné entre les deux meilleurs modèles (les configurations D et E) pour réduire le taux d'erreur à 7.0 %. [Zeiler & Fergus 2014] ont combiné entre 6 modèles de différentes configurations pour améliorer le taux d'erreur sur la base de test par 1.6 %. Ces modèles diffèrent dans le nombre des neurones des couches entièrement connectées. En histopathologie, [Chen *et al.* 2016a] ont utilisé la même stratégie, où ils ont combiné entre 3 CNN dont le nombre des neurones au niveau des FC est 1024-256-2, 1024-512-2, 512-256-2 respectivement. Dans une autre contribution plus récente, [Nanni *et al.* 2019a] ont combiné entre plusieurs modèles pré-entrainés de type CNN. Ces modèles ont été réajustés sur des bases d'apprentissage histopathologiques. Dans ce cadre, ils ont exploité les CNN réajustés en tant qu'extracteurs de caractéristiques. Ensuite, ils ont combiné entre les caractéristiques extraites. Enfin, ils ont entrainé le réseau SVM sur les résultats obtenus.

[Ju *et al.* 2018] ont critiqué l'utilisation de la méthode de combinaison par moyenne non pondérée. Cette méthode est influencée par les mauvais modèles appartenant au groupe (weak learners). En plus, elle est plus adaptée aux réseaux caractérisés par une structure et des performances similaires, et elle est sensible à la présence des modèles biaisés par rapport aux autres composants du groupe. Dans ce cadre, [Ju *et al.* 2018] ont présenté une étude comparative entre les performances de quatre méthodes

ensemblistes connues : vote majoritaire, classificateur bayésien optimal, la généralisation empilée, et super learner, où ils ont exploité les réseaux CNN comme des modules de base. Dans cette étude, ils ont utilisé des ensembles de différente nature : (a) des réseaux de même architecture enregistrés dans différents points d'apprentissage (training checkpoints), (b) le même réseau entrainé à plusieurs reprises, (c) des réseaux de différentes structures, (d) des modèles très confiants (over confident), (e) les mauvais modèles, (f) tous les réseaux entrainés précédemment. En résumé, les résultats obtenus ont prouvé l'efficacité de la méthode super learner par rapport aux autres méthodes ensemblistes. Cette méthode attribue un poids à chaque classificateur de base. Ensuite, ces poids sont regroupés dans une couche de convolution et ajustés par un processus d'apprentissage sur la base de validation. D'autre part, les résultats ont montré l'efficacité de la méthode de combinaison par moyenne non pondérée par rapport au vote majoritaire. Cependant, cette méthode est vulnérable aux mauvais modèles et sensible aux modèles très confiants. Par conséquent, l'étape de sélection des modèles à combiner est une étape très délicate et nécessite une attention considérable. Dans ce cadre, nous avons proposé une nouvelle méthode de sélection dynamique qui exploite la métaheuristique optimisation par essaim de particules (PSO) dans la phase de sélection.

Les méthodes de combinaison par vote et moyenne non pondérée sont parmi les méthodes simples qui ont été largement exploitées en apprentissage profond. D'autre part, il existe d'autres méthodes ensemblistes plus complexes, comme : le stacking et le boosting.

Dans la technique ensembliste de boosting, un nouveau modèle est ajouté dans chaque itération pour corriger les erreurs des modèles précédents. Dans le cadre de l'apprentissage profond, peu d'études ont examiné la technique de boosting en vision par ordinateur à cause de sa complexité élevée en termes de calcul. Par exemple, [Mosca & Magoulas 2017] ont proposé la méthode deep incremental boosting (DIB) qui est basée sur Adaboost et la technique de l'apprentissage transféré. Dans la première itération, ils ont commencé par la phase d'apprentissage du réseau CNN. Ensuite, dans le reste des itérations, ils ont transféré les couches du réseau CNN de l'itération précédente et ils ont rajouté une couche de convolution supplémentaire au nouveau réseau. Le but de l'apprentissage transféré est de réduire le temps de traitement considérable de la méthode boosting et d'éviter le problème de sur-apprentissage des réseaux CNN.

Généralement, les méthodes ensemblistes citées précédemment nécessitent d'effectuer plusieurs apprentissages afin de générer les modèles de base appartenant à l'ensemble. D'autre part, la complexité élevée de calcul et les exigences en termes de ressources sont parmi les problèmes principaux des réseaux DL (tableau 5.8). Ces caractéristiques ont limité l'utilisation des méthodes ensemblistes en apprentissage profond. Afin de résoudre ces limitations, une solution simple consiste à exploiter le processus itératif d'apprentissage des ANN. Cette stratégie permet de produire un ensemble de modèles dans un seul processus d'apprentissage, où les modèles sont enregistrés dans différents points d'apprentissage. Cette méthode a été exploitée dans plusieurs domaines, comme : les systèmes de traduction [Sennrich *et al.* 2016, Vaswani *et al.* 2017], la génération des résumés [Kobayashi 2018], la détection des programmes mal-

veillants [Sang *et al.* 2018a], la classification des images [Ju *et al.* 2018], la segmentation des images médicales [Fok *et al.* 2018, Jung *et al.* 2018], la reconnaissance des émotions faciales [Sang *et al.* 2018b], et l'étiquetage des vidéos à grande échelle [Skalic *et al.* 2017]. Le tableau 5.9 résume les architectures DNN utilisées dans les méthodes ensemblistes à base des points d'apprentissage (checkpoints).

Réseau	Temps	Matériel
AlexNet [Krizhevsky *et al.* 2012]	Cinq à six jours	Deux GPUs NVIDIA GTX580 3GB
ZFNet [Zeiler & Fergus 2014]	12 jours	Une seule GPU NVIDIA GTX580
Inception [Szegedy *et al.* 2015]	Une semaine (estimation)	Peu de GPU haut de gamme
VGGNet [Simonyan & Zisserman 2014b]	2–3 semaines selon l'architecture.	Quatre GPUs NVIDIA Titan Black
Xception [Chollet 2017]	3 jours	60 NVIDIA K80 GPUs

TABLE 5.8 – *Les exigences matérielles et le temps d'exécution pour l'apprentissage des réseaux de neurones convolutifs sur la base d'apprentissage ImageNet.*

Référence	Architecture
[Chen *et al.* 2017a]	MLP CNN Long LSTM
[Sennrich *et al.* 2016]	Réseau Encoder−decoder
[Vaswani *et al.* 2017]	Réseau Transformer basé sur le réseau encoder-decoder
[Ju *et al.* 2018]	NIN, VGGNet, ResNet
[Sang *et al.* 2018b]	DenseNet
[Skalic *et al.* 2017]	Mixture of Neural-Network Experts (MoNN) LSTM GRU
[Kobayashi 2018]	LSTM encoder−decoder
[Fok *et al.* 2018]	ResNet34
[Sang *et al.* 2018a]	RNSALL basé sur le modèle ResNet

TABLE 5.9 – *Les architectures DNN précédemment combinées à base de plusieurs points d'apprentissage.*

Différentes stratégies ont été exploitées afin de sélectionner les points d'apprentissage appropriés. Par exemple, [Chen *et al.* 2017a] ont combiné entre les trois meilleurs modèles. Ils ont prouvé l'efficacité de la moyenne des prédictions par rapport à la moyenne des poids. De même, [Fok *et al.* 2018] ont combiné entre les meilleurs 2 à 5 modèles et [Sang *et al.* 2018a] ont combiné entre les 25 meilleurs modèles. D'autres travaux ont suggéré la combinaison des modèles enregistrés dans les

derniers points d'apprentissage [Sennrich *et al.* 2016, Vaswani *et al.* 2017, Ju *et al.* 2018]. Par exemple, [Sennrich *et al.* 2016] ont proposé de combiner entre les 4 derniers modèles enregistrés dans chaque 30 000 mini-batch. Dans une autre contribution, [Vaswani *et al.* 2017] ont combiné entre les derniers 5 et 20 modèles enregistrés dans des intervalles de 10 minutes. D'autres travaux proposent de combiner les modèles générés dans les dernières époques [Sang *et al.* 2018b, Sang *et al.* 2018b].

Dans le cadre de la combinaison des modèles enregistrés dans plusieurs points d'apprentissage, nous avons implémenté une méthode qui combine entre plusieurs modèles MobileNet enregistrés dans des intervalles de 3 minutes. Nous avons exploité les techniques de vote majoritaire et de moyenne non pondérée pour combiner entre ces modèles. En plus, nous avons comparé entre les deux méthodes statiques de sélection du sous ensemble : les N derniers modèles et les N meilleurs modèles.

5.3.2 L'apprentissage transféré

Dans les dernières années, les CNN ont connu un intérêt considérable en vision par ordinateur. Dans ce cadre, différentes architectures optimisées ont été proposées pour réduire les problèmes de sur-apprentissage [Simonyan & Zisserman 2014b, Szegedy *et al.* 2015]. Le nombre des paramètres de ces architectures est considéré comme un facteur clé qui influence la performance du réseau. Un nombre important de paramètres exige une quantité élevée de données pour éviter les problèmes de sur-apprentissage. Cela illustre les inconvénients des CNN sur les données médicales à cause du nombre limité des images et la difficulté de leur annotation. En plus, l'apprentissage à partir des initialisations aléatoires est exigeant en termes de temps de traitement et de capacité de calcul. Pour résoudre ces limitations, des travaux récents ont prouvé l'efficacité de la stratégie de fine tuning et de l'exploitation des CNN comme des modules d'extraction de caractéristiques sur les petits volumes de données [Tajbakhsh *et al.* 2016]. Le chapitre 2 détaille le principe de la technique d'apprentissage transféré.

Plusieurs efforts ont été faits pour mesurer la transférabilité entre différentes bases d'apprentissage. Par exemple, [Yosinski *et al.* 2014] ont démontré l'efficacité de l'apprentissage transféré entre les bases d'apprentissage similaires par rapport aux bases distantes en se basant sur deux bases d'apprentissage extraites de la base imageNet. Tandis que, [Azizpour *et al.* 2015] ont remarqué une amélioration dans les performances lors d'un apprentissage transféré entre des tâches distantes. Leur contribution a étudié la transférabilité des réseaux entrainés sur les bases d'apprentissage ImageNet et Places à d'autres tâches de reconnaissance visuelle. L'objectif principal de leur analyse était d'identifier l'effet de différents facteurs sur l'apprentissage transféré, comme : la profondeur du réseau, l'arrêt prématuré, la nature des classes, et la taille des données sources. Leur étude expérimentale indique que l'augmentation de la largeur, de la profondeur, et de la taille a un effet positif sur les tâches les plus proches. Dans une autre contribution, [Chopra *et al.* 2013] ont exploité l'apprentissage transféré pour l'adaptation du domaine entre des bases d'apprentissage similaires. Dans ce cadre, les échantillons sources

ont été collectés à partir d'Amazon et les images cibles de Dslr et Webcam. Le but principal était d'améliorer la généralisation sur les nouvelles situations.

Dans le domaine de l'analyse des images médicales, les modèles entrainés sur la base d'apprentissage ImageNet ont connu un grand intérêt. L'objectif des travaux proposés était de transférer la connaissance d'un domaine non médical (ImageNet) vers un autre domaine médical, comme : la classification du cancer du sein [Ferreira *et al.* 2018, Zhi *et al.* 2017, Vesal *et al.* 2018, Khan *et al.* 2019, Mehra *et al.* 2018], la classification du cancer colorectal [Mehra *et al.* 2018], et la détection des polypes et des embolies pulmonaires [Tajbakhsh *et al.* 2016]. La majorité de ces travaux ont comparé l'apprentissage à partir des initialisations aléatoires, fine-tuning, et l'exploitation des CNN comme des modules d'extraction des caractéristiques. Par exemple, [Shin *et al.* 2017] ont montré l'efficacité de l'apprentissage à partir des initialisations aléatoires et de l'apprentissage transféré par rapport à l'exploitation des CNN comme des modules d'extraction des caractéristiques. Dans une autre contribution, [Malik *et al.* 2019] ont prouvé l'efficacité de l'apprentissage à partir des initialisations aléatoires en détection et de la technique de fine-tuning en classification, où ils ont proposé une approche adaptative inspirée du modèle inceptionV3. Dans une autre recherche, [Mehra *et al.* 2018] ont prouvé l'importance de l'apprentissage transféré par rapport à l'apprentissage à partir des initialisations aléatoires en se basant sur 3 architectures de type CNN : VGG16, VGG19, et ResNet50. Les modèles générés ont été utilisés comme des modules d'extraction de caractéristiques. De même, l'étude présentée par [Tajbakhsh *et al.* 2016] a démontré l'efficacité du réajustement profond par rapport à l'apprentissage à partir des initialisations aléatoires. D'autre part, ils ont remarqué l'efficacité de l'apprentissage à partir des initialisations aléatoires par rapport au réajustement peu profond. Leurs résultats valident l'hypothèse sur le rapport entre la profondeur de réajustement et la distance entre les tâches source et cible.

Plusieurs stratégies ont été exploitées pour réajuster les modèles entrainés sur ImageNet. Par exemple, [Ferreira *et al.* 2018] ont fixé les premières 678 couches du modèle pré-entrainé ResnetV2. Ensuite, le modèle généré a été réajusté sur la tâche cible. De même, [Vesal *et al.* 2018] ont exploité les modèles pre-entrainés Inception-V3 et ResNet50, où ils ont montré l'efficacité du modèle ResNet50 par rapport à Inception-V3. Dans une autre contribution, [Zhi *et al.* 2017] ont proposé une architecture basée les 6 premières couches de l'architecture VGGNet. Leur étude expérimentale identifie l'efficacité de l'apprentissage transféré de l'architecture proposée par rapport à la version originale de VGGNet.

Dans cette thése, nous avons exploité les techniques d'apprentissage transféré dans les trois contributions expérimentales proposées, notamment la dernière, qui est une nouvelle méthode d'apprentissage transféré entre les bases d'apprentissage histopathologiques.

5.4 Conclusion

Dans ce chapitre, Nous avons présenté et comparé les méthodes DL proposées en histopathologie, et l'état de l'art des contributions proposées dans cette thèse.

La première section est une synthèse sur les travaux DL en détection de la mitose à partir des images histopathologiques du cancer du sein. Dans ce cadre, nous avons présenté et comparé entre les méthodes DL proposées. Malgré le nombre important des travaux de recherches, plusieurs défis restent à relever, comme : la collecte des bases d'apprentissage à partir d'un maximum de centres de pathologie, la nécessité de validation des nouvelles annotations par des experts dans le domaine afin de créer des systèmes robustes, et la résolution des problèmes liés à la perte de l'information contextuelle.

La deuxième section présente l'état de l'art des travaux qui exploitent les stratégies des méthodes ensemblistes et d'apprentissage transféré en apprentissage profond. L'étude effectuée a mené à plusieurs conclusions, comme : les inconvénients des méthodes de sélection statique en apprentissage ensembliste, et les différentes questions liées à l'apprentissage transféré à partir des modèles entrainés sur ImageNet. Toutes ces limitations ont fait l'objet des contributions proposées dans cette thèse. Le chapitre suivant détaille les trois contributions expérimentales proposées.

Contributions de la thèse 6

Sommaire

Les méthodes d'apprentissage profond sont caractérisées par leur capacité d'extraction de caractéristiques par rapport aux algorithmes d'apprentissage automatique classiques. Cependant, ces méthodes ont un problème de sur-apprentissage sur les volumes limités de données, car ces derniers exigent un grand volume de données pour ajuster équitablement les paramètres du réseau. En plus, ils sont caractérisés par leur variance élevée et biais faible.

En histopathologie, les WSD ont facilité la collecte des données par la numérisation des lames de verre à des WSI. Malgré leur disponibilité, le nombre des échantillons collectés reste limité pour les applications de type DL. En plus, l'annotation des centaines à des milliers de données collectées exige un temps considérable et une grande charge de travail. Pour éviter les problèmes de sur-apprentissage et pour réduire la variance élevée des réseaux DL, certains travaux ont suggéré l'utilisation des techniques d'apprentissage transféré, de régularisation, et des méthodes ensemblistes.

Dans cette thèse, nous nous intéressons à la résolution des différentes limitations citées auparavant en classification des images histopathologiques. Dans ce cadre, nous avons proposé trois travaux de recherche [Dif & Elberrichi 2020d, Dif & Elberrichi 2020c, Dif & Elberrichi 2020b]. Les deux premières contributions exploitent les techniques de régularisation, d'apprentissage transféré, et des méthodes ensemblistes [Dif & Elberrichi 2020c, Dif & Elberrichi 2020b]. La dernière contribution [Dif & Elberrichi 2020d] est une nouvelle méthode de fine tuning entre les bases d'apprentissage histopathologiques.

La première contribution [Dif & Elberrichi 2020b] est un framework de régularisation pour la classification des images histopathologiques. Dans ce travail de recherche, nous avons proposé un framework basé sur différentes techniques de régularisation : l'augmentation des données, les petits (small) modèles (MobileNetV1, MobileNetV2), la sélection de la méthode d'optimisation appropriée en apprentissage (SGD, RmsProp), et les méthodes ensemblistes.

Dans le cadre des techniques ensemblistes, nous avons exploité deux stratégies de sélection statiques (meilleurs et derniers modèles) afin de sélectionner les modèles à combiner à partir de plusieurs points d'apprentissage. Ensuite, nous avons combiné les modèles sélectionnés par les techniques de vote et de moyenne non pondérée. Malgré les résultats prometteurs dans cette contribution, la diversité et la bonne coopération entre ces modèles ne sont pas assurées, car, ces modèles peuvent avoir presque les mêmes erreurs et donc leur combinaison ne conduit pas toujours aux meilleurs résultats. Pour résoudre ces problèmes, nous avons proposé une nouvelle stratégie de sélection basée sur la métaheuristique optimisation par essaim de particules (PSO) [Dif & Elberrichi 2020c]. Cette stratégie prend en considération la qualité du sous ensemble sélectionné au lieu de considérer séparément la qualité de chaque modèle appartenant à cet ensemble.

Dans le deuxième travail de recherche [Dif & Elberrichi 2020c], l'objectif principal était de répondre à deux problématiques : le sur-apprentissage et les inconvénients des méthodes de sélection statique. Dans ce cadre, nous avons proposé une méthode de sélection dynamique basée sur les techniques d'apprentissage transféré et la métaheuristique PSO. Premiè-

rement, cette technique génère un ensemble de modèles par la méthode d'apprentissage transféré. Ensuite, la métaheuristique optimisation par essaim de particules (PSO) sélectionne un sous ensemble de modèles pertinents à partir de l'ensemble de modèles générés. Enfin, ces modèles sont combinés par un vote majoritaire ou par une moyenne non pondérée.

Les deux premières contributions exploitent les techniques d'apprentissage transféré et les méthodes ensemblistes. Dans ce cadre, le nombre de modèles à combiner peut engendrer un problème de capacité de stockage dans certaines situations, par exemple : les systèmes de vision intégrés. En plus, la méthode d'apprentissage transféré précédemment utilisée se base sur les modèles pré-entrainés sur la base ImageNet. Malgré le succès de cette technique dans l'état de l'art, il n'existe pas de principes théoriques sur le fonctionnement interne de cette stratégie et beaucoup de questions se posent sur la relation entre la base ImageNet et les bases d'apprentissage histopathologiques. Pour cela, nous avons proposé une méthode qui se base sur l'apprentissage transféré entre des bases d'apprentissage de même domaine au lieu de transférer la connaissance à partir de la base ImageNet [Dif & Elberrichi 2020d].

Enfin, le troisième travail de recherche [Dif & Elberrichi 2020d] présente une nouvelle stratégie de fine-tuning pour l'analyse des images histopathologiques. Contrairement aux solutions précédemment proposées, où les modèles entrainés sur la base d'apprentissage ImageNet sont réutilisés pour la classification d'une nouvelle tâche, cette étude propose d'effectuer l'apprentissage transféré à partir des modèles précédemment entrainés sur des bases d'apprentissage histopathologiques. L'objectif principal est d'exploiter les hypothèses liées à l'efficacité de l'apprentissage transféré entre les bases non distantes et d'examiner pour la première fois ces suggestions sur les images histopathologiques.

Les expérimentations ont été effectuées sur un ordinateur portable équipé d'une carte graphique de type NVIDIA GeForce GTX 1060 (6 GB) et la librairie d'apprentissage profond Keras.

6.1 Un framework de régularisation pour la classification des images histopathologiques à l'aide des réseaux de neurones convolutifs.

6.1.1 Résumé

Les méthodes d'apprentissage profond sont caractérisées par leur capacité d'extraction de caractéristiques par rapport aux algorithmes d'apprentissage automatique classiques. Cependant, ces méthodes ont un problème de sur-apprentissage sur les volumes limités de données. En plus, elles sont caractérisées par leur variance élevée et biais faibles.

L'objectif de cette contribution est de réduire ces problèmes par l'optimisation des modèles entrainés en termes de généralisation. Dans ce cadre, nous avons proposé un framework basé sur différentes techniques de régularisation : l'augmentation des données, les petits (small) modèles (MobileNetV1, MobileNetV2), la sélection de la méthode d'optimisation appropriée en apprentissage (SGD, RmsProp), et les méthodes ensemblistes. Dans le cadre des techniques ensemblistes, nous avons exploité deux stratégies de sélection statiques (meilleurs et derniers modèles) afin de sélectionner les modèles à combiner à partir de plusieurs points d'apprentissage. Ensuite, nous avons combiné les modèles sélectionnés par les techniques de vote et de moyenne non pondérée.

Afin d'évaluer le framework proposé, nous l'avons testé sur la base d'apprentissage histopathologique lymphoma. Les résultats obtenus confirment l'efficacité du modèle MobileNetV2 et la méthode d'optimisation SGD en termes de généralisation. Les meilleurs résultats ont été obtenus par la combinaison des trois meilleurs modèles.

En résumé, ces résultats prouvent l'efficacité de la technique de combinaison des modèles générés à partir de plusieurs points d'apprentissage, en classification des images histopathologiques.

6.1.2 Problématique

En histopathologie, le diagnostic des biopsies est une tâche difficile et nécessite des années d'expérience, et cela peut causer une grande variance entre les diagnostics des pathologistes. Afin de réduire cette variabilité, les systèmes d'aide au diagnostic sont largement utilisés. Ces systèmes sont généralement basés sur les méthodes ML et DL.

Dans les chapitres précédents, nous avons largement discuté les avantages et les apports des méthodes DL par rapport aux méthodes ML classiques, ainsi, dans cette thèse nous nous intéressons à l'exploitation des techniques DL pour la classification des images histopathologiques. Cela est justifié par la haute résolution des WSI numérisés par les WSD. Cependant, malgré les avantages des réseaux DL, le nombre limité des images histopathologiques et la difficulté de leur annotation peuvent causer des problèmes de sur-apprentissage. En plus, ces stratégies sont caractérisées par leur variance élevée. Afin de résoudre ces problèmes, plusieurs travaux ont suggéré l'exploitation des méthodes de régularisation, comme la régularisation par abandon [Srivastava *et al.* 2014] et les méthodes ensemblistes [Ju *et al.* 2018].

6.1.3 Motivation

L'objectif principal de cette contribution est d'améliorer l'efficacité des réseaux de neurones convolutifs en terme de généralisation, pour la classification des sous types de lymphomes. Dans ce cadre, nous avons proposé un framework basé sur plusieurs méthodes de régularisation : l'augmentation de données, les petits (small) modèles, la sélection de la méthode d'optimisation appropriée en apprentissage, et les méthodes ensemblistes.

Augmentation de données

Cette technique permet d'augmenter le nombre des images en entrée afin d'éviter les problèmes de sur-apprentissage et de s'échapper aux attaques contradictoires. En plus, l'entrainement des réseaux CNN sur un volume important et varié de données permet d'améliorer leur efficacité en généralisation.

Exploitation des petits (small) modèles

Les réseaux DL sont caractérisés par leurs problèmes de sur-apprentissage à cause du nombre élevé de paramètres. L'utilisation des réseaux peu profonds est parmi les techniques exploitées pour éviter les problèmes de sur-apprentissage. Ces réseaux sont caractérisés par un nombre réduit de paramètres par rapport aux réseaux profonds. D'autre part, les réseaux profonds sont connus par leur efficacité dans la résolution des problèmes non linéaire. Dans ce cadre, plusieurs travaux ont proposé des réseaux profonds basés sur des techniques de réduction de dimensionnalité, où ils ont remplacé les modules de convolution classiques par d'autres modules qui exploitent les techniques de réduction de dimensionnalité, comme : les modules d'Inception [Szegedy *et al.* 2015] et les depthwise separable convolutions [Howard *et al.* 2017]. Le but principal de ces contributions était de proposer des architectures moins couteuses en termes de capacité de stockage et de calcul, afin de réduire le temps d'inférence et d'adapter les modèles générés aux restrictions des systèmes de vision intégrés.

Dans cette contribution, nous avons utilisé les architectures MobileNetV1 et MobileNetV2 pour éviter les problèmes de sur-apprentissage sur les volumes limités des données médicales.

Méthode d'optimisation

L'apprentissage en MLP consiste à ajuster l'ensemble des paramètres θ (poids w_{ij} et biais b) afin de minimiser une fonction du coût C. Pour optimiser ces paramètres, l'algorithme de rétropropagation du gradient est exploité pour ajuster ces paramètres à base de différent optimiseurs, comme : SGD [Robbins & Monro 1951], Adam [Kingma & Ba 2015], RmsProp [Tieleman & Hinton 2012], et Adagrad [Duchi *et al.* 2011]. Dans ce cadre, plusieurs travaux ont essayé d'optimiser la méthode SGD. D'autre part, la nature stochastique de ces méthodes n'assure pas l'efficacité d'une seule méthode par rapport aux autres. Dans ce cadre, nous avons testé deux méthodes d'optimisations (SGD et RmsProp) et comparé leurs

graphes de convergences sur les données d'apprentissage, de validation et de test. Le but de cette comparaison est de sélectionner le modèle le plus efficace en généralisation pour la tâche traité dans cette contribution.

Méthode ensembliste

Les méthodes ensemblistes sont généralement utilisées afin de combiner la décision de plusieurs modèles. Cette combinaison permet d'améliorer la généralisation et dans certains cas la performance du groupe. Dans ce cadre, nous avons exploité les méthodes ensemblistes afin de réduire la grande variance entre les performances sur les données de test et d'apprentissage et donc réduire les risques de sur-apprentissage. Cette contribution utilise les méthodes de sélection statiques à partir d'un ensemble de modèles enregistrés dans plusieurs points d'apprentissage. Cette technique combine entre plusieurs modèles générés dans un seul apprentissage, et donc réduit le temps de traitement par rapport aux autres méthodes qui exigent d'effectuer plusieurs apprentissages.

Cette étude présente la première évaluation de la méthode ensembliste présentée ci-dessus dans le cadre de classification des images histopathologiques par les réseaux MobileNet.

6.1.4 Etat de l'art des méthodes proposées pour la classification de la base d'apprentissage lymphoma

Les lymphomes sont des tumeurs qui affectent les lymphocytes (cellules T ou B) [Orlov *et al.* 2010]. Ils sont catégorisés en lymphomes non hodgkiniens (NHL) et lymphomes hodgkin (HL). Les NHL présentent un taux de 90 % des lymphomes [Orlov *et al.* 2010] et ils sont catégorisés en différents sous-types, comme : le lymphome diffus à grandes cellules B (DLBCL) qui représente la forme la plus connue, la leucémie lymphoïde chronique (CLL), le lymphome folliculaire (FL), et le lymphome à cellules du manteau (MCL). Ces sous-types représentent des NHL agressifs (DLBCL, MCL) ou indolents (CLL, FL), où les lymphomes agressifs progressent rapidement par rapport aux indolents.

Plusieurs travaux dans l'état de l'art ont proposés d'automatiser la segmentation [Tosta *et al.* 2017] et la classification des NHL à partir des images histopathologiques. Dans ce cadre, [Shamir *et al.* 2008] ont proposé 9 bases d'apprentissage biologiques. Ces bases sont composées d'un ensemble d'images dont leur taille varie de 25×25 à 1388×1040. L'objectif majeur de cette contribution était d'encourager les experts en vision par ordinateur à proposer des méthodes performantes pour la classification des images biologiques. L'étude expérimentale basée sur l'approche de distance de voisinage pondérée (WND-CHARM) a montré la difficulté de la classification automatique des sous types des NHL. Cette complexité a encouragé la communauté de vision par ordinateur à développer d'autres méthodes plus robustes pour résoudre cette tâche.

Le tableau 6.1 résume les travaux liés à la classification automatique des sous types de NHL. Les méthodes proposées sont de deux types ML et DL. Précédemment, la majorité des travaux étaient basés sur les méthodes ML. Dans ce type d'applications, la phase d'extraction des caractéristiques

a connu un intérêt considérable à cause de la morphologie complexe des images histopathologiques, où différentes stratégies d'extraction d'attributs ont été proposées. Par exemple, [Orlov *et al.* 2010] ont utilisé la méthode de transformation basée sur les attributs globaux. Premièrement, ils ont transformé les pixels bruts en plaines spectrales. Ensuite, ils ont extrait différentes caractéristiques globales : texture, polynôme, et autres caractéristiques statistiques. Enfin, ils ont exploité différentes méthodes de sélection d'attributs afin de sélectionner les attributs discriminants : analyse discriminante de fisher (FLD), minimum redundancy maximum relevance (mRmR), et transformation de fisher (F/C).

Method	Article	Classifier	Feature extration	Feature selection
ML	[Shamir *et al.* 2008]	WND-CHARM		-
	[Meng *et al.* 2010]	C-RSPM + WMVA	Caractéristiques de couleur et de texture	Chi-square
	[Orlov *et al.* 2010]	WND	Caractéristiques globales (texture, polynôme et statistiques)	FLD, mRmR, F/C.
	[Di Ruberto *et al.* 2015]	SVM	GLCM modifié	-
	[Nava *et al.* 2016]	KFDA	DOMs	RELIEF
	[Song *et al.* 2016]	SVM	Descripteurs de texture visuels (IFV, LBP, HOG, GIST, CENTRIST)	SDT
	[Tosta *et al.* 2017]	SVM	Segmentation with GA	-
ML + DL	[Codella *et al.* 2016]	SVMs Non linéaire	Caractéristiques de bas niveau (histogramme des couleurs LBP, ... , Un CNN transféré à partir d'un modèle ImageNet)	-
	[Song *et al.* 2017a]	SVM	VGG-VD pré-entrainé	CFV
	[Song *et al.* 2017b]	SVM	Caractéristiques locales (SHIFT, VGG-VD pré-entrainé)	SDR
	[Bai *et al.* 2019]	Random forest	Caractéristique de texture et statistiques GoogLeNet pré-entrainé	-
	[Nanni *et al.* 2018]	SVM	LBP, Représentations de texture et caractéristiques statistiques. Caractéristiques extraites par des CNNs.	-
DL	[Janowczyk & Madabhushi 2016]	AlexNet (La version de Cifar-10)		-

TABLE 6.1 – *Les méthodes proposées dans l'état de l'art pour la classification des soustypes de lymphomes.*

Dans une autre contribution, [Meng *et al.* 2010] ont développé une approche basée sur collateral representative subspace projection modeling (C-RSPM). Premièrement, ils ont divisé chaque image en 25 blocs au total. Ensuite, ils ont extrait un ensemble de 505 attributs (couleur et texture) à partir de chaque bloc. Enfin, ils ont sélectionné 50 attributs pertinents à partir de l'ensemble des attributs par la méthode de sélection d'attributs Chi-square. Dans une autre étude, [Di Ruberto *et al.* 2015] ont adapté trois modèles de texture en niveau de gris à l'extraction des caractéristiques à partir des images colorées des lymphomes. L'étude présentée par [Nava *et al.* 2016] propose la combinaison des moments orthogonaux discrets (DOMs) des images histopathologiques prétraitées par la méthode de déconvolution de couleur. Dans une autre contribution,

[Song *et al.* 2016] ont suggéré l'exploitation de la méthode de subcategory discriminant transform (SDT) sur les attributs extraits afin de minimiser la variance intra-classes et optimiser la différence entre les classes.

Le tableau 6.1 résument les différentes méthodes d'extraction de caractéristiques utilisées. Selon l'étude comparative, il est difficile de définir la méthode d'extraction pertinente en raison de leur nature stochastique. En plus, ces caractéristiques sont sélectionnées en fonction de la tâche traitée. Afin de standardiser la phase d'extraction de caractéristiques, plusieurs travaux ont suggéré l'utilisation des réseaux CNN comme des modules d'extraction de caractéristiques. Par exemple, [Codella *et al.* 2016] ont proposé une approche d'apprentissage visuel en plusieurs étapes, où ils ont combiné entre des caractéristiques de bas niveaux et des caractéristiques extraites par des CNN pré-entrainés. Dans d'autres contributions, [Song *et al.* 2017a] et [Song *et al.* 2017b] ont proposé une méthode basée sur les attributs extraits par un Convnet-based FV (C-FV). Ensuite, comme extension à cette méthode, ils ont combiné le vecteur de caractéristiques résultant avec d'autres types d'attributs locales.

Les approches citées ci dessus ont proposé la combinaison des attributs extraits par différentes méthodes, ensuite, les attributs résultants sont classifiés par un seul algorithme d'apprentissage. D'autres approches suggèrent de classifier séparément les sous ensembles des attributs extraits, ensuite, les modèles résultants sont combinés par les méthodes ensemblistes. Dans ce cadre, [Bai *et al.* 2019] ont combiné entre le classificateur random forest entrainé sur les caractéristiques texturales et statistiques et le classificateur softmax entrainé sur les attributs extraits par le modèle Inception.

Dans l'état actuel de nos connaissances, peu de travaux ont utilisé les méthodes DL en classification des sous-types de lymphomes. Dans ce cadre, [Janowczyk & Madabhushi 2016] étaient les premiers qui ont entrainé un réseau CNN à partir des initialisations aléatoires.

6.1.5 La méthode proposée

La figure 6.1 présente le schéma général du framework proposé. Ce schéma est composé de 3 modules principaux : prétraitement, apprentissage et méthodes ensemblistes. Dans ce qui suit, nous détaillerons chaque module séparément.

Prétraitement

La Figure 6.2 illustre le schéma général de prétraitement. Ce schéma est basé sur différentes techniques de prétraitement : normalisation moyenne, extraction des patchs, rotations, et corps aléatoires.

Les images colorées sont représentées par un ensemble de pixels sous forme d'une matrice, où chaque matrice est associée à un canal de couleur (rouge, vert, et bleu). Les valeurs des pixels sont des entiers compris entre 0 et 255. La standardisation des pixels à des valeurs comprises entre 0 et 1 est une étape importante, car les grandes valeurs de pixels peuvent perturber ou ralentir le processus d'apprentissage des réseaux DL si ces derniers sont associés à des poids de petites valeurs. La méthode la plus

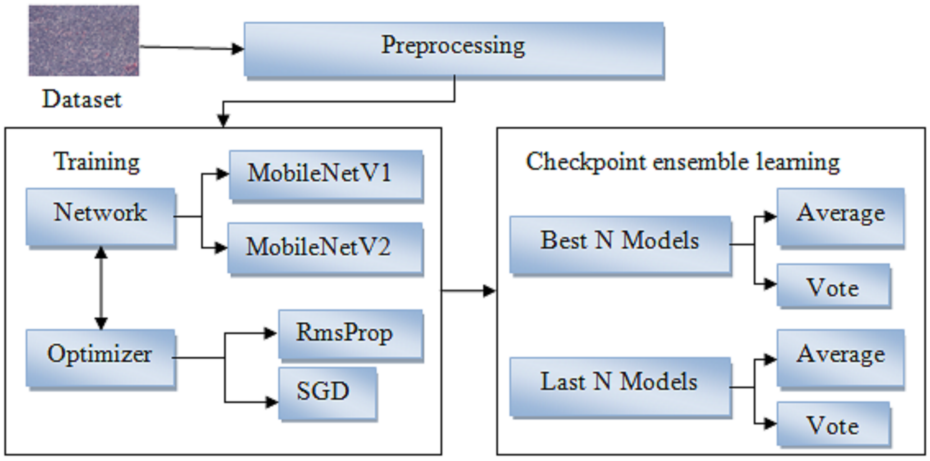

FIGURE 6.1 – *Les composants du framework proposé.*

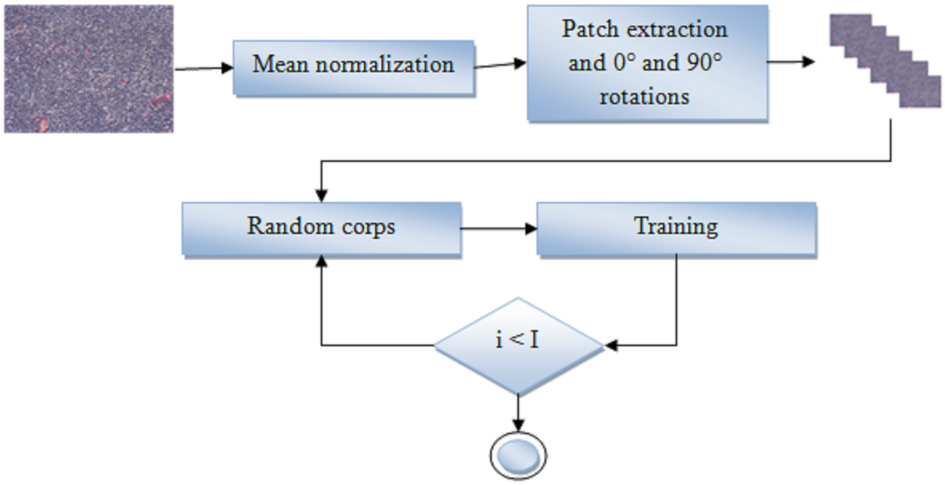

FIGURE 6.2 – *Le schéma de prétraitement.*

simple de normalisation divise chaque pixel par la valeur 225. Dans cette contribution, nous avons exploité une autre méthode de normalisation. Cette méthode consiste à soustraire la moyenne des valeurs des pixels de l'ensemble des pixels. Elle permet de centrer la distribution des valeurs des pixels autour de 0.

La méthode de centrage nécessite de calculer la valeur moyenne des pixels. Cela peut être réalisé en suivant différentes stratégies : par images, par mini batch, par base d'apprentissage. Dans cette étude, nous avons exploité la technique de centrage par base d'apprentissage. Premièrement, nous avons calculé la moyenne des images appartenant à la base d'apprentissage. Ensuite, nous avons soustrait cette moyenne de toutes les autres images. L'équation 6.1 présente le processus de normalisation, où $m_{i,j}$ est la valeur de pixel de l'image m et N est le nombre des images.

$$m_{i,j}(t) = m_{i,j}(t-1) - \frac{\sum_{k=1}^{N} m_{i,j}^{(k)}(t-1)}{N} \tag{6.1}$$

La deuxième étape de prétraitement consiste à augmenter le nombre des images en entrée par la méthode d'extraction des patchs (Figure 6.3)

et par la technique de rotation. Les images histopathologiques sont caractérisées par une haute résolution, par exemple, les images appartenant à la base d'apprentissage lymphoma ont une taille de 1388×1040.

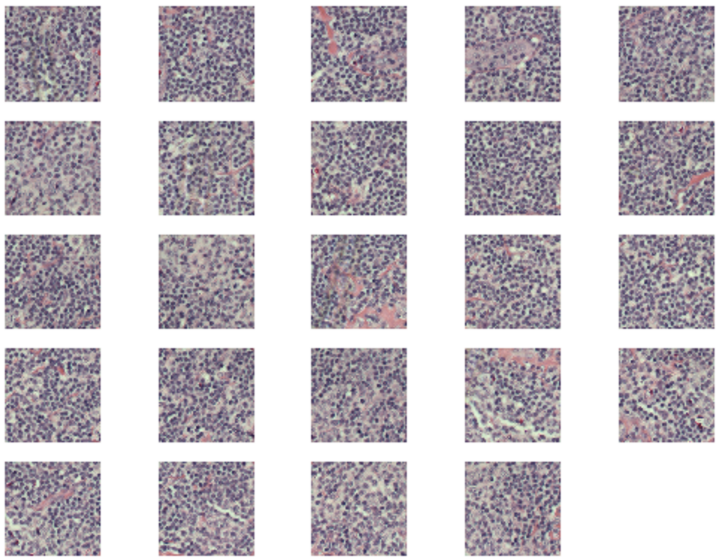

FIGURE 6.3 – *Le résultat d'extraction des patchs par la méthode de fenêtre coulissante (Lymphoma).*

L'exploitation directe de ces images peut causer une augmentation exponentielle dans le nombre de paramètres du réseau, et cela accroît les exigences des réseaux entrainés en termes de stockage et de capacité de calcul. En plus, les réseaux composés d'un nombre important de paramètres ont plus de risque de sur-apprentissage. Afin de résoudre ces limitations et d'éviter le problème de sur-apprentissage sur les volumes limités des images histopathologiques, nous avons exploité la méthode de fenêtre coulissante, cette méthode permet d'extraire un nombre important de patchs à partir d'une seule image.

Dans cette contribution, nous avons choisi de diviser chaque image à des patchs non superposés de taille 144×144. La grande différence entre la taille du patch et l'image originale peut causer une perte d'information dans le cas des images naturelles, car le cadre du patch doit être centré autour de l'objet d'intérêt. Tandis que, dans le cas des images histopathologiques, la distribution régulière de certaines formes biologiques de base autorise l'exploitation des patchs de petite taille.

Après l'étape d'extraction, nous avons pivoté les patchs par des rotations de 0 et 90 degrés. La rotation est parmi les méthodes d'augmentation de données qui ont été largement exploitées. Elle permet d'améliorer le processus d'analyse de ces images, car le pathologiste observe les biopsies sous différents angles. En plus, elle protège les modèles générés des attaques contradictoires.

Enfin, nous avons extrait des échantillons aléatoires de taille 128×128 durant l'apprentissage. Cette technique assure la sélection de différentes

images dans le processus de génération des lots, car elle génère des images distinctes à partir d'une seule image durant différentes itérations. Le but principal de cette méthode est d'améliorer la généralisation durant le processus d'apprentissage.

Apprentissage et méthodes ensemblistes

Le temps de traitement et l'espace mémoire sont deux paramètres importants dans le déploiement réel des systèmes de vision par ordinateur. Ces deux paramètres présentent des défis majeurs dans le cas des réseaux CNN. D'autre part, ces réseaux risquent les problèmes de sur-apprentissage sur les volumes limités de données. Afin de résoudre ces limitations, l' exploitation des petits (small) modèles est parmi les solutions proposées [Howard *et al.* 2017].

Dans cette contribution, nous avons utilisé les réseaux MobileNetV1 et MobileNetV2. Ces deux réseaux ont été conçus pour les appareils mobiles et les systèmes de vision intégrés, et ils sont basés sur deux types de modules de base : DSC et IRLB, respectivement. Le but principal de ces deux modules est de diminuer le nombre total des paramètres par des techniques de réduction de dimensionnalité. En plus de ces deux modules, les réseaux MobileNet utilisent les deux hyper-paramètres : le multiplicateur de largeur et le multiplicateur de résolution afin de réduire le nombre des paramètres et le coût de calcul, respectivement.

Pour plus de détails sur les architectures MobileNetV1 et MobileNetV2, le chapitre 2 explique leurs principes. D'après l'étude comparative entre différentes architectures (tableau2.1), nous avons constaté que le nombre de paramètres des réseaux MobileNet est réduit par rapport aux autres. Par exemple, le réseau MobileNet contient 28 couches au total et 4.2 millions de paramètres, tandis que l'architecture AlexNet a 60 millions paramètres et seulement 8 couches.

Le tableau 6.2 illustre le nombre de paramètres des réseaux MobileNet en fonction de la taille des patchs (128×128) utilisés dans cette contribution.

Paramètres	MobileNetV1	MobileNetV2
Total	3 231 939	2 261 827
Entraînables	3 210 051	2 227 715
Non entraînables	21 888	34 112

TABLE 6.2 – *Le nombre de paramètres des architectures MobileNet utilisées.*

Les modèles robustes sont caractérisés par leur bonne performance sur les données non exploitées durant l'apprentissage et l'optimisation. Afin d'éviter les problèmes de sur-apprentissage, nous avons sélectionné les bons modèles en fonction de leur efficacité en termes de généralisation. La stratégie proposée consiste à suivre le processus d'apprentissage et de mesurer l'écart entre les courbes de performance des modèles sur les bases d'apprentissage, de validation, et de test. Afin de générer plusieurs modèles, nous avons utilisé différents optimiseurs durant le processus d'apprentissage : SGD et RmsProp.

La performance des modèles sur les données d'apprentissage et de validation était mesurée sur les patchs extraits séparément. Cependant, durant l'inférence la classe de l'image entière doit être prédite. Pour cela, dans le cas des données de test, nous avons généré les classes par un vote majoritaire entre les classes des patchs appartenant à cette image. L'algorithme 2 illustre le processus de vote, où C est le nombre des classes, P est le nombre des patchs : si le t éme patch est classifié par la classe j alors $d_{i,j} = 1$ si non $d_{i,j} = 0$.

Algorithme 2 : Le calcul des classes des données de test

Input : Test, Modèle
Output : Test_Précision
$Patches \leftarrow \varnothing$;
foreach *(Image* $\in$ *Test)* **do**
  $Classes \leftarrow \varnothing$;
  $Patchs \leftarrow Patch\_Extraction(Image)$;
  foreach *(Patch* $\in$ *Patchs)* **do**
    $Classes \leftarrow Classes \cup Modele.Classifier(Patch)$;
  end
  $Class \leftarrow argmax_{j\epsilon\{1,2,..,c\}} \sum_{t=1}^{P} d_{t,j}$;
end
Test_Précision ← Calculer_Test_Précision(Test) ;

Le dernier processus dans cette contribution combine plusieurs modèles MobileNet afin de réduire la grande variance de ces réseaux. En plus, la combinaison des décisions permet d'améliorer la généralisation.

Une méthode ensembliste est composée de 3 traitements : la génération des modèles, la sélection des modèles à combiner, et la combinaison de ces modèles.

Généralement, afin de générer un ensemble de modèles, plusieurs apprentissages sont effectués. Ce processus est couteux en termes de temps d'exécution, où le temps de traitement de chaque modèle est multiplié par le nombre total des modèles appartenant à l'ensemble. Afin de réduire la complexité temporelle, nous avons enregistré les états du modèle en cours d'apprentissage dans plusieurs points d'apprentissage. Dans le processus d'apprentissage des réseaux de neurones, les poids sont mets à jour dans chaque itération, et cela permet de générer différents modèles dans un seul processus d'apprentissage. De cette façon, la nature itérative du processus d'apprentissage des ANN peut assurer la construction d'un ensemble de modèles dans un seul apprentissage, et donc optimise la complexité temporelle. Dans ce cadre, nous avons enregistré l'ensemble de modèles dans des intervalles de 3 minutes.

Ensuite, nous avons sélectionné un sous ensemble de modèles à partir de l'ensemble généré dans l'étape précédente. Le but de cette sélection est de choisir un sous ensemble de modèles pertinents afin d'assurer l'efficacité de la combinaison à réaliser. Dans ce traitement, il est préférable de diminuer le nombre de modèles à combiner afin de réduire les exigences en termes de stockage et temps d'inférence.

Dans cette contribution, nous avons exploité deux techniques de sélec-

tion statique : les modèles les plus performants sur la base de validation et les derniers modèles enregistrés. Nous avons choisi cette sélection, car les meilleurs modèles sont caractérisés par leur bonne performance et les derniers modèles sont connus par leur stabilité.

Dans la dernière étape, nous avons combiné entre les modèles sélectionnés par les techniques de vote et de moyenne non pondérée. L'équation 6.2 illustre le processus de vote majoritaire, où J est la classe sélectionnée, C est le nombre des classes, et N est le nombre des modèles : si le t éme classificateur choisit la classe j alors $d_{i,j} = 1$ si non $d_{i,j} = 1$. L'équation 6.3 présente le processus de moyenne non pondérée, où $y_t(x)$ est le vecteur des poids du t éme classificateur pour l'échantillon x.

$$J = argmax_{j\epsilon\{1,2,..,c\}} \sum_{t=1}^{N} d_{t,j} \tag{6.2}$$

$$y(x) = \frac{\sum_{t=1}^{N} y_t(x)}{N} \tag{6.3}$$

L'algorithme 3 résume le processus de la méthode ensembliste proposée dans cette contribution.

Algorithme 3 : La combinaison entre plusieurs modèles MobileNet.

Input : Modèles, Validation, Test, Sélection ∈ {Meilleurs, derniers} , Combinaison ∈ {Vote, moyenne} , N : nombre de modèles sélectionnés.
Output : Test_Précision
if *(Sélection = Meilleurs)* **then**
  Modèles ← Descending_Sort (Models, Validation) ;
Sous_Modèles ← Modèles$[1, N]$;
if *(Combinaison = Vote)* **then**
  Test_Précision ← vote(Sous_Modèles, Test) ;
else
  Test_Précision ← moyenne(Sous_Modèles, Test) ;

6.1.6 L'étude expérimentale

Dans cette contribution, l'approche proposée a été évaluée sur la base d'apprentissage histopathologique lymphoma [Shamir *et al.* 2008]. Pour plus d'informations sur cette base, le chapitre 4 détaille sa configuration. Le tableau 6.3 décrit le nombre des images/patchs par catégorie, où les patchs sont extraits de chaque image en se basant sur les techniques d'augmentation décrites dans la section précédente.

Dans la phase d'apprentissage, nous avons employé les réseaux MobileNetV1 et MobileNetV2 et les deux optimiseurs : SGD et RmsProp. Pour évaluer les modèles générés, nous avons exploité la technique d'évaluation stratified hold out. Cette technique divise équitablement la base en trois sous bases : base d'apprentissage, base de validation, et base de test.

La base d'apprentissage permet de construire un modèle prédictif pour l'inférence. La base de validation est utilisée pour réajuster les pa-

Nombre total des images / patchs	Nombre des images / patchs par catégorie		
375/41860	CLL 113/12600	FL 140/15680	MCL 122/13580

TABLE 6.3 – *Le nombre des images/patchs dans la base d'apprentissage lymphoma.*

ramètres et les hyper-paramètres du modèle entrainé ou pour sélectionner le modèle le plus performant. Elle permet de mesurer la performance du modèle durant l'apprentissage afin d'éviter les problèmes de surapprentissage sur les données d'apprentissage. Dans ce cadre, nous avons utilisé la base de validation dans l'étape d'optimisation : la sélection du meilleur modèle (MobileNetV1 ou MobileNetV2 entrainés par SGD ou RmsProp), et la sélection des meilleurs modèles à partir de l'ensemble des modèles enregistrés dans plusieurs points d'apprentissage. Enfin, la base de test est exploitée pour valider la performance du modèle ou de l'ensemble des modèles sur des données non utilisées durant l'apprentissage et l'optimisation. Elle permet de tester l'efficacité du modèle entrainé en termes de généralisation, où le manque de généralisation est justifié par la grande variance entre les performances sur les données de test et d'apprentissage. Dans cette étude, nous avons divisé la base lymphoma en trois parties : 20 % pour le test, 20 % du reste de la base pour la validation, et 80 % pour l'apprentissage.

Le tableau 6.4 présente les paramètres d'apprentissage. Nous avons entrainé tous les modèles dans 60 000 itérations avec un lot de 128 et 100 échantillons en apprentissage et en validation, respectivement. La taille du lot présente le nombre d'échantillons d'apprentissage à considérer dans chaque itération afin de mettre à jour les poids du réseau. Cette taille est définie en fonction de la taille mémoire, où il est important d'assurer un compromis entre les exigences du matériel et la taille du lot, car les petits lots ont une mauvaise influence sur la précision du gradient.

Paramètre		Valeur
Nombre maximum d'itérations		60 000
Taille du lot (batch) (apprentissage / évaluation)		128/100
Taux d'apprentissage (η)	η initiale	0.05
	Taux de dégradation de η	0.9
	η final	$10^{-4} * \eta$
RmsProp	Constante de décroissement	0.9
	Momentum	0.9
	Epsilon	1.0
Fonction d'erreur		Entropie croisée
Taux de régularisation par abandon		0.2

TABLE 6.4 – *Les hyper-paramètres de l'apprentissage.*

Durant l'apprentissage à base de SGD, nous avons utilisé un taux d'apprentissage (η) décroissant. Le choix de la valeur du taux d'apprentissage est une tâche difficile, car une petite valeur peut ralentir le processus d'apprentissage, tandis qu'une grande valeur entraine des problèmes d'oscillations et un processus d'apprentissage instable. Le décroissement du taux

d'apprentissage permet d'adapter sa valeur à l'état d'apprentissage, où il est préférable de dégrader sa valeur pour assurer la stabilité du processus d'apprentissage. Généralement, Les réseaux entrainés se stabilise dans les itérations avancées, donc il devient important de ralentir la vitesse d'apprentissage.

Le tableau 6.4 illustre les valeurs initiale et finale du taux d'apprentissage et le taux de décroissement de l'optimiseur SGD, ainsi que les hyper-paramètres de l'optimiseur RmsProp. Pour plus d'information sur ces deux optimiseurs, le chapitre 1 explique leurs principes. Durant l'apprentissage, nous avons calculé l'erreur par la fonction d'entropie croisée catégorique. Cette fonction est conçue pour les problèmes de classification à plusieurs classes (équation 6.4 : S_p est le score de la classe p et C est le nombre de classes).

$$CE = -log(\frac{e^{S_p}}{\sum_j^C x^{S_j}}) \tag{6.4}$$

Enfin, nous avons utilisé la technique de régularisation par abandon avec un taux de 0.2 dans l'avant dernière couche entièrement connectée du réseau MobileNet. Cette technique ignore l'étape d'apprentissage au niveau d'un sous ensemble de neurones sélectionnés aléatoirement.

Dans cette contribution, nous avons commencé par la comparaison entre les techniques d'apprentissage transféré et d'apprentissage à partir des initialisations aléatoires (training from scratch). L'objectif principal de cette comparaison est de poursuivre le reste des étapes en se basant sur la technique appropriée. Dans ce cadre, nous avons transféré les poids du modèle MobileNetV1 entrainé sur la base d'apprentissage source ImageNet et réajusté les couches entièrement connectées sur la base d'apprentissage cible Lymphoma. Afin de réaliser ce traitement, il est important d'adapter la taille des images de cette base à la taille des entrées du modèle exploité (224×224). Pour cela, nous avons utilisé deux techniques : le redimensionnement des images à 224×224 et l'extraction des patchs de taille 224×224. Le tableau 6.5 présente les résultats obtenus par ces deux techniques. Nous avons remarqué l'efficacité de la stratégie de classification par patch. Cela est justifié par les limites de la technique de redimensionnement, car le réajustement des images de 1388×1040 à 224×224 cause une perte d'information microscopique, et cela implique une incertitude dans les décisions du modèle généré.

Taille du patch	Précision (Test)
Redimensionnement de l'image	0.6533
Des patchs de 224×224	0.7333

TABLE 6.5 – *Les résultats obtenus par la technique d'apprentissage transféré*

Ensuite, nous avons entrainé les réseaux MobileNetV1 et MobileNetV2 à partir des initialisations aléatoires. Nous avons commencé par MobileNetV1 associé aux optimiseurs SGD, SGD avec dropout, et RmsProp. Ensuite, nous avons entrainé le réseau MobileNetV2 à base de l'optimiseur SGD. La figure 6.4 illustre la courbe de convergence de l'erreur de prédiction du réseau mobileNetV1 entrainé par (a) SGD et (b) RmsProp. Cette

erreur est calculée par la fonction d'entropie croisée. L'étude comparative entre les deux courbes montre que RmsProp a atteint une erreur (0.054) plus réduite par rapport à SGD (0.3).

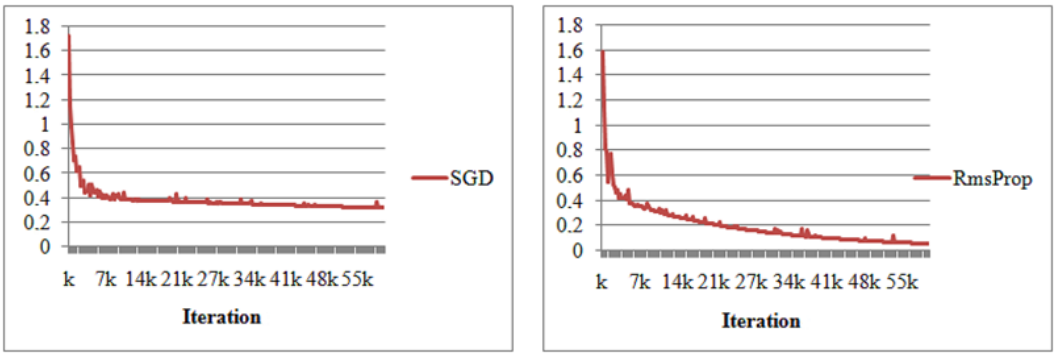

FIGURE 6.4 – *La convergence de l'erreur de prédiction des modèles MobileNetV1 sur la base d'apprentissage.*

La figure 6.5 illustre les courbes de convergence de la précision des modèles MobileNetV1 et MobileNetV2 sur les bases d'apprentissage, de validation, et de test : (a) MobileNetV1 + RmsProp, (b) MobileNetV1 + SGD, (c) MobileNetV1 avec dropout + SGD, et (d) MobileNetV2 + SGD.

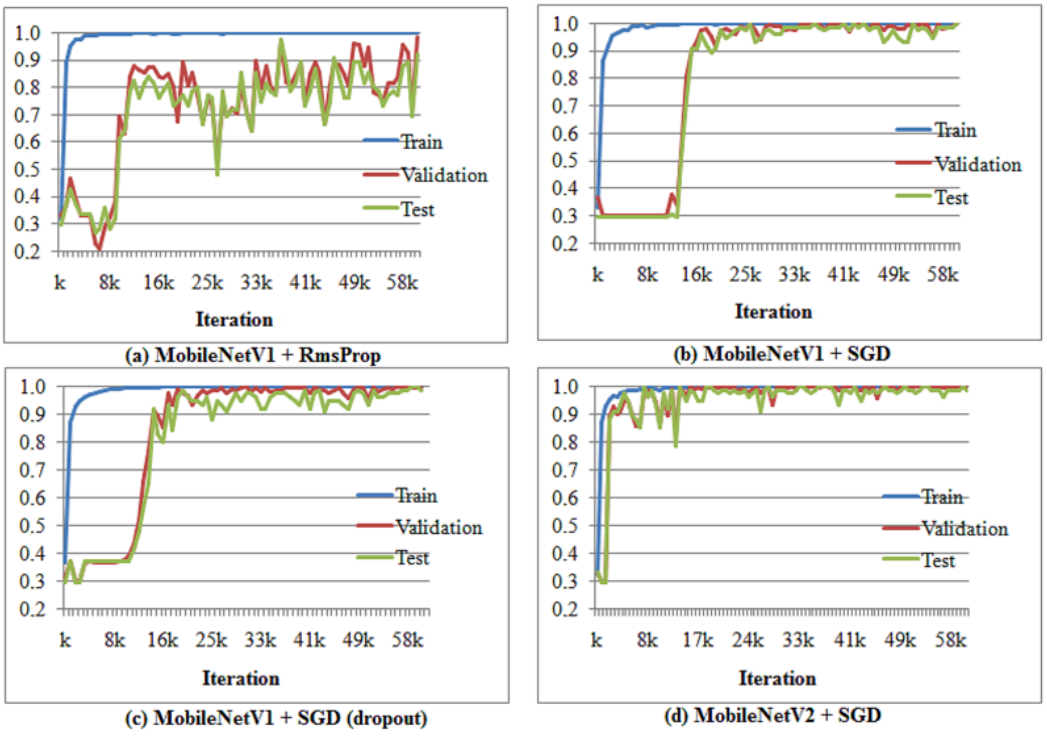

FIGURE 6.5 – *La convergence de la précision des modèles MobileNetV1 et MobileNetV2 sur les bases d'apprentissage, de validation, et de test.*

L'analyse de ces résultats a montré que les courbes du modèle MobileNetV1 + RmsProp indiquent une oscillation et une grande variance entre les résultats sur les bases d'apprentissage, de validation, et de test. Malgré l'erreur réduite de ce modèle par rapport à MobileNetV1 + SGD, MobileNetV1 + SGD a une variance faible entre les courbes de précision des 3 bases à partir de l'itération 20 000. Ces résultats indiquent que le modèle

MobileNetV1+ RmsProp a oscillé autour du minimum global à cause de sa convergence rapide vers un minimum local. D'autre part, le modèle MobileNetV1+ SGD a prouvé son efficacité en généralisation.

Pour valider ces hypothèses, nous avons augmenté le taux de régularisation par abandon de 0.01 à 0.2. La comparaison entre les deux modèles (MobileNetV1 avec et sans dropout) approuve son efficacité en généralisation, où nous avons remarqué que les courbes de précision sont pratiquement les mêmes. Sur la base de cette analyse, nous avons effectué l'apprentissage du réseau MobileNetV2 par l'optimiseur SGD. L'étude comparative entre les deux modèles MobileNetV1 + SGD et MobileNetV2 + SGD montre que le modèle MobileNetV2 est plus performant en terme de généralisation, car il existe moins de variances entre les courbe de précision des bases d'apprentissage, de validation, et de test.

Dans la phase de test, nous avons sélectionné le modèle qui maximise la performance de la base de validation à partir de l'ensemble des modèles enregistrés dans plusieurs points d'apprentissage. Le tableau 6.6 présente la précision obtenue sur les bases d'apprentissage, de validation, et de test par le modèle sélectionné. Nous avons remarqué une légère variance entre les performances d'apprentissage et de validation. Tandis, que les résultats sur les données test ont diminués légèrement en raison de leur différence par rapport aux données d'apprentissage et de validation. En plus, dans la partie test, nous avons adapté une stratégie d'évaluation différente qui consiste à effectuer un vote majoritaire entre les classes des patchs appartenant à la même image, tandis que dans l'apprentissage et la validation, nous avons évalué chaque patch séparément.

Réseau	Optimiseur	Apprentissage	Validation	Test
	RmsProp	0.9995	0.9866	0.9200
MobilenetV1	SGD	0.9998	0.9990	1
	SGD (Abandon)	0.9996	0.9981	0.9733
MobileNetV2	SGD	0.9996	0.9996	0.9867

TABLE 6.6 – *La précision des réseaux MobilenetV1 et MobileNetV2 sur les bases d'apprentissage, de validation et de test.*

Dans les expérimentations précédentes, nous avons utilisé la méthode d'évaluation stratified hold out. Cette méthode a été largement utilisée pour l'évaluation des réseaux de neurones en raison de sa rapidité et simplicité. Cependant, elle peut causer une grande variance entre les résultats obtenus sur les données d'apprentissage et de test. En plus, le choix des données d'apprentissage et de test appropriées est une tâche difficile. Afin de résoudre ces limitations, la technique de validation croisée est exploitée. Elle divise la base traitée à N groupes, ensuite, plusieurs apprentissages sont effectués pour générer N modèles. Enfin, la moyenne des prédictions des modèles est calculée pour prédire la classe en question.

Contrairement à la méthode hold out, cette technique effectue plusieurs apprentissages pour générer l'ensemble des modèles, et cela augmente les exigences du système en termes de stockage et de capacité de calcul. D'autre part, elle permet de générer des modèles confidents par rapport aux méthodes basées sur un seul apprentissage, et cela peut également améliorer l'efficacité des modèles en généralisation, et donc di-

minue les différents problèmes liés au sur-apprentissage. Le tableau 6.7 présente les résultats obtenus par la méthode de validation croisée (N =5). L'étude comparative montre une différence importante entre les résultats obtenus par les modèles 2 et 3. Cela valide l'efficacité de cette méthode d'évaluation dans la réduction de la grande variance entre les résultats des différents réseaux. Nous avons remarqué que les fausses prédictions du modèle 2 sont à cause de sa confusion entre les classes MCL et CLL, où 2 images de type MCL et 2 autres de type CLL ont été mal classées.

Division	1	2	3	4	5	Moyenne
Validation	0.9996	0.9981	0.9977	0.9983	0.9987	0.9985
Test	0.9867	0.9467	1	1	0.9600	0.9787

TABLE 6.7 – *La précision du réseau MobileNetV2 à base de la méthode d'évaluation 5-validations-croisées*

Afin d'améliorer la performance de l'ensemble des modèles générés par la méthode de validation croisée, nous avons exploité les méthodes ensemblistes. Dans ce cadre, nous avons enregistré des modèles dans plusieurs points d'apprentissage. Ensuite, nous avons utilisé deux méthodes de sélection statiques : la sélection des modèles les plus performants sur la base de validation et la sélection des derniers modèles enregistrés. Enfin, nous avons combiné les modèles sélectionnés par les méthodes de vote et de moyenne non pondérée.

Le tableau 6.8 présente la précision obtenue sur la base de test. Nous avons varié dans le nombre des modèles (de 2 à 50) à combiner pour mesurer son effet sur les résultats. Les résultats suggèrent qu'un sous ensemble de 3 à 4 modèles est suffisant pour améliorer la précision de 97.87% à 98.67% dans la stratégie de sélection par vote. D'autre part, nous avons remarqué une chute dans la précision lors de la combinaison des 15 à 20 meilleurs modèles et des 5 à 10 derniers modèles. Dans la stratégie de de moyenne non pondérée, nous avons identifié une amélioration dans la précision de 97.87% à 98.67% par la combinaison des 2 et 5 derniers modèles et une performance qui varie de 98.13% à 98.40% dans le reste des cas. Ces résultats prouvent que le choix du nombre des modèles à combiner est un paramètre important, où un nombre considérable de modèles n'assure pas toujours les meilleurs résultats.

Modèles	Méthode	Meilleurs modèles	Derniers modèles enregistrés	Méthode	Meilleurs modèles	Derniers modèles enregistrés
2	Vote	0.9813	0.9786	Moyenne	0.9840	0.9813
3		0.9867	0.9813		0.9840	0.9867
4		0.9867	0.9867		0.9840	0.9813
5		0.9867	0.9840		0.9840	0.9867
10		0.9867	0.9840		0.9813	0.9840
15		0.9840	0.9867		0.9840	0.9840
20		0.9840	0.9867		0.9840	0.9840
50		0.9867	0.9867		0.9813	0.9813

TABLE 6.8 – *La précision des méthodes ensembliste basées sur la combinaison des N derniers et meilleurs modèles enregistrés dans plusieurs points d'apprentissage.*

L'étude comparative entre les résultats obtenus par la méthode de moyenne non pondérée montre que la technique de combinaison des derniers modèles est plus performante par rapport à la technique de combinaison des meilleurs modèles. Cela peut être lié au manque de diversité entre les meilleurs modèles et confirme qu'un bon ensemble n'est pas toujours lié à la présence des modèles performants, mais plutôt à leur diversité et bonne collaboration.

En résumé, les résultats obtenus dans cette contribution suggèrent l'efficacité de SDG par rapport à Rmsprop dans l'apprentissage du réseau MobileNet sur la base Lymphoma. En plus, ces résultats valident les hypothèses sur l'efficacité des méthodes ensemblistes à base des modèles enregistrés dans plusieurs points d'apprentissage. Cette contribution examine cette méthode de sélection dans la classification des images histopathologique par le réseau MobileNet.

Le tableau 6.9 compare les résultats obtenus dans cette contribution et les résultats de l'état de l'art. Les meilleurs résultats dans la catégorie ML étaient obtenus par la méthode de segmentation à base des algorithmes génétiques (GA) [Tosta *et al.* 2017] et l'approche basée sur les descripteurs de texture visuels [Song *et al.* 2016]. Dans la catégorie des méthodes qui hybrident les techniques ML et DL, nous avons remarqué que l'hybridation entre les attributs locaux et les attributs extraits par les réseaux CNN pré-entrainé est plus performante par rapport à certaines méthodes ML [Shamir *et al.* 2008, Meng *et al.* 2010, Nava *et al.* 2016]. Enfin, la comparaison entre les méthodes DL valide l'efficacité du réseau MobileNetV2 par rapport à AlexNet. En plus, ces résultats montrent que d'autres réseaux plus profonds (ResNet50 et DenseNet) étaient moins performants ou avaient des performances équivalentes (InceptionV3) aux réseau MobileNetV2.

Méthode	Référence	Méthode d'évaluation	Précision (%)
ML	[Shamir *et al.* 2008]	-	85
	[Meng *et al.* 2010]	3-validation-croisée	92.70
	[Nava *et al.* 2016]	10-validation-croisée	93.83
	[Song *et al.* 2016]	5-validation-croisée	96.8
	[Di Ruberto *et al.* 2015]	validation-croisée	96.4 $+-$ 1.6
	[Tosta *et al.* 2017]	10-validation-croisée	98.14
	[Orlov *et al.* 2010]	8-validation-croisée	98-99
ML + DL	[Codella *et al.* 2016]	3-validation-croisée	95.5
	[Song *et al.* 2017a]	4-validation-croisée	96.5
	[Nanni *et al.* 2018]	-	97.33
	[Song *et al.* 2017b]	4-validation-croisée	97.9
	[Bai *et al.* 2019]	Hold out	99.1
	[Nanni *et al.* 2019a]	5-validation-croisée	96.87
DL	[Janowczyk & Madabhushi 2016]	5-validation-croisée	96.58
	[Nanni *et al.* 2019b]	5-validation-croisée	ResNet50 : 92.00 DenseNet : 93.60
	Inception-v3	Hold out	97.78
	MobileNetV2 [Dif & Elberrichi 2020d]	5-validation-croisée	97.87
	Ensemble de MobileNetV2 [Dif & Elberrichi 2020d]	5-validation-croisée	98.67

TABLE 6.9 – *La comparaison entre les résultats de l'état de l'art et les résultats obtenus.*

6.1.7 Conclusion

Dans cette contribution, nous avons présenté un framework basé sur les réseaux MobileNet. L'objectif principal de cette étude était d'exploiter un maximum de méthodes de régularisation afin d'éviter les problèmes de sur-apprentissage. Dans ce cadre, nous avons utilisé les méthodes d'augmentation de données, les petits (small) modèles, la méthode d'évaluation de validation croisée, et les méthodes ensemblistes.

Premièrement, nous avons appliqué des méthode d'augmentation de données sur la base d'apprentissage. Ensuite, nous avons effectué l'apprentissage et sélectionné le modèle approprié en termes de généralisation à partir de l'ensemble des modèles générés. Enfin, nous avons combiné entre plusieurs modèles enregistrés dans différents points d'apprentissage par les méthodes de vote et de moyenne non pondérée. Nous avons testé ces techniques sur la base d'apprentissage histopathologique lymphoma, et les meilleurs résultats étaient obtenus par un vote majoritaire entre 3 modèles de type MobileNetV2.

L'étude comparative aux autres résultats de l'état de l'art prouve l'efficacité de la combinaison des modèles enregistrés dans plusieurs points d'apprentissage pour la classification des images histopathologiques. Cependant, les méthodes de moyenne non pondérée et de vote majoritaire sont sensibles aux modèles faibles (week learners), et cela peut entraîner une chute dans la précision de l'ensemble, ainsi, il faut sélectionner les modèles appropriés pour la combinaison.

La suite de ce travail de recherche se base sur les méthodes de sélection dynamique. Dans ce cadre, nous avons proposé une stratégie de sélection basé sur la métaheuristique optimisation par essaim de particules (PSO) [Dif & Elberrichi 2020c].

6.2 Une nouvelle méthode de sélection dynamique des modèles d'apprentissage profond pour la classification du cancer colorectal

6.2.1 Résumé

Les méthodes d'apprentissage profond étaient largement utilisées dans les systèmes d'aide au diagnostic. Ces systèmes présentent un bon outil pour l'analyse des images histopathologiques. Malgré leur efficacité, le nombre limité des images peut entrainer un problème de sur apprentissage. Pour résoudre ce problème, plusieurs travaux ont proposé l'utilisation des méthodes ensemblistes basées sur des techniques de sélection statique.

Dans cette contribution, nous avons proposé une nouvelle stratégie basée sur la sélection dynamique à partir d'un ensemble de modèles de type DL. Premièrement, cette technique génère un ensemble de modèles par la méthode d'apprentissage transféré. Ensuite, la métaheuristique optimisation par essaim de particules (PSO) sélectionne un sous ensemble de modèles pertinents à partir de l'ensemble de modèles générés. Enfin, ces modèles sont combinés par un vote majoritaire ou par une moyenne

non pondérée. L'approche proposée a été testée sur une base d'apprentissage histopathologique spécialisé en classification du cancer colorectal. Les meilleurs résultats sont obtenus par le modèle Resnet121 et la stratégie de vote majoritaire. Ces résultats indiquent l'efficacité de la stratégie de sélection dynamique proposée en apprentissage profond.

6.2.2 Problématique

Récemment, les WSD sont utilisés pour la numérisation des WSI. Cette technologie a assisté les experts en imagerie médicale dans l'acquisition des données et leur exploitation dans les systèmes d'aide au diagnostic. Malgré l'efficacité de ces outils, le nombre des images collectées reste limité pour les algorithmes DL, car, ces derniers exigent un grand volume de données pour éviter les problèmes de sur-apprentissage. En plus, l'annotation des centaines à des milliers de données collectées exige un temps considérable et une grande charge de travail. Pour résoudre ces problèmes, certains travaux ont suggéré l'utilisation des techniques de régularisation et des méthodes ensemblistes.

Plusieurs travaux ont exploité les techniques de sélection statique dans les méthodes ensemblistes. Dans le cadre des travaux proposés dans la compétition ILSVRC, la stratégie de sélection des meilleurs modèles a été largement utilisée. Malgré leur efficacité, la diversité et la bonne coopération entre ces meilleurs modèles ne sont pas assurées, car, ces modèles peuvent avoir presque les mêmes erreurs et donc leurs combinaisons ne conduit pas toujours aux meilleurs résultats. Par exemple, dans la contribution présentée dans la section précédente [Dif & Elberrichi 2020b], nous avons remarqué que la combinaison par vote majoritaire des 4 derniers modèles (enregistrés dans les dernières itérations) est plus performante par rapport à la combinaison des meilleurs modèles.

D'autre part, la stratégie de sélection statique utilise la même technique de sélection pour répondre à tout type de problèmes. Cependant, en apprentissage automatique, il n'existe pas une méthode spécifique de haute qualité qui peut résoudre efficacement tous les problèmes.

6.2.3 Motivation

L'objectif principal de cette contribution est de répondre à deux problématiques : le sur-apprentissage et les inconvénients des méthodes de sélection statique. Dans ce cadre, nous avons proposé une méthode de sélection dynamique basée sur les techniques d'apprentissage transféré et la métaheuristique PSO.

L'apprentissage transféré

Dans les méthodes ensemblistes, la première étape génère un ensemble de modèles par un seul ou plusieurs apprentissages. La première technique enregistre l'état du modèle dans plusieurs points d'apprentissage [Dif & Elberrichi 2020b]. Malgré l'efficacité de cette méthode en termes de complexité de calcul, l'utilisation de plusieurs variantes d'un apprentissage effectué à partir d'une seule initialisation peut réduire la différence entre ces modèles. Cela peut donc diminuer la coopération entre

les membres de l'ensemble, car la diversité est un critère important dans le cadre des méthodes ensemblistes. D'autre part, la deuxième technique est couteuse en termes de complexité temporelle, car elle exige d'effectuer plusieurs apprentissages. Afin de résoudre ces problèmes, nous avons choisi d'opter pour la deuxième technique et d'exploiter la stratégie d'apprentissage transféré pour générer un maximum de modèles dans un temps de traitement raisonnable. En plus, l'apprentissage transféré permet de réduire les différents problèmes liés au sur-apprentissage sur les volumes limités de données.

La sélection dynamique

Cette technique permet de sélectionner automatiquement un sous ensemble de modèles pertinent et qui dépend du problème traité en entrée. L'avantage de cette technique se résume dans sa capacité dans la sélection des modèles appropriés en fonction de l'efficacité de l'ensemble plutôt que sur l'efficacité de chaque membre appartenant à cet ensemble. Dans cette situation, un modèle faible peut être bon s'il réussit à renforcer certaines bonnes décisions.

La stratégie de base d'une sélection dynamique consiste à créer un ensemble composé d'un nombre important de modèles, et ensuite, effectuer toutes les combinaisons possibles entre ces modèles. Ce problème est considéré comme un problème d'optimisation combinatoire, car il exige d'effectuer 2^N combinaisons, et cela peut entrainer une explosion dans la complexité temporelle si le nombre des modèles est important. En plus, augmenter le temps de calcul est délicat dans le cadre des méthodes DL, car ces derniers ont une complexité temporelle très élevée en apprentissage. Afin d'éviter les problèmes des méthodes exactes, nous avons exploité la métaheurisitques PSO pour réduire le temps de calcul.

Dans l'état actuel de nos connaissances, les techniques de sélection dynamique n'ont jamais été exploitées dans le cadre des méthodes DL à cause de leurs problèmes liés à la complexité temporelle élevée. L'objectif de cette contribution est de tester l'effet de la sélection dynamique en apprentissage profond. Dans ce cadre, nous avons exploité 7 architectures de type CNN, la métaheustique PSO, et la technique d'apprentissage transféré pour optimiser la complexité temporelle.

6.2.4 Etat de l'art des méthodes proposées pour la classification du cancer colorectal à partir des images histopathologiques

La classification automatique des sous types du cancer colorectal a connu un grand intérêt en vision par ordinateur. L'objectif de cette automatisation était de réduire l'inter et intra-variabilité entre pathologistes. Malgré les efforts faits, la disponibilité des données médicales et la difficulté de leurs annotations présentent des défis majeurs dans la conception de ces systèmes. Afin d'encourager la communauté de l'IA à développer des systèmes pertinents pour automatiser la classification des sous types du cancer colorectal, [Kather *et al.* 2016] ont proposé une base d'apprentissage histopathologique publique.

Le tableau 6.10 résume les méthodes proposées dans l'état de l'art

pour la classification de cette base d'apprentissage. Ces travaux sont de trois types ML, DL, et des méthodes hybrides (ML + DL). Les méthodes ML [Kather *et al.* 2016, Ly *et al.* 2017] sont basées sur trois modules principaux : extraction des caractéristiques, sélection des attributs, et classification. Contrairement aux méthodes ML, les méthodes DL effectuent implicitement l'extraction des caractéristiques à travers les couches de convolution. Dans ce cadre, [Ciompi *et al.* 2017] ont utilisé le réseau VGGNet pour la classification du cancer colorectal. Dans une autre contribution plus récente, [Rączkowski *et al.* 2019] ont proposé une architecture de type CNN (ARA-CNN) qui est basée sur les connexions résiduelles et une mesure bayésienne. Malgré les avantages des méthodes DL, ces derniers risquent le problème de sur-apprentissage sur cette base à cause du nombre limité des images histopathologiques. Pour résoudre ces limitations, des travaux ont proposé l'hybridation entre les techniques ML et DL [Cascianelli *et al.* 2018, Wang *et al.* 2017a, Pham 2017]. Ces travaux utilisent les techniques DL pour l'extraction des caractéristiques (VGGNet, ResNet, BCNN, et Autoencoder) et les algorithmes ML pour la classification (KNN, SVM, et Softmax).

Méthode	Référence	Extraction des caractéristiques	Classication	Normalisation de couleurs	Nombre de classes
ML	[Kather *et al.* 2016]	Caractéristiques d'histogramme d'ordre inférieur et supérieur, Filtre de Gabor, LBP, GLCM	SVM	-	8
	[Ly *et al.* 2017]	Transformation de distribution cumulative (CDT)	Automata	-	7
	[Bianconi *et al.* 2019]	Descripteurs de texture et de couleur : FullHist, MargHists, GLCM, LBP, Gabor + GLCM + LBP	SVM	-	8
DL	[Ciompi *et al.* 2017]	VGGNET - 11 couches		Normalization (SN1)	6
	[Rączkowski *et al.* 2019]	ARA-CNN		-	8
ML + DL	[Cascianelli *et al.* 2018]	VGGF-F, VGG-S, VGG-VD-16. Feature selection : GRP, CBFS, PCA	KNN	-	8
	[Wang *et al.* 2017a]	BCNN	SVM	Décomposition (H & E)	8
	[Pham 2017]	Autoencoder	Softmax	-	8
	[Bianconi *et al.* 2019]	ResNet-50, ResNet-101, VGGNet16, et VGGNetT19	1-NN	Normalization : Macenko et Reinhard	8

TABLE 6.10 – *Les méthodes proposées dans l'état de l'art pour la classification des soustypes du cancer colorectal.*

L'inter-variabilité entre laboratoires présente un autre défi dans l'analyse des images histologiques, où les modèles entrainés sur les données collectées d'un seul laboratoire risquent le problème de manque de généralisation sur les données provenant des autres laboratoires. Pour diminuer ce manque de généralisation, les techniques de normalisation de

couleurs [Bejnordi *et al.* 2015] et l'exploitation des bases d'apprentissages multi-sources [Ciompi *et al.* 2017] ont été proposés.

6.2.5 L'optimisation par essaims particulaires

Le terme métaheuristique a été introduit par Glover en 1986 pour différencier la recherche tabou des autres heuristiques [Hao *et al.* 1999].

— méta : indique le passage à un niveau " supérieur " pour étudier des informations de niveau inférieur.

— heuristique : signifie " trouver ".

Les métaheuristiques sont des algorithmes stochastiques qui permettent de trouver une solution efficace dans des délais raisonnables. Ces méthodes sont conçues pour résoudre des problèmes d'optimisation difficiles qui ne peuvent pas être facilement traités par les méthodes d'optimisation classiques. Contrairement aux méthodes exactes, les métaheuristiques n'assurent pas la meilleur solution. Le but principal est de trouver un bon optimum ou minimum local à proximité de l'optimum ou du minimum global. Le choix de l'optimum ou du minimum est défini en fonction de l'objectif du problème traité où il faut distinguer les problèmes de minimisation des problèmes de maximisation.

Différentes métaheuristiques ont été proposées dans l'état de l'art pour la résolution des problèmes NP-difficiles. Par exemple, les algorithmes génétiques (GA) étaient les premiers algorithmes évolutionnaires proposés qui s'inspirent de l'évolution naturelle. L'optimisation par essaims particulaires (PSO) [Eberhart & Kennedy 1995] est parmi les métaheuristiques les plus exploitées en raison de leur facilité et temps de traitement réduit. Cette méthode s'inspire de l'intelligence des essaims (oiseaux, poissons...). L'optimisation par colonies de fourmis (ACO) s'inspire du comportement des fourmis dans la recherche du plus court chemin [Dorigo *et al.* 1996]. Plus récemment, plusieurs métaheuristiques ont été développées : la recherche coucou (CS) [Yang & Deb 2009], l'algorithme des lucioles (FA) [Yang 2009], et l'algorithme des chauves-souris (BAT) [Yang 2010].

Les métaheuristiques sont basées sur deux concepts : l'exploitation et l'exploration de l'espace de recherche. En général, les algorithmes d'optimisation par essaims sont caractérisés par une bonne exploitation, tandis que les algorithmes évolutionnaires ont une bonne exploration. Dans cette contribution, nous nous intéressons à la métaheuristiques PSO.

L'optimisation par essaims particulaires [Eberhart & Kennedy 1995] est une métaheuristique inspirée de l'intelligence des essaims (oiseaux et poissons). Cette métaheuristique imite le comportement des particules dans leur processus de collaboration pour atteindre un objectif. Contrairement aux autres métaheuristiques, comme les GA, PSO est basée sur la coopération et la communication entre les particules plutôt que sur la compétition. Le comportement de chaque individu dans l'essaim est influencé par sa propre expérience et celle de ses voisins.

PSO exploite l'espace de recherche par le concept de mémoire, où chaque particule mémorise la meilleur solution dans son historique, sa position actuelle, et sa vitesse. Premièrement, une population de N particules est initialisée par des positions x_d et des vitesses v_d aléatoires et chaque

particule est évaluée selon une fonction fitness f(x). Dans chaque itération, les positions et les vitesses sont mises à jour selon la valeur de cette fonction. Á l'itération t+1, les valeurs des positions $x_d^{(t+1)}$ et des vitesses $v_d^{(t+1)}$ sont calculées selon l'équation 6.5, où , c est la constante d'accélération, Q est le poids de la vitesse, $l_d^{(t)}$ est la meilleure position dans l'historique de la particule, et $g^{(t)}$ est la position de la meilleure particule dans l'essaim. Chaque particule mets à jour sa vitesse selon deux composantes : cognitive $l_d^{(t)} - x_d^{(t)}$ et sociale $g^{(t)} - x_d^{(t)}$.

$$\begin{cases} v_d^{(t+1)} = (1-Q)v_d^{(t)} + Q(c_1(l_d^{(t)} - x_d^{(t)}) + c_2(g^{(t)} - x_d^{(t)})) \\ x_d^{(t+1)} = x_d^{(t)} + v_d^{(t+1)} \end{cases} \tag{6.5}$$

Pour les problèmes de maximisation, les valeurs de $l_d^{(t)}$ et $g_d^{(t)}$ sont définies selon les équations 6.6 et 6.7 respectivement.

$$l_d^{(t+1)} = \begin{cases} l_d^{(t)} sif(x_d^{(t+1)}) < f(l_d^{(t)}) \\ x_d^{(t+1)} sinon \end{cases} \quad d = 1, ..., N \tag{6.6}$$

$$g_d^{(t+1)} = \begin{cases} g_d^{(t)} sif(x_d^{(t+1)}) < f(g^{(t)}) \\ x_d^{(t+1)} sinon \end{cases} \quad d = 1, ..., N \tag{6.7}$$

Malgré l'efficacité de la version classique de PSO, plusieurs travaux ont proposé de l'améliorer [Garcia-Gonzalo & Fernandez-Martinez 2012], où l'objectif principal était d'éviter le problème de convergence prématuré.

Dans cette contribution, nous avons exploité la version de PSO qui utilise la notion des particules informatrices. Chaque particule de l'essaim est associée à M particules informatrices qui sont sélectionnées aléatoirement à partir de l'essaim. Cette version remplace le terme $l_d^{(t)}$ par $p_d^{(t)}$ (équation 6.8) afin d'éviter d'être piégé rapidement autour d'un minimum local, où $p_d^{(t)}$ est la meilleure solution dans l'historique des particules informatrices de la particule $x_d^{(t)}$.

$$\begin{cases} v_d^{(t+1)} = (1-Q)v_d^{(t)} + Q(c_1(p_d^{(t)} - x_d^{(t)}) + c_2(g^{(t)} - x_d^{(t)})) \\ x_d^{(t+1)} = x_d^{(t)} + v_d^{(t+1)} \end{cases} \tag{6.8}$$

6.2.6 La méthode proposée

La figure 6.6 présente le schéma de l'approche proposée. Cette méthode ensembliste est basée sur 3 modules principaux : l'apprentissage transféré, la génération d'un ensemble de modèles, et la sélection dynamique.

Premièrement, nous avons transféré les couches de convolution des modèles entrainés sur la base ImageNet aux nouveaux modèles. Cette étape permet de construire un module automatique d'extraction des caractéristiques et de produire une base d'apprentissage structurée structurée en attributs et instances. L'exploitation de la technique d'apprentissage transféré et l'utilisation des couches de convolution comme des modules d'extraction de caractéristiques permet de réduire le temps de traitement

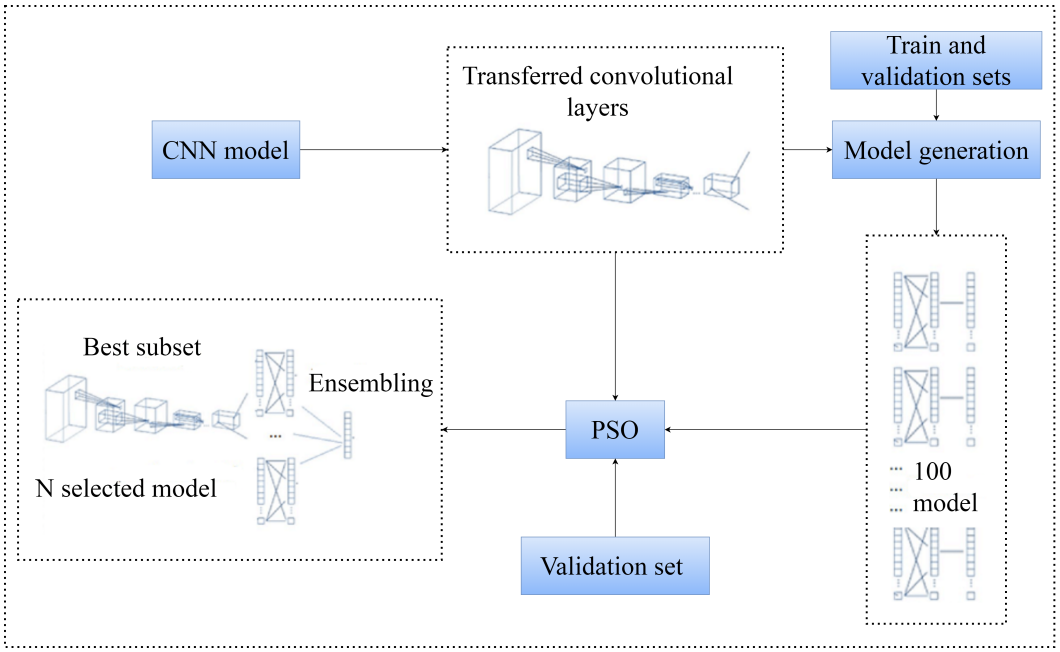

FIGURE 6.6 – *Les composants de la méthode ensembliste dynamique proposée.*

et diminue la capacité de stockage, car chaque image de taille 150×150 est transformée à un vecteur de caractéristiques de taille 2080. Ce vecteur est composé des valeurs de la dernière couche de convolution. En résumé, cette étape produit une nouvelle base structurée de taille $M \times 2080$ à partir d'une base d'apprentissage source non structurée de taille $M \times N \times N$, où $N \times N$ est la taille de l'image en entrée et M est le nombre d'instances.

Le deuxième module génère un ensemble de modèles pour la méthode ensembliste. Dans ce cadre, il existe plusieurs techniques : la variation dans la base d'apprentissage, la variation dans les conditions initiales, la variation dans l'architecture du réseau, et la variation dans la technique de combinaison.

Dans cette contribution, nous avons utilisé la technique de variation dans la base d'apprentissage car les ANN sont sensibles au changement de données, ainsi, cette technique peut générer des modèles variables.

La bonne performance de l'ensemble et la diversité de ses composants sont deux critères importants dans la construction d'un bon ensemble. Pour réaliser ce compromis, nous avons exploité deux stratégies. La première consiste à entrainer chaque réseau sur 90% des données sélectionnées aléatoirement à partir de la base d'apprentissage. Cette stratégie permet d'assurer la diversité et la généralisation. La seconde stratégie exclut les modèles faibles de l'ensemble afin d'assurer la qualité, où nous avons gardé seulement les modèles dont la précision est supérieure à un certain seuil. L'algorithme 4 explique le processus de génération de l'ensemble des modèles. Ces modèles sont entrainés sur la base d'apprentissage générée de la phase précédente. afin d'assurer la généralisation, nous avons exploité la base de validation pour tester la performance des modèles sur des données non utilisées durant l'apprentissage .

La troisième étape et qui constitue le corps de cette contribution sélectionne un sous ensemble pertinent de modèles à partir de l'ensemble des modèles générés de la partie précédente. Contrairement aux autres mé-

Algorithme 4 : Le processus de génération de l'ensemble des modèles.

Input : Base_Apprentissage, Base_Validation, Modèles = ∅, Seuil.
Output : Modèles
Réseau ← Initialisez aléatoirement des couches entièrement connectées ;
for $i \leftarrow 1, 100$ **do**
    Données ← Sélectionner aléatoirement 90% de données à partir de Base_Apprentissage ;
    Modèle ← Apprentissage(Réseau,Données) ;
    Précision ← Performance (Modèle,Base_Validation) ;
    if *(Précision* $\geqslant$ *Seuil)* **then**
        Modèles ← Modèles ∪ Modèle ;

thodes proposées dans l'état de l'art, dans cette contribution, nous avons choisi d'effectuer une sélection dynamique au lieu d'une sélection statique. Cette stratégie prend en considération la qualité du sous ensemble sélectionné au lieu de considérer séparément la qualité de chaque modèle appartenant à cet ensemble. En plus, son aspect dynamique lui permet de s'adapter à tout type de problèmes en entrée. Afin de sélectionner le sous ensemble le plus optimal, nous devons traiter 2^N combinaisons, où N = 100 est le nombre des modèles appartenant à l'ensemble généré dans l'étape précédente. En outre, ce nombre de combinaison est multiplié par le temps d'évaluation de chaque sous ensemble, et cela peut entrainer une complexité temporelle très élevée. Afin de réduire le temps de traitement, nous avons choisi d'exploiter les avantages des méthodes stochastiques dans l'optimisation de la complexité temporelle. Dans ce cadre, nous avons utilisé la métaheuristique PSOen raison de sa facilité et du temps de traitement optimisé. Le but principal était d'imiter le comportement des particules et de créer une collaboration et une coopération entre les différents modèles appartenant au groupe.

L'algorithme 5 présente le processus de sélection dynamique par la metaheuristique PSO. Dans cet algorithme, chaque particule désigne une solution. Ces solutions sont caractérisées par une position , une vitesse , et une valeur de la fonction fitness. Dans cette contribution, nous avons représenté la position par un vecteur composé de n éléments : $\{x_i^1, x_i^2, .., x_i^n\}$, $x_i^j \in [1, 100]$, $n \leq 100$, où x_i^j est une valeur discrète qui représente l'index du modèle j appartenant à l'ensemble i. La vitesse est représentée par un vecteur de n éléments : $\{v_i^1, v_i^2, .., v_i^n\}$, $v_i^j \in [0, 1]$, $n \leq 100$.

Dans le processus de sélection dynamique par PSO, l'algorithme commence par l'initialisation aléatoire de M particules et la meilleure solution Best. Ensuite, pour un certain nombre d'itérations Nmax, chaque particule mit à jour sa position et sa vitesse en fonction de l'équation 6.8, et chaque position est évaluée par la fonction fitness afin de décider de la meilleure solution dans l'historique des particules informatrices et la meilleure solution Best. La fonction nettoyer permet de convertir les valeurs continues x_i générées de l'équation 6.8 à des valeurs discrètes qui respectent la plage des valeurs , car l'ensemble original des modèles contient 100 modèles au

total. Afin d'effectuer ce processus, nous avons proposé d'arrondir les valeurs continues et de supprimer ou de remplacer les variables superflues selon une variable aléatoire $r \in [0, 1]$.

Algorithme 5 : Le processus de sélection dynamique par PSO.

```
Input : Nombre d'itérations Nmax, Nombre de particules M,
        Fonction_Fitness, Base_Validation.
Output : La meilleur solution
Initialiser une population aléatoire de M particules.;
Initialisez la meilleure solution Best;
Fonction_Fitness(Base_Validation,Best);
while (N < Nmax) do
    foreach (particle p ∈ pop ) do
        Mettre à jour la vitesse et la position de p (l'équation 6.8) );
        Nettoyer(p);
        Fonction_Fitness(Base_Validation, p);
        Mettre à jour la meilleure solution de p;
    Mettre à jour Best ;
```

Dans le processus de sélection dynamique, l'efficacité d'une solution dépend de la valeur de la fonction fitness. Cette fonction est choisie en fonction du problème traité. Dans cette contribution, nous avons exploité la valeur de f-mesure du sous ensemble sélectionné. L'algorithme 6 explique le processus d'évaluation des solutions. Premièrement, les prédictions des modèles appartenant au vecteur x_i sont combinées par un vote majoritaire ou une moyenne non pondérée. Ensuite, la f-mesure est calculée en fonction des valeurs prédites et réelles. Le but principal de la métaheuristique PSO est de trouver un sous ensemble de modèles qui maximise la valeur de la f-mesure sur la base de validation. L'équation 6.9 illustre la procédure de vote majoritaire, où $\widehat{y}$ est la classe prédite, C est le nombre des classes, si le modèle $x_i^{(t)}$ choisit la classe j alors $d_{i,j} = 1$, sinon $d_{i,j} = 0$. L'équation 6.10 présente le processus de moyenne non pondérée, où $y_t(s)$ est le vecteur des poids du modèle $x_i^{(t)}$ pour l'instance s.

6.2.7 L'étude expérimentale

L'approche proposée a été testée sur la base d'apprentissage de classification du cancer colorectal CRC [Kather *et al.* 2016]. Pour plus d'informations sur cette base histopathologique, le chapitre 4 détaille sa configuration.

Premièrement, nous avons commencé par la phase d'extraction des caractéristiques à base de 7 architectures de type CNN : Xception, Inception, Inception-ResNet, VGG (16, 18), ResNet (50, 101, 201), DenseNet (121, 169, 201), et MobileNet. Dans ce cadre, nous avons exploité les modèles précédemment entrainés sur la base d'apprentissage ImageNet, où les couches de convolution ont été transférées et utilisées comme des extracteur de caractéristiques. En résumé, cette étape génère 7 bases d'apprentissage structurées à partir de la base cible non structurée.

Algorithme 6 : Le processus de la fonction d'évaluation des solutions

Input : Base_Validation, Position x_i, Ensemble.
Output : Valeur_Fitness
Classes = $\leftarrow \phi$;
foreach *(Instance s ∈ Base_Validation.instances)* **do**
  if *(Ensemble = Vote)* **then**

$$\widehat{y} = argmax_{j\epsilon\{1,2,..,c\}} \sum_{t=1}^{N} d_{t,j} \tag{6.9}$$

  else if *(Ensemble = Moyenne)* **then**

$$\widehat{y} = argmax_{j\epsilon\{1,2,..,c\}} \frac{\sum_{t=1}^{N} y_t(s)}{N} \tag{6.10}$$

  Classes $\leftarrow$ Classes $\cup \widehat{y}$;
Fitness_value = f-measure (Classes, Base_Validation.classes) ;

Ensuite, nous avons entrainé un Perceptron composé d'une seule couche cachée et une couche de sortie sur chaque base. Le réseau a été entrainé par l'optimiseur RmsProp dans 10000 itérations avec un lot de taille 128. Pour l'évaluation, nous avons utilisé la méthode d'évaluation $5-$validation$-$croisée : $3/5$ pour l'apprentissage, $1/5$ pour la validation, et $1/5$ pour le test. Comme nous l'avons justifié dans la contribution précédente [Dif & Elberrichi 2020b], cette méthode d'évaluation permet d'améliorer la généralisation et de générer des modèles plus robustes et résistants aux problèmes de sur-apprentissage, où la base d'apprentissage est utilisée pour entrainer les modèles, la base de validation pour guider le processus d'optimisation de la métaheuristique PSO, et la base de test pour l'évaluation finale des modèles sur des instances non exploitées durant l'apprentissage et l'optimisation. Le tableau 6.11 présente les valeurs des paramètres de PSO utilisés dans cette contribution.

Paramètre	valeur
Nombre de particules informatives	4
Poids de la vitesse (Q)	0.73
Constante d'accélération (c)	2.05
Nombre d'itérations (Nmax)	50
Nombre de particules (M)	100

TABLE 6.11 – *Les valeurs des paramètres de la métaheuristique PSO.*

Le tableau 6.12 illustre les résultats obtenus en termes de précision sur les bases de validation et de test, où l'architecture désigne le modèle exploité pour l'extraction des caractéristiques. L'étude comparative montre une similarité entre les résultats obtenus par les modules Xception, Inception, et Inception-ResNet, et cela est justifié par la similitude de leurs blocs d'Inception. En plus, nous avons remarqué que les résultats obtenus par DenseNet sont moins performants par rapport aux autre résultats obtenus par des réseaux moins profonds (Xception et ResNets). Cela peut être justifié par une perte d'informations lors de l'extraction des caracté-

ristiques. En résumé, ces résultats démontrent qu'il n'existe pas de lien entre la profondeur et la performance.

Architecture	Validation (%)	Test (%)
Xception	92.55	92.44
Inception	91.25	91.08
Inception-ResNet	90.85	91.62
VGG16	89.9	89.68
VGG19	88.93	88.92
Resnet50	92.8	93.52
Resnet101	93.35	92.66
Resnet201	92.9	93.12
DenseNet121	83.05	83.96
DenseNet169	73.65	73.78
DenseNet201	91.33	91.88
MobileNet	90.62	90.92

TABLE 6.12 – *La précision du Perceptron à une seule couche cachée sur la base CRC.*

En raison de la bonne performance des réseaux ResNet dans l'étape précédente (92 % à 93 %), nous avons choisi de poursuivre les expérimentations à base de ces extracteurs de caractéristiques et d'ajouter une couche cachée supplémentaire dans le Perceptron pour l'apprentissage. Le tableau 6.13 illustre les résultats obtenus et la figure 6.7 compare entre les résultats du Percepton 1 (une seule couche cachée) et du Percepton 2 (deux couches cachées). Ces résultats illustrent l'efficacité du Percepton 2 par rapport à Percepton 1, où la précision à base du module d'extraction Resnet201 a été améliorée de 92.9% à 94%.

Architecture	Validation (%)	Test (%)
Resnet50	93.6	93.54
Resnet101	93.9	93.56
Resnet201	94	93.88

TABLE 6.13 – *La précision du Perceptron à deux couches cachées sur la base CRC.*

En se basant sur l'étude comparative effectuée dans les étapes précédentes et des bons résultats obtenus, nous avons choisi de poursuivre le reste des expérimentations à base de l'architecture du Perceptron 2 et les bases obtenues par les modules d'extraction de caractéristiques de type ResNet. Dans ce cadre, nous avons commencé par une étude comparative entre deux méthodes de sélection statique et la méthode de sélection dynamique proposée dans cette contribution (PSO).

La première méthode de sélection statique sélectionne le modèle le plus performant sur la base de validation à partir de l'ensemble original, tandis que la deuxième combine entre tous les modèles appartenant à cet ensemble. Pour chaque module ResNet, nous avons généré un ensemble composé de 100 modèles en suivant la méthode présentée dans l'algorithme 4. Ensuite, nous avons appliqué les 3 méthodes de sélection sur cet ensemble pour générer les sous ensembles. Enfin, nous avons com-

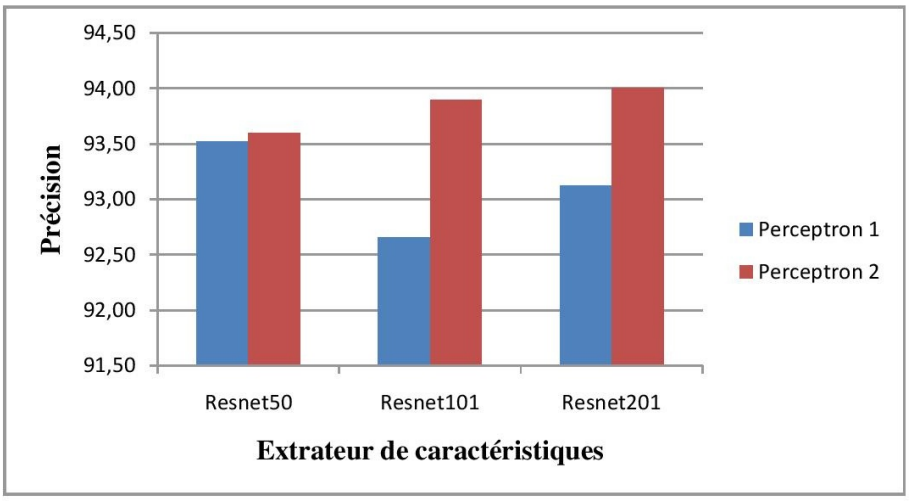

FIGURE 6.7 – *La comparaison entre les précision des Perceptron 1 et Perceptron 2.*

biné entre ces sous ensembles par la méthode de vote majoritaire ou de moyenne non pondérée.

Le tableau 6.14 compare les résultats obtenus par ces 3 méthodes de sélection en se basant sur le module d'extraction Resnet50. Ces résultats indiquent l'efficacité de la méthode de sélection du meilleur modèle par rapport à la combinaison de tous les modèles sur la base de validation, tandis que la stratégie de sélection de tous les modèles était plus performante sur la base de test, et cela valide l'efficacité de la méthode de sélection de tous les modèles en termes de généralisation. D'autre part, l'étude comparative entre les deux stratégies de sélection statique et la méthode de sélection dynamique proposée prouve l'efficacité de la sélection dynamique. En plus, nous avons remarqué l'efficacité de la méthode de vote majoritaire par rapport à la méthode de moyenne non pondérée sur la base de test.

Architecture		Validation (%)	Test (%)
Meilleur modèle		94.67	93.88
Moyenne non pondérée	PSO	94.97	93.98
	Tous les modèles	93.85	93.88
Vote	PSO	94.92	94.02
	Tous les modèles	94.25	94

TABLE 6.14 – *La comparaison entre les résultats des méthodes de sélection statique et dynamique (PSO) à base de Resnet50.*

Le tableau 6.15 illustre les résultats obtenus à base du module d'extraction Resnet101. Ces résultats montrent l'efficacité de la méthode de combinaison de tous les modèles par rapport à la stratégie de sélection du meilleur modèle en termes de généralisation. D'autre part, l'étude comparative valide les hypothèses sur l'efficacité de PSO en sélection dynamique, tandis que dans ces expérimentations, la méthode de moyenne non pondérée était plus adaptée à la stratégie de sélection dynamique.

Le tableau 6.16 présente les résultats obtenus à base du module d'extraction Resnet121. Ces résultats prouvent l'efficacité de la stratégie de

Architecture		Validation (%)	Test (%)
Meilleur modèle		94.85	93.44
Moyenne	PSO	95.2	93.66
non pondérée	Tous les modèles	94.03	93.46
Vote	PSO	95.12	93.54
	Tous les modèles	94.27	93.66

TABLE 6.15 – *La comparaison entre les résultats des méthodes de sélection statique et dynamique (PSO) à base de Resnet101.*

sélection de tous les modèles par rapport au meilleur modèle et l'efficacité de la méthode de sélection dynamique par rapport à la sélection statique. En plus, nous avons observé l'efficacité de la méthode de vote majoritaire pour les deux stratégies de sélection.

Architecture		Validation (%)	Test (%)
Meilleur modèl		94.52	93.9
Moyenne	PSO	95	94.48
non pondérée	Tous les modèles	94.02	94.12
Vote	PSO	94.93	94.52
	Tous les modèles	94.15	94.45

TABLE 6.16 – *La comparaison entre les résultats des méthodes de sélection statique et dynamique (PSO) à base de Resnet121.*

L'étude comparative entre les méthodes de sélection statique et dynamique prouve l'efficacité de PSO en sélection dynamique en se basant sur les deux méthodes de combinaison (vote et moyenne non pondérée). Les résultats obtenus montrent qu'un sous ensemble sélectionné par PSO (composé de cinq à huit modèles) est plus performant par rapport à l'ensemble original (composé de 100 modèles). Dans toutes les expérimentations effectuées, le meilleur résultat a été obtenu par PSO (94.52 %) à base de l'architecture Resnet121 et la méthode de combinaison de vote majoritaire, où PSO a amélioré les performances de 93.88 % à 94.52 %. Ces résultats valident les hypothèses sur l'importance du degré de coopération entre les modèles appartenant à l'ensemble et la qualité de leur combinaison.

En résumé, les résultats obtenus illustrent l'importance du processus de sélection en apprentissage ensembliste et valident l'efficacité de la stratégie de sélection dynamique en apprentissage profond. La méthode proposée a prouvé son efficacité en généralisation en raison de la stratégie d'apprentissage transféré et la méthode de sous-échantillonnage utilisée dans le processus de génération des modèles.

Le tableau 6.17 compare entre les résultats de cette contribution [Dif & Elberrichi 2020c] et les résultats de l'état de l'art sur la base d'apprentissage CRC. Les méthodes proposées sont de trois types : ML [Kather *et al.* 2016, Ly *et al.* 2017, Bianconi *et al.* 2019], DL [Ciompi *et al.* 2017, Rączkowski *et al.* 2019] et des hybridations entre ces deux stratégies [Cascianelli *et al.* 2018, Wang *et al.* 2017a, Pham 2017]. L'étude comparative entre les résultats montre une grande différence entre les résultats des méthodes ML, cela peut être justifié par leurs diffé-

rentes stratégies d'extraction de caractéristiques. D'autre part, la méthode hybride proposée par [Wang *et al.* 2017a] est plus performante par rapport aux autres méthodes ML et DL. En résumé, ces résultats prouvent l'efficacité de la méthode de sélection dynamique proposée, où nous avons obtenus des résultats satisfaisants sur la base d'apprentissage CRC.

	Méthode d'évaluation	Résultats (%)
[Kather *et al.* 2016]	10-cross-validation	Précision : 87.4
[Ly *et al.* 2017]	Hold out	Précision : 47.33
[Ciompi *et al.* 2017]	5-validations-croisées	Précision : 79.66
[Wang *et al.* 2017a]	5-validations-croisées	Précision : 92.6
[Pham 2017]	5-validations-croisées	Erreur : 16
[Cascianelli *et al.* 2018]	Hold out	Précision : 85
[Dif & Elberrichi 2020c]	5-validations-croisées	Précision : 94.52

TABLE 6.17 – *La comparaison entre les résultats de l'état de l'art et les résultats obtenus pour la classification du cancer colorectal.*

6.2.8 Conclusion

Dans cette contribution, nous avons proposé une méthode de sélection dynamique en apprentissage profond. La méthode proposée est basée sur les techniques d'apprentissage transféré et la métaheuristique PSO. L'objectif de l'apprentissage transféré est de générer et d'entrainer un ensemble de modèles dans un temps raisonnable. D'autre part, PSO est exploitée afin d'effectuer une sélection dynamique à partir de l'ensemble des modèles générés. Dans ce cadre, nous avons utilisé 7 architectures de type CNN pré-entrainées sur la base ImageNet pour l'apprentissage transféré.

L'étude comparative entre l'approche proposée et deux autres stratégies de sélection statique a prouvé l'efficacité de la sélection dynamique. Cette méthode est caractérisée par son aspect dynamique qui permet d'assurer un compromis entre la bonne performance et la coopération ente les modèles de l'ensemble. Malgré l'efficacité de la technique de sélection dynamique en terme de précision, son temps de traitement est très élevé en cas d'apprentissage à partir des initialisations aléatoires au lieu de l'apprentissage transféré.

La suite de ce travail de recherche se base sur l'apprentissage transféré entre des bases d'apprentissage de même domaine au lieu de transférer la connaissance à partir de la base ImageNet [Dif & Elberrichi 2020d]. Dans ce cadre, nous avons testé l'effet de l'apprentissage transféré et de la profondeur de fine tuning sur différentes bases d'apprentissage histopathologiques.

6.3 Une nouvelle stratégie de fine-tuning entre les bases d'apprentissage histopathologiques en apprentissage profond

6.3.1 Résumé

Cette contribution présente une nouvelle stratégie de fine-tuning pour l'analyse des images histopathologiques. Contrairement aux solutions précédemment proposées, où les modèles entrainés sur la base d'apprentissage ImageNet sont réutilisés pour la classification d'une nouvelle tâche, cette étude propose d'effectuer l'apprentissage transféré à partir des modèles précédemment entrainés sur des bases d'apprentissage histopathologiques. L'objectif principal est d'exploiter les hypothèses liées à l'efficacité de l'apprentissage transféré entre les bases non distantes et d'examiner pour la première fois ces suggestions sur les images histopathologiques. Dans ce cadre, nous avons utilisé trois modules principaux : le réseau Inception-v3, 6 bases d'apprentissage histopathologiques source, et 4 bases cibles. Les résultats obtenus illustrent que les modèles pré-entrainés sur des bases histopathologiques sont plus performants par rapport aux modèles pré-entrainés sur la base ImageNet. En particulier, le modèle entrainé sur la base 2018-A a prouvé son efficacité en apprentissage transféré sur différente tâche histopathologique cible. L'étude comparative avec les autres résultats obtenus dans l'état de l'art montre que la méthode proposée a atteint les meilleurs résultats sur les deux bases cibles : CRC (95.28 %) et KIMIA-PATH (98.18 %).

6.3.2 Problématique

Les réseaux CNN ont connu un grand intérêt en vision par ordinateur en raison de leur efficacité en traitement des images. Malgré leurs avantages, ces réseaux risquent le problème de sur-apprentissage sur les volumes limités de données en raison de leur nombre de paramètres élevé. D'autre part, la quantité limitée des données histopathologiques et la difficulté de leur annotation ont limité l'exploitation de ces techniques pour l'analyse de ces images. Plusieurs travaux ont suggéré l'utilisation de la technique d'apprentissage transféré. Elle permet de transférer de la connaissance à partir d'un modèle entrainé sur la base d'apprentissage volumineuse ImageNet à un nouveau modèle. Ensuite, le nouveau modèle est réajusté sur la nouvelle tâche histopathologique à classifier. La quantité élevée des données appartenant à la base ImageNet et le nombre important des catégories ont rendu cette base un bon outils pour l'apprentissage transféré. D'autre part, la nature hiérarchique des DNN favorise l'exploitation de la technique d'apprentissage transféré par rapport aux autres méthodes d'apprentissage automatique classiques. L'apprentissage transféré entre deux tâches distantes est justifié par la similarité entre les attributs de bas niveau qui sont représentés dans les premières couches cachées (bords et coins). Malgré ces hypothèses, il n'existe pas de principes théoriques sur le fonctionnement interne de cette stratégie et beaucoup de questions se posent sur la relation entre la base ImageNet et les bases d'apprentissage histopathologiques.

6.3.3 Motivation

L'objectif principal de cette contribution est de partager de la connaissance entre des bases d'apprentissage histopathologiques au lieu de l'extraire à partir de la base d'apprentissage ImageNet. Notre but est d'exploiter les avantages de l'apprentissage transféré à partir des bases d'apprentissage non distantes. Le partage entre les modèles entrainés sur des bases appartenant au même domaine peut être plus utile en raison de l'apparence similaire des images histopathologiques par rapport à ceux de la base ImageNet.

Dans cette étude, nous avons exploité les techniques de l'apprentissage transféré et de fine tuning. Premièrement, nous avons entrainé le réseau Inception-V3 sur des bases d'apprentissage histopathologiques sources qui sont caractérisées par un nombre important d'images. Ensuite, nous avons fixé les paramètres des premières couches et réajusté les paramètres des couches profondes sur les nouvelles bases histopathologiques cibles. Les premières couches sont fixées, car elles incluent des caractéristiques générales, tandis que les couches profondes sont plus spécifiques à la nouvelle tâche à traiter. Ce processus est inspiré par l'étude présentée par [Zeiler & Fergus 2014], où ils ont illustré la nature hiérarchique des réseaux CNN par la visualisation du contenu interne des couches cachées. Ils ont remarqué que les premières couches présentent des structures simples : coin, bord et couleur, alors que les couches profondes sont caractérisées par des structures plus complexes générées par la combinaison des structures des couches précédentes. Malgré ces hypothèses, il est difficile de définir le nombre de couches à fixer et à réajuster.

En résumé, cette étude présente la première contribution qui teste la transférabilité entre deux tâches histopathologiques. L'objectif principal est de comparer la performance du processus d'apprentissage transféré à partir de la base ImageNet à la performance d'apprentissage transféré à partir des bases histopathologiques. Dans ce cadre, nous avons étudié l'effet de la profondeur de fine tuning sur les résultats.

6.3.4 La méthode proposée

Ce travail de recherche présente une méthode d'apprentissage transféré entre les bases d'apprentissage appartenant au même domaine. L'objectif de cette contribution est d'évaluer l'apprentissage transféré à partir des base d'apprentissage histopathologique au lieu de la base ImageNet. Dans ce cadre, nous avons exploité un ensemble de bases d'apprentissage sources et cibles. Les bases sources sont sélectionnées en fonction du nombre des images et leur taille, où un large volume de données est exigé afin de générer des modèles robustes au problème de surapprentissage. Tandis que les bases d'apprentissage cibles sont caractérisées par un nombre limité de données afin de démontrer l'efficacité des techniques d'apprentissage transféré et de fine tuning en terme généralisation.

La figure 6.8 présente le schéma de la méthode proposée. Il est composé de deux modules principaux : prétraitement et apprentissage transféré. Premièrement, les bases d'apprentissage sources sont prétraitées par les techniques d'augmentation de données, et ensuite le résultat du pré-

traitement est fourni au réseau Inception-V3 afin de générer un modèle de base pour l'apprentissage transféré. Enfin, ce modèle est utilisé dans le processus d'apprentissage transféré et de fine tuning sur une nouvelle base d'apprentissage cible. Le processus de fine tuning permet de générer un nouveau modèle adapté à la tâche de classification cible.

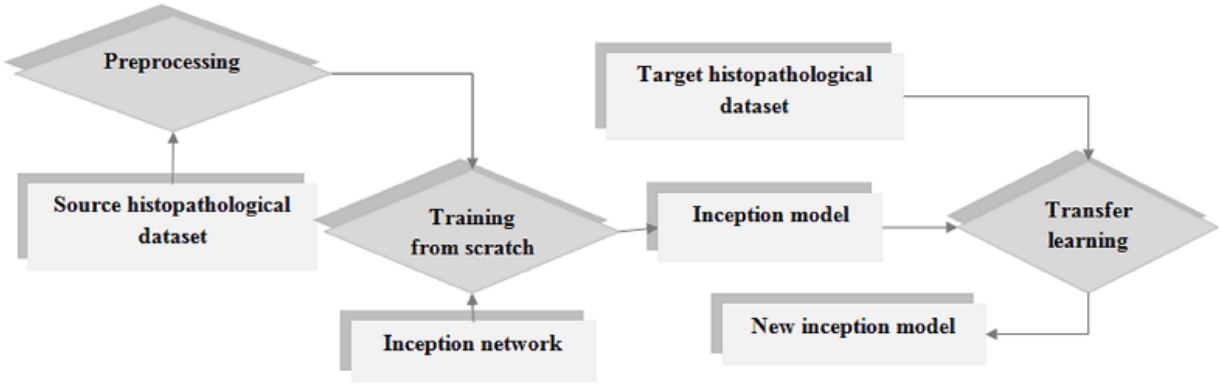

FIGURE 6.8 – *Les composants de la stratégie de fine-tuning utilisée entre les bases d'apprentissage histopathologiques.*

Prétraitement

En histopathologie, les WSI numérisées par les WSD sont caractérisées par une haute résolution. Le processus d'apprentissage des réseaux CNN sur ces images peut conduire à une explosion dans le nombre des paramètres, car ces derniers dépendent de la taille de l'image en entrée. D'autre part, le nombre des images histopathologiques est limité, et cela peut entrainer un problème de sur-apprentissage. Afin de résoudre ces problème, les techniques d'augmentation de données sont utilisées pour créer plusieurs variantes de taille réduite à partir d'une seule image.

Dans cette contribution, nous avons exploité la stratégie d'extraction des patchs. Ces patchs sont générés par la méthode de fenêtres coulissantes avec un taux de chevauchement $r \in \{0, 0.5, 0.3\}$. La figure 6.9 illustre le processus de la fenêtre coulissante. Premièrement, une fenêtre de taille 224×224 est créée, ensuite, elle est glissée sur l'image cible et déplacée selon le taux de chevauchement, où $r = 0.5$ signifie que le patch courant et le patch suivant ont 50 % de données communes.

La deuxième étape consiste à effectuer des rotations de 90° et 180° sur les patchs résultants. Cette étape s'inspire du processus manuel effectué par les pathologistes lors de leurs observations de la lame de verre sous différents angles. En plus, elle permet d'éviter les problèmes d'attaques contradictoires sur les réseaux CNN et d'améliorer leur efficacité en termes de généralisation. Le tableau 6.18 présente le nombre des images après l'augmentation de données. Nous avons choisi d'augmenter les bases d'apprentissage dont la taille de l'image source est supérieure à la taille du patch, et pour le reste des bases, nous avons redimensionné les images de taille réduite à 224×224.

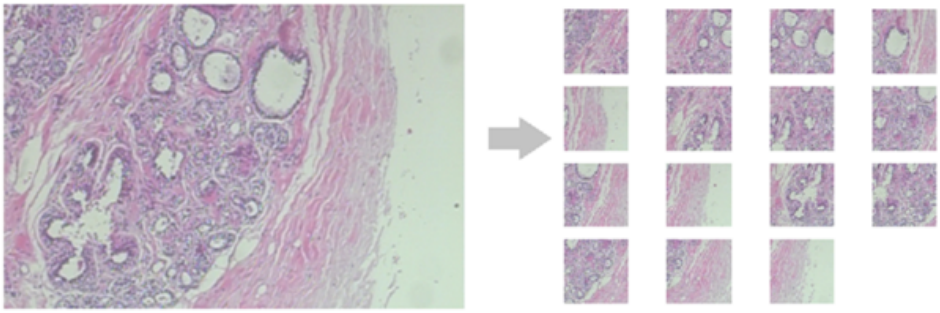

FIGURE 6.9 – *Le résultat d'extraction des patchs par la méthode de fenêtre coulissante (ICAR-2018).*

Dataset	Taux de chevauchement	Apprentissage	Validation
Breakhis-2c	0.3	114 912	38 304
Lymphoma	0.3	157950	52650
MITOS-Atypia	0	76896	25 704
ICIAR 2018-A	0.5	146 880	48 960
BBHC-2015	0	9396	2430

TABLE 6.18 – *Le nombre des patchs après l'augmentation de données.*

Apprentissage transféré

En Apprentissage profond, l'objectif de l'apprentissage transféré est de réutiliser les modèles pré-entrainés sur une base source S_T sur une autre base cible T_T. Cette technique permet de réduire la complexité temporelle des réseaux de neurones profonds et d'éviter les problèmes de sur-apprentissage sur les volumes limités de données. La méthode d'apprentissage transféré réutilise les poids $P(X_s \mid Y_s)$ du modèle pré-entrainé sur la base S_T, ensuite, un sous ensemble des premières couches (L_i, $i \in \{1, 2, .., k\}$, $k <$ nombre de couches) est fixé, et le reste des couches (L_i, $i \in \{k+1, \ldots, N\}$, $N =$ nombre de couches) sont réajustées sur la nouvelle tâche de classification afin d'obtenir les nouveaux poids $P(X_T \mid Y_T)$ de T_T.

Les premières couches sont fixées en raison de la similitude des caractéristiques de bas niveau entre les différentes tâches, comme les bords et les coins. Tandis que ces caractéristiques deviennent plus spécifiques à la tâche traitée, au niveau des couches profondes. Enfin, la dernière couche softmax est remplacée par une nouvelle couche adaptée au nombre de classes dans T_T.

Dans cette contribution, nous avons proposé d'effectuer un fine tuning entre des bases d'apprentissage sources et cibles, où $S_T \in$ (Lymphoma, Breakhis, MITOS-Atypia, ICIAR 2018-A, Pcam, NCT-CRC-HE-100K-NONORM, ImageNet) et $T_T \in$ (CRC, BBHC-2015, CRC-VAL-HE-7K, KIMIA-PATH960). La structure des bases d'apprentissage S_T et T_T est détaillée dans le chapitre 4.

L'algorithme 7 explique le processus d'apprentissage transféré. La variable Nombre_couches_entraînables $\in \{5, 10, 50, 100\}$ présente le nombre des dernières couches à réajuster. En résumé, l'objectif de cet algorithme est de mesurer l'effet du nombre de couches à réajuster sur les résultats et de comparer les résultats d'apprentissage transféré à partir des bases dis-

tantes (ImageNet) et non distantes (histopathologiques). Les bases histopathologiques sont catégorisées en tant que bases non distantes en raison de l'apparence similaire de leurs structures microscopiques, et cela peut réduire la profondeur de fine tuning et améliore par la suite la précision du réseau entrainé.

Algorithme 7 : Le processus d'apprentissage transféré entre les bases histopathologiques

Input : Bases_histopathologique_cibles,
Nombre_couches_entraînables, Optimiseur,
Modèles_Inception-v3 = {Lymphoma, Breakhis,
MITOS−Atypia, ICIAR 2018−A, Pcam,
NCT−CRC−HE−100K−NONORM, ImageNet }
Output : Modèles
Modèles ← ϕ ;
foreach *(Modèle ∈ Modèles_Inception−v3)* **do**
  foreach *(Couche∈ Modèle.couches[:-Nombre_couches_entraînables])* **do**
    Couche.entrainable ← Faux;
  Nouveau_modèle ← Modèle.apprentissage(Bases_histopathologique_cibles, Optimiseur);
  Modèles ← Modèles ⋃ Nouveau_modèle ;

Optimiseur et métriques

Dans cette contribution, nous avons exploité la descente de gradient accélérée de Nesterov (NAG) pour l'optimisation en apprentissage à partir des initialisations aléatoires et en apprentissage transféré. Le processus de mise à jour des poids de NAG est détaillé dans le chapitre 1 de cette thèse. Pour l'évaluation, nous avons utilisé la fonction d'entropie colisée (CE) et la précision. Le chapitre 1 détaille le processus de calcul de CE, et l'équation 6.11 illustre le calcul de la précision, où TP est le taux des vrais positifs, TN est le taux des vrais négatifs, FP est le taux des faux positifs, et FN est le taux des faux négatifs. Pour les bases dont la taille des images > 224, nous avons effectué un vote majoritaire entres les classes des patchs en suivant la même méthode expliquée dans la première contribution [Dif & Elberrichi 2020b] (algorithme 2).

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{6.11}$$

6.3.5 L'étude expérimentale

Le but de cette section est d'évaluer et de discuter les résultats obtenus par la technique d'apprentissage transféré entre les bases d'apprentissage histhopatologiques sources et cibles.

Le tableau 6.19 présente les valeurs des paramètres utilisés pour l'apprentissage et l'apprentissage transféré. Nous avons utilisé l'optimiseur

NAG avec 0.9 de momentum et un taux d'apprentissage (η = 0.01) variable avec une décroissance exponentielle. La valeur de η dépend de l'état des poids entrainés, où il doit être diminué dans les époques avancées afin d'éviter les problèmes d'oscillation.

Paramètre		Valeur
Nombre d'époques	Apprentissage	20
	Apprentissage transféré	100
Taille du lot		64
NAG	Momentum	0.9
	Taux d'apprentissage (η)	0.01
	Taux de décroissement	10^{-6}

TABLE 6.19 – *Les hyper-paramètres de l'apprentissage.*

Pour l'évaluation, nous avons utilisé la méthode hold out : 3/5 pour l'apprentissage, 1/5 pour la validation, et 1/5 pour le test. Nous avons évalué le modèle généré sur la base de validation en se basant sur les patchs extraits séparément, tandis que, pour le test, nous avons effectué la méthode de vote détaillé dans la section précédente. La technique de vote est valable pour les bases d'apprentissage dont la taille de l'image est supérieure à la taille du patch (224×224) (Lymphoma, Breakhis, MITOS-Atypia, ICIAR 2018-A, BBHC-2015). Pour le reste des bases, nous avons combiné les bases de test et validation pour l'évaluation (Pcam, NCT-CRC-HE-100K-NONORM, CRC, KIMIA-PATH960, CRC-VAL-HE-7K).

Apprentissage à partir des initialisations aléatoires

L'apprentissage permet de préparer les modèles sources qui seront exploités par la suite dans la phase d'apprentissage transféré. Dans ce cadre, nous avons entrainé le réseau inception-v3 sur les bases sources $S_T \in$ (Lymphoma, Breakhis, MITOS-Atypia, ICIAR 2018-A, Pcam, NCT-CRC-HE-100K-NONORM). Les figures 6.10 et 6.11 présentent les courbes de convergence de la précision des modèles sur les bases d'apprentissage et de validation.

L'objectif principal de ce traitement n'était pas de maximiser la performance des modèles entrainés sur les bases sources, mais plutôt de préparer ces modèles pour l'étape de l'apprentissage transféré. Pour cela, nous avons sélectionné le modèle enregistré dans la dernière itération au lieu d'enregistrer le modèle le plus performant sur la base de validation, car il n'existe aucune hypothèse sur le comportement du meilleur modèle sur la base cible.

Le tableau 6.20 présente les résultats obtenus en termes de précision sur les bases de validation et de test. Ce tableau illustre l'efficacité de l'apprentissage sur les bases lymphoma et breakhis-2c.

L'étude comparative a montré une grande différence entre les performances des modèles entrainés sur une même base d'apprentissage associé à un nombre différent de classes (breakhis-2c et breakhis-8c). Ces résultats suggèrent la complexité de la classification multitâches par rapport à la classification binaire. D'autre part, nous avons observé que les modèles

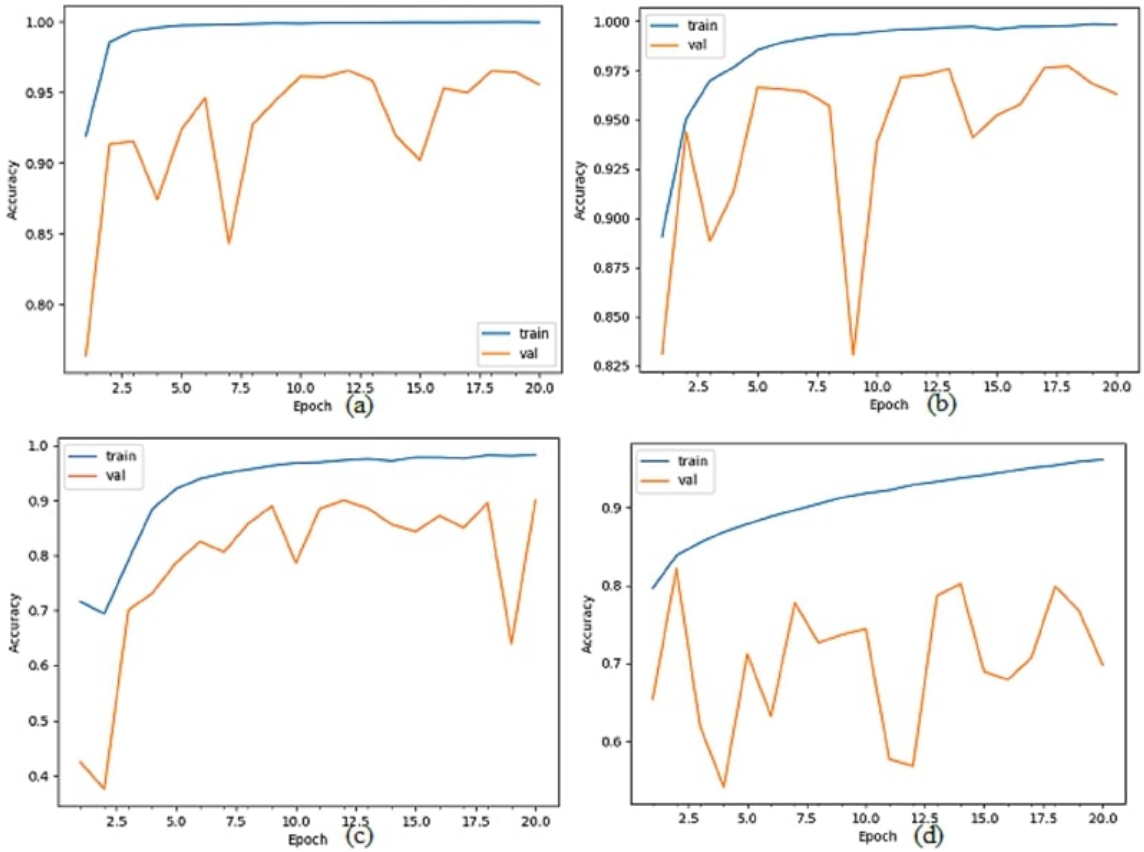

FIGURE 6.10 – *La convergence de la précision des modèles entrainés sur les bases (a) Lymphoma, (b) Breakhis-2c, (c) Breakhis-8c, et (d) MITOS-Atypia.*

enregistrés dans la dernière itération ne sont pas assez performants sur les bases MITOS-Atypia et ICIAR 2018-A, où les meilleurs résultats sur la base de validation ont été obtenus dans la deuxième (81 %) et la quatrième époque (92 %), respectivement (figures 6.10 et 6.11). Enfin, la stratégie de votre entre les patchs a prouvé son efficacité, où nous avons remarqué une différence considérable entre les résultats sur les bases de validation et de test pour les bases lymphoma (95.57 %, 100 %), breakhis-8c (89.98 %, 93.73 %) et MITOS-Atypia (69.78 %, 74.79 %).

Base d'apprentissage	Validation (%)	Test (%)
Lymphoma	95.57	100
Breakhis-2c	96.31	96.74
Breakhis-8c	89.98	93.73
MITOS-Atypia	69.78	74.79
Pcam	84.83	–
NCT-CRC-HE-100K-NONORM	74,64	–
ICIAR2018-A	69.84	62.5

TABLE 6.20 – *La précision des modèles entrainés à partir des initialisations aléatoires.*

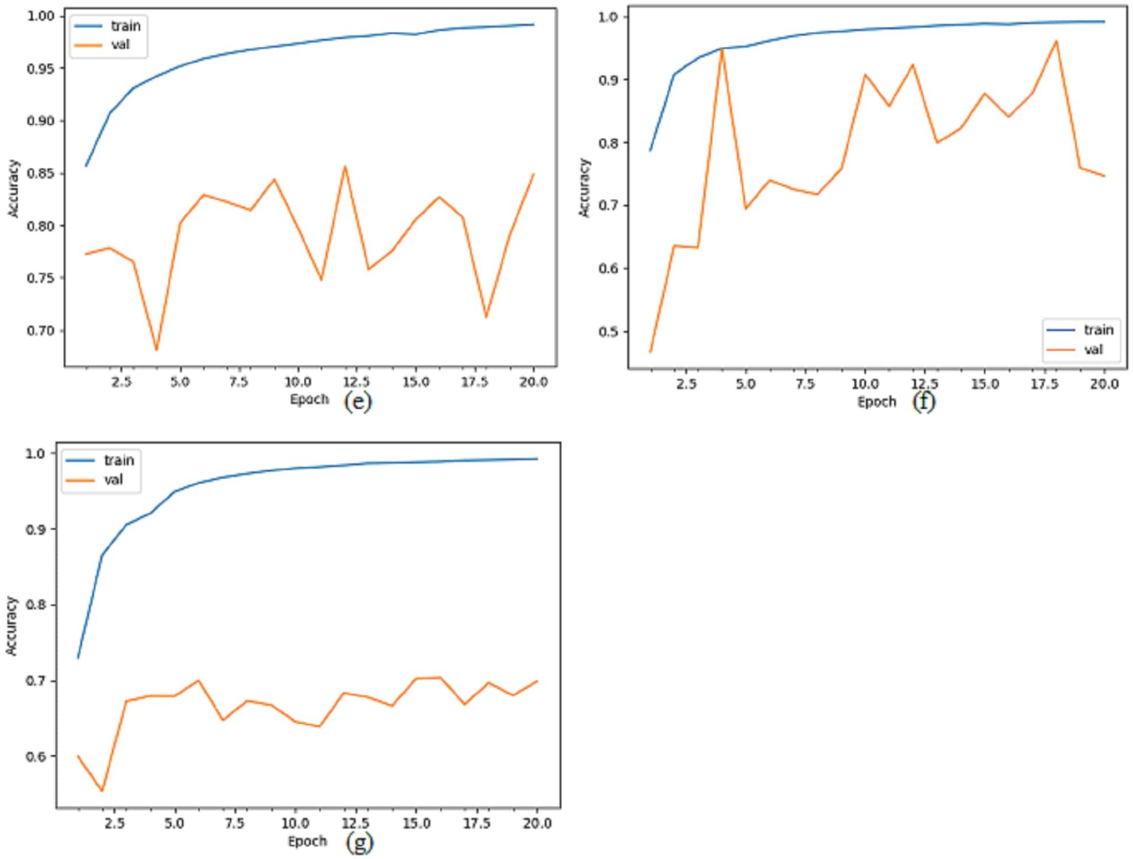

FIGURE 6.11 – *La convergence de la précision des modèles entrainés sur les bases (e) Pcam, (f) NCT-CRC-HE-100K-NONORM, et (g) ICIAR 2018-A.*

Apprentissage transféré

Dans cette partie, nous avons exploité les modèles générés par l'apprentissage à partir des initialisations aléatoires. Dans l'étape de fine tuning, nous avons varié dans le nombre N des couches à réajuster, où $N \in \{5, 10, 50, 100\}$ et $N \leq 312$. L'objectif de cette variation est d'étudier l'influence de la profondeur de fine tuning à bases modèles entrainés sur des bases distantes (ImageNet) et non distantes S_T.

1. **Apprentissage transféré à la base d'apprentissage cible CRC**
   Premièrement, nous avons testé les modèles pré-entrainés sur la base cible CRC. Le tableau 6.21 présente les résultats obtenus en termes de précision sur les bases d'apprentissage et de validation. Ces résultats indiquent une relation entre le nombre de couches réajustées et la différence entre les résultats sur les bases d'apprentissage et de validation, où cette différence est plus large pour les grandes valeurs de $N \in \{50, 100\}$.

   La figure 6.12 résume les résultats obtenus sur la base de validation. Les courbes valident cette relation pour tous les modèles sources sauf breakhis-8c et NCT-CRC-HE-100K-NONORM, où nous avons remarqué une légère diminution dans la précision de N = 50 à N = 100.

   Cette figure indique aussi une variance remarquable entre les courbes des modèles breakhis-8c et breakhis-2c. Ces résultats montrent l'in-

Nombre de couches réajustées		5	10	50	100
Lymphoma	Apprentissage	85.76%	83.71%	94.97%	98.18%
	Validation	84.78%	84.72%	87.56%	89.94%
Breakhis-2c	Apprentissage	83.51%	83.07%	94.20%	98.27%
	Validation	82.33%	82.61%	84.33%	87.44%
Breakhis-8c	Apprentissage	92.60%	93.37%	96.18%	96.61%
	Validation	92.44%	92.50%	94.00%	93.56%
MITOS-Atypia	Apprentissage	89.34%	91.79%	96.98%	98.98%
	Validation	88.89%	88.94%	91.78%	92.83%
Pcam	Apprentissage	84.5%	85.09%	98.18%	99.63%
	Validation	83.72%	83.89%	89.61%	90.67%
ICIAR2018-A	Apprentissage	94.16%	94.22%	98.77%	99.11%
	Validation	93.39%	93.44%	95.11%	95.28%
NCT-CRC-HE-100K-NONORM	Apprentissage	93.95%	93.98%	99.44%	99.72%
	Validation	93%	93.06%	94.11%	93.28%
ImageNet	Apprentissage	99.07%	98.58%	99.97%	99.85%
	Validation	86.56%	86.94%	91.50%	92.61%

TABLE 6.21 – *La précision des modèles sources réajustés sur la base d'apprentissage CRC.*

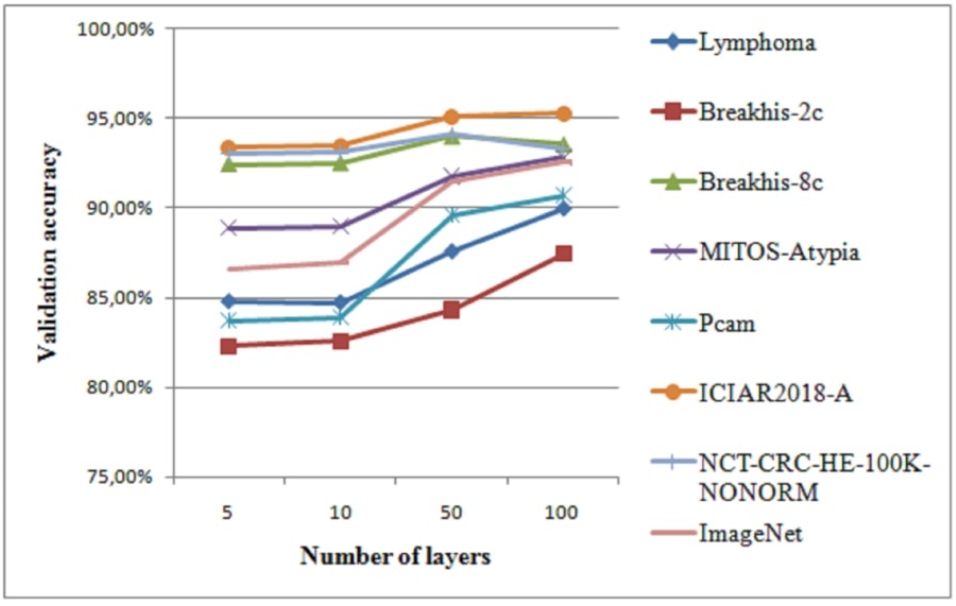

FIGURE 6.12 – *Les courbes de précision des modèles réajustés sur la base d'apprentissage CRC.*

fluence du nombre de classes sur le processus d'apprentissage transféré. Pour une première expérimentation, nous avons suggéré que les

bons résultats sont obtenus entre les tâches similaire en termes de nombre de classes (C = 8).

L'étude comparative entre les courbes de convergence prouve l'efficacité des modèles sources : ICIAR2018-A, Breakhis-8c, NCT-CRC-HE-100K-NONORM, et MITOS-ATYPIA par rapport au modèle ImageNet, où le modèle ICIAR2018-A a atteint les meilleurs résultats (95.28 %).

Les résultats obtenus indiquent aussi que, malgré la ressemblance des bases cible CRC et source NCT-CRC-HE-100K-NONORM, le processus d'apprentissage transféré à CRC était plus efficace à partir de la base source ICIAR2018-A par rapport à NCT-CRC-HE-100K-NONORM. Cela prouve l'efficacité du modèle source ICIAR2018-A dans le processus l'apprentissage transféré en termes généralisation.

2. **Apprentissage transféré à la base d'apprentissage cible BBHC-2015**
   Deuxièmement, nous avons testé les modèles pré-entrainés sur la base cible BBHC-2015.

   Le tableau 6.22 résume les résultats obtenus sur les bases d'apprentissage et de validation, et les résultats du vote majoritaire entre les patchs de la base de validation. Dans ce cadre, nous avons remarqué une grande différence entre les bases d'apprentissage et de validation lors du réajustement profond N=100.

   La figure 6.13 présente les résultats obtenus sur la base de validation. Elle illustre une perte dans la précision lors du réajustement des 10 ou des 50 dernières couches, tandis que le réajustement des 100 dernières couches a amélioré les résultats.

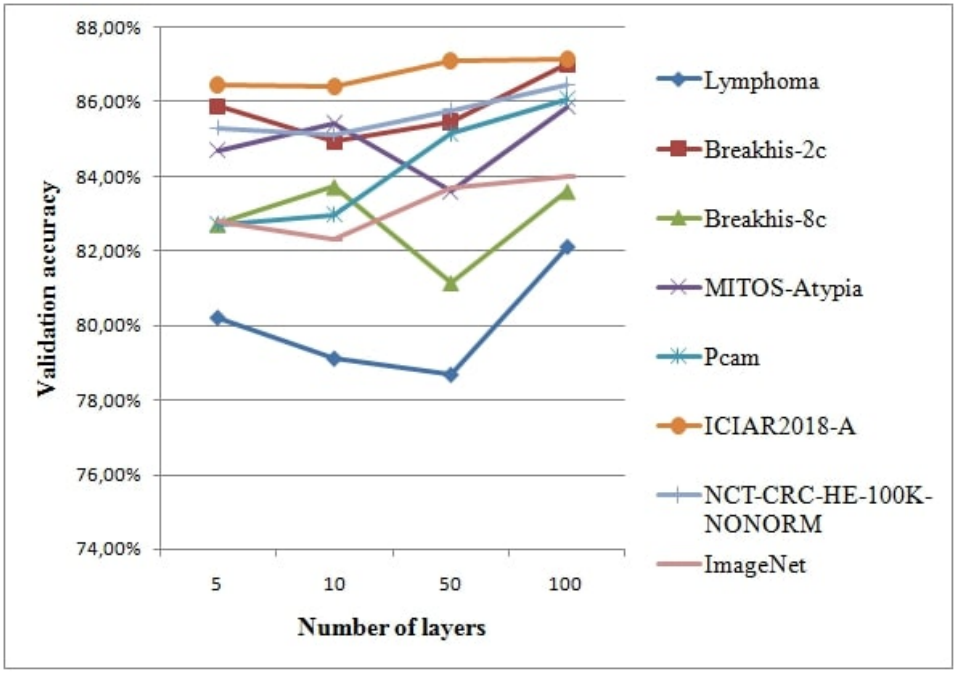

FIGURE 6.13 – *Les courbes de précision des modèles réajustés sur la base d'apprentissage BBHC-2015.*

L'étude comparative entres les courbes montre l'efficacité des modèles sources : breakhis-2c, MITOS-Atypia, ICIAR 2018-A, Pcam, et NCT-CRC-HE-100K-NONORM par rapport au modèle ImageNet. Ces ré-

Nombre de couches réajustées		5	10	50	10
Lymphoma	Apprentissage	91.31%	95.18%	99.09%	99.54%
	Validation	80.21%	79.12%	78.69%	82.10%
	Vote	86.84%	81.58%	81.58%	86.84%
Breakhis-2c	Apprentissage	96.00%	95.84%	98.91%	99.65%
	Validation	85.88%	84.95%	85.48%	87.00%
	Vote	89.47%	89.47%	86.84%	92.11%
Breakhis-8c	Apprentissage	99.35%	99.28%	99.87%	99.94%
	Validation	82.72%	83.73%	81.14%	83.61%
	Vote	86.84%	89.47%	86.84%	86.84%
MITOS-Atypia	Apprentissage	96.95%	97.30%	98.72%	99.35%
	Validation	84.71%	85.42%	83.61%	85.86%
	Vote	89.47%	92.11%	89.47%	89.47%
Pcam	Apprentissage	94.61%	93.43%	99.58%	99.86%
	Validation	82.72%	82.97%	85.15%	86.07%
	Vote	84.21%	81.58%	92.11%	92.11%
ICIAR2018-A	Apprentissage	99.51%	99.37%	95.39%	99.95%
	Validation	86.47%	86.42%	87.09%	87.15%
	Vote	92.11%	89.47%	89.47%	89.47%
NCT-CRC-HE-100K-NONORM	Apprentissage	96.39%	97.47%	99.34%	99.85%
	Validation	85.29%	85.13%	85.78%	86.47%
	Vote	89.47%	89.47%	92.11%	92.11%
ImageNet	Apprentissage	99.55%	99.62%	99.99%	99.98%
	Validation	82.81%	82.32%	83.71%	84.01%
	Vote	86.84%	86.84%	86.84%	89.47%

TABLE 6.22 – *La précision des modèles sources réajustés sur la base d'apprentissage BBHC-2015.*

sultats valident les hypothèses liées à l'efficacité de la stratégie d'apprentissage transféré entre les bases histopathologiques. De même que les expérimentations précédentes sur CRC, nous avons remarqué une grande différence entre les courbes breakhis-8c et breakhis-2c. Dans cette expérimentation, le modèle breakhis-2c était plus efficace par rapport breakhis-8c. Cela est justifié par la similarité des bases breakhis-2c et BBHC-2015 en termes de nombre de classes.

Enfin, le tableau comparatif illustre l'efficacité de la technique de vote majoritaire entre patchs. Par exemple, les résultats sur la base Pcam-50 se sont améliorés de 85.15% à 92.11% par le processus de vote. En résumé, ces expérimentations valident l'efficacité du modèle de base ICIAR 2018-A en apprentissage transféré.

3. **Apprentissage transféré à la base d'apprentissage cible KIMIA-**

PATH960

De la même manière, nous avons testé les modèles pré-entrainés sur la base cible KIMIA-PATH960.

Le tableau 6.23 résume les résultats de fine tuning sur les bases d'apprentissage et de validation et la figure 6.14 présente les résultats obtenus sur la base de validation. Cette figure illustre aussi la grande variance entre les courbes des modèles breakhis-8c et breakhis-2c, où le modèle breakhis-8c était plus performant. Ces résultats indiquent l'importance du nombre de classes lors du processus d'apprentissage transféré. Le modèle breakhis-8c est plus adapté aux bases cibles caractérisées par un nombre important de classes. D'autre part, le modèle breakhis-2c est plus approprié aux bases cibles qui sont caractérisées par un nombre réduit de classes.

Nombre de couches réajustéess		5	10	50	10
Lymphoma	Apprentissage	88.38%	90.78%	97.52%	98.06%
	Validation	88.02%	89.84%	92.45%	94.79%
Breakhis- 2c	Apprentissage	85.74%	84.10%	98.49%	96.85%
	Validation	83.85%	83.85%	91.41%	92.97%
Breakhis-8c	Apprentissage	96.79%	93.75%	100%	100%
	Validation	94.01%	93.75%	95.57%	95.05%
MITOS-Atypia	Apprentissage	92.71%	93.07%	98.39%	97.75%
	Validation	92.71%	92.97%	96.09%	96.88%
Pcam	Apprentissage	87.52%	87.08%	98.73%	99.83%
	Validation	85.42%	85.68%	93.23%	93.75%
ICIAR2018-A	Apprentissage	95.68%	95.78%	100%	98.46%
	Validation	94.27%	95.31%	98.18%	98.18%
NCT-CRC-HE-100K-NONORM	Apprentissage	95.84%	97.05%	99.16%	99.4%
	Validation	95.31%	95.31%	96.88%	97.14%
ImageNet	Apprentissage	99.83%	99.83%	100.00%	100.00%
	Validation	90.63%	91.67%	95%	94.53%

TABLE 6.23 – *La précision des modèles sources réajustés sur la base d'apprentissage KIMIA-PATH96.*

En résumé, dans ces expérimentations, le modèle ICIAR2018-A était plus performant sur la base de validation par rapport aux autres modèles sources.

4. **Apprentissage transféré à la base d'apprentissage cible CRC-VAL-HE-7K**

Enfin, nous avons testé les modèles pré-entrainés sur la base cible CRC-VAL-HE-7K.

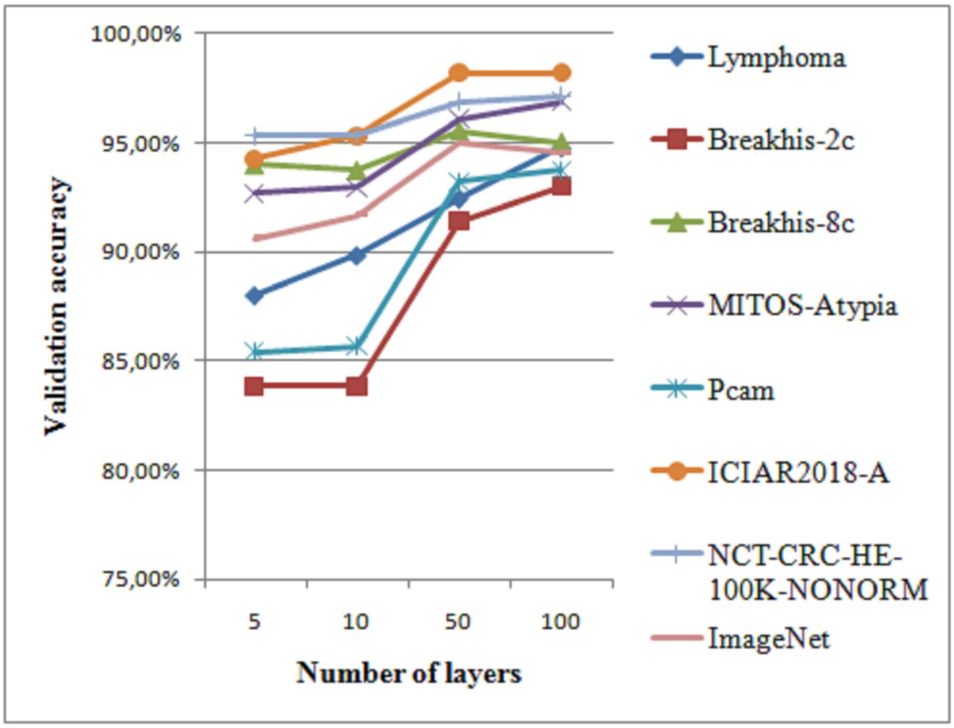

FIGURE 6.14 – *Les courbes de précision des modèles réajustés sur la base d'apprentissage KIMIA-PATH960.*

Le tableau 6.24 présente les résultats de réajustement sur les bases d'apprentissage et de validation, et la figure 6.15 résume les résultats obtenus sur la base de validation. Cette figure indique que l'augmentation de la profondeur de réajustement améliore la précision de tous les modèles sauf le modèle breakhis-8c. En plus, ces expérimentations valident les hypothèses présentées auparavant sur le rapport entre le nombre de classes dans la base source (breakhis-8c : 8) et cible (CRC-VAL-HE-7K : 9), et cela explique l'efficacité du modèle breakhis-8c par rapport à breakhis-2c.

Dans ces expérimentations, nous avons pensé que les meilleurs résultats seront obtenus par le modèle source NCT-CRC-HE-100K-NONORM, car les dataset NCT-CRC-HE-100K-NONORM et CRC-VAL-HE-7K ont été collectées à partir des mêmes laboratoires et annotées par les mêmes catégories. En revanche, il était surprenant que le modèle ICIAR2018-A a atteint les meilleurs résultats. Ces résultats indiquent l'importance de généralisation même en apprentissage transféré entre des bases d'apprentissage apparentant au même domaine d'application. En résumé, le modèle ICIAR2018-A a atteint la meilleure précision (99,13%) par rapport au modèle ImageNet et les autres modèles entrainés sur les bases histopathologiques.

Temps de traitement

Le tableau 6.25 résume le temps de traitement en apprentissage à partir des initialisations aléatoires et en apprentissage transféré sur les bases sources et cibles, respectivement. Il illustre la grande complexité temporelle de apprentissage à partir des initialisations aléatoires, où le réseau

Nombre de couches réajustéess		5	10	50	10
Lymphoma	Apprentissage	92.05%	92.39%	98.05%	99.65%
	Validation	91.82%	91.78%	95.72%	96.59%
Breakhis- 2c	Apprentissage	87.93%	91.83%	97.31%	99.37%
	Validation	90.60%	90.84%	95.09%	95.51%
Breakhis-8	Apprentissage	97.97%	98.78%	99.70%	99.88%
	Validation	97.98%	99.12%	98.40%	98.61%
MITOS-Atypia	Apprentissage	96.47%	96.38%	97.17%	99.54%
	Validation	95.75%	95.68%	97.11%	97.81%
Pcam	Apprentissage	93.29%	93.46%	99.65%	99.98%
	Validation	92.44%	92.93%	96.59%	97.53%
ICIAR2018-A	Apprentissage	97.73%	98.65%	99.54%	99.7%
	Validation	97.49%	97.77%	98.89%	99.13%
NCT-CRC-HE-100K-NONORM	Apprentissage	98.24%	98.58%	99.46%	99.7%
	Validation	97.74%	97.6%	98.15%	98.5%
ImageNet	Apprentissage	99.07%	99.41%	99.95%	100.00%
	Validation	96.45%	96.62%	98.54%	99.10%

TABLE 6.24 – *La précision des modèles sources réajustés sur la base d'apprentissage CRC-VAL-HE-7K.*

inception-v3 exige une durée de un à deux jours dans 20 époques. D'autre part, l'apprentissage transféré est moins exigeant en termes de complexité temporelle, où le réajustement de 5 à 100 couches dans 100 époques dure moins de 3 heures. Ces résultats valident l'efficacité des techniques d'apprentissage transféré et de fine tuning dans l'optimisation de la complexité temporelle.

Comparaison et discussion

En résumé, ces résultats valident l'efficacité de la méthode d'apprentissage transféré entre les bases histopathologiques.

Les modèles ICIAR 2018-A, NCT-CRC-HE-100K-NONORM, et MITOS-Atypia étaient plus performants par rapport au modèle ImageNet lors du réajustement sur les bases cibles CRC, BBHC-2015, et KIMIA-PATH960. D'autre part, la performance des modèles breakhis-8c et breakhis-2c dépend de leur corrélation à la tâche cible. Dans cette étude, nous avons remarqué que le modèle source (breakhis-8c ou breakhis-2c) est plus performant sur la base cible si le nombre de classes dans la base source est proche du nombre de classes dans la base cible.

La figure 6.16 illustre la différence entre les précisions des modèles ICIAR 2018-A et ImageNet. Elle montre que le modèle ICIAR 2018-A a atteint les meilleurs résultats sur toutes les tâches cibles, notamment sur

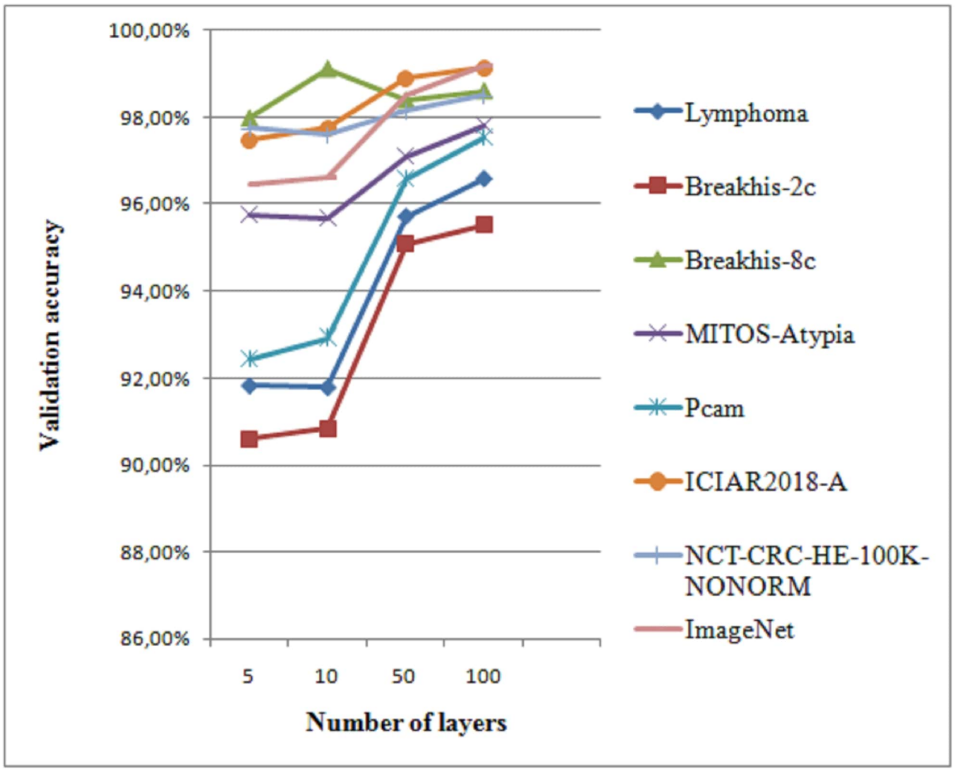

FIGURE 6.15 – *Les courbes de précision des modèles réajustés sur la base d'apprentissage CRC-VAL-HE-7K.*

Base d'apprentissage		Temps de traitement (jours/H :MIN :S)
Apprentissage à partir des initialisations aléatoires	Lymphoma	2 jours 05 :40 :57
	Breakhis- 2c	1 jour 15 :05 :46
	Breakhis-8c	1 jour 15 :12 :43
	Pcam	1 jour 01 :46 :32
	ICIAR2018-A	2 jours 14 :09 :17
	NCT-CRC-HE-100K	2 jours 00 :21 :04
	ImageNet	22 :35 :44
Apprentissage transféré	CRC	00 :57 :56
	BBHC-2015	02 :47 :18
	KIMIA-PATH96	00 :12 :15
	CRC-VAL-HE-7K	01 :22 :34

TABLE 6.25 – *Le temps de traitement de l'apprentissage à partir des initialisations aléatoires et de l'apprentissage transféré.*

les bases CRC et BBHC-2015. En plus, les courbes indiquent que l'augmentation de la profondeur de réajustement réduit la variance entre la performance de ces modèles. Nous expliquons ces résultats par la distance entre les tâches sources et cibles, où les modèles entrainés sur des bases distantes telles que ImageNet nécessitent un réajustement profond pour obtenir des résultats satisfaisants.

Malgré la faible précision du modèle ICIAR 2018-A (69.84 %) sur la base source, ce modèle a prouvé son efficacité en tant qu'un modèle de

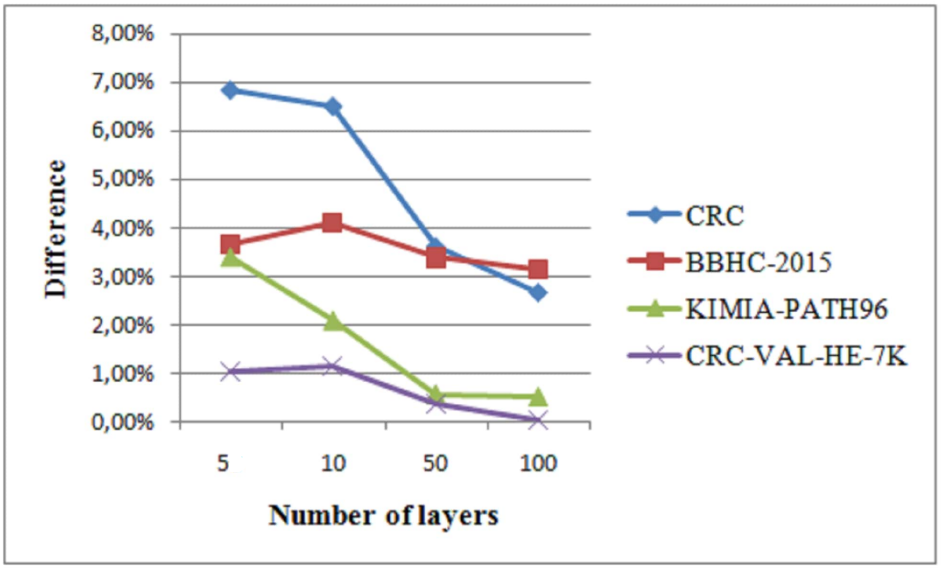

FIGURE 6.16 – *La différence entre les résultats obtenus à base des modelés sources entrainés sur les bases ICIAR 2018-A et ImageNet.*

base sur les autres tâches histopathologiques cibles. D'autre part, les modèles très performants sur leurs bases sources comme Lymphoma (100 %) ont obtenu de mauvais résultats sur les bases cibles. En général, l'étude comparative entre les résultats de tous les modèles suggèrent que les modèles performants sur les bases sources ont subi à un sur-apprentissage sur la tâche source, et cela a réduit leur efficacité en terme de généralisation lors d'un réajustement sur les autres tâches cibles. En plus, nous avons remarqué que le réajustement du modèle NCT-CRC-HE-100K-NONORM sur des bases cibles similaires (CRC-VAL-HE-7K, CRC) était moins efficace par rapport au réajustement à partir modèle ICIAR 2018-A, et cela indique l'importance de la généralisation dans le processus d'apprentissage transféré entre les bases histopathologiques.

Enfin, nous avons comparé les précisions obtenues sur les bases cibles (CRC, BBHC-2015, CRC-VAL-HE-7K, et KIMIA-PATH960) et les autres résultats de l'état de l'art (tableau 6.26). Le but principal de cette contribution est de construire des modèles de base pour le processus d'apprentissage transféré, pour cette raison, nous nous somme intéressés seulement aux bases cibles lors de la comparaison.

Les stratégies testées sur la base CRC sont de trois types : ML [Kather *et al.* 2016, Ly *et al.* 2017], DL [Ciompi *et al.* 2017], et des méthodes hybrides [Cascianelli *et al.* 2018, Wang *et al.* 2017a]. Dans ce cadre, différentes techniques d'extraction de caractéristiques ont été exploitées : handcrafted features [Kather *et al.* 2016] et les réseaux CNN en tant qu'extracteurs de caractéristiques ((BBCN [Wang *et al.* 2017a], VGGF-F, VGG-S, VGG-VD-16 [Cascianelli *et al.* 2018]).

En plus, plusieurs travaux ont automatisé la classification de la base d'apprentissage BBHC-2015 par les techniques ML [Hammoudi *et al.* 2018] et différentes architectures de types DL : CNN pour l'extraction des caractéristiques [Vo *et al.* 2019], Deep Spatial Fusion Network [Huang & Chung 2018], Incremental Boosting Convolution

Base d'apprentissage	Référénce	Précision(%)
CRC	[Kather *et al.* 2016]	87.4
	[Ly *et al.* 2017]	47.33
	[Ciompi *et al.* 2017]	79.66
	[Cascianelli *et al.* 2018]	85
	[Wang *et al.* 2017a]	92.6
	[Dif & Elberrichi 2020d]	95.28
BBHC-2015	[Araújo *et al.* 2017]	83.3
	[Hammoudi *et al.* 2018]	75
	[Vo *et al.* 2019]	99.5
	[Huang & Chung 2018]	88.9
	[Alom *et al.* 2019]	99.05
	[Dif & Elberrichi 2020d]	92.11
KIMIA-PATH960	[Kumar *et al.* 2017]	94.74
	[Alhindi *et al.* 2018]	90.52
	[Dif & Elberrichi 2020d]	98,18
CRC-VAL-HE-7K	[Kather *et al.* 2019]	99
	[Dif & Elberrichi 2020d]	99.13

TABLE 6.26 – *La comparaison entre les résultats de l'état de l'art et les résultats obtenus.*

Networks (ICBN) [Alom *et al.* 2019], et Inception Recurrent Residual Convolutional Neural Network [Araújo *et al.* 2017].

Enfin, pour la base KIMIA-PATH960, les techniques proposées ont prouvé l'efficacité de LBP par rapport aux CNN en tant qu'extracteurs de caractéristiques [Alhindi *et al.* 2018] et de la technique bag-of-visual words (BoVW) par rapport à LBP [Kumar *et al.* 2017].

En résumé, l'étude comparative a prouvé l'efficacité de la méthode proposée en termes de généralisation et la haute qualité de la stratégie d'apprentissage transféré entre les bases d'apprentissage histopathologiques, où nous avons atteint des résultats satisfaisants sur les bases CRC (95.28 %) et KIMIA-PATH960 (98.18 %).

6.3.6 Conclusion

Cette contribution présente une nouvelle méthode de fine tuning entre les bases histopathologiques. L'objectif principal était d'examiner pour la première fois l'effet de l'apprentissage transféré entre des bases histopathologiques et de tester les hypothèses liées à l'efficacité de cette technique entre des bases non distantes. Dans ce cadre, nous avons entrainé le réseau profond InceptionV3 sur 6 bases d'apprentissages histopathologiques sources (Lymphoma, Breakhis, MITOS-Atypia, ICIAR 2018-A, Pcam, et NCT-CRC-HE-100K-NONORM), ensuite, nous avons exploité les modèles générés en apprentissage transféré sur d'autres bases histopathologiques cibles (CRC, BBHC-2015, CRC-VAL-HE-7K, et KIMIA-PATH960), où un sous ensemble des dernières couches N a été réajusté ($N \in \{5, 10, 50, 100\}$).

Les expérimentations effectuées ont confirmé l'efficacité de l'apprentissage transféré entre les bases non distantes, où les modèles ICIAR 2018-A,

MITOS-Atypia, et NCT-CRC-HE-100K-NONORM ont atteint de bons résultats par rapport aux modèles entrainés sur la base distante ImageNet. En plus, l'étude comparative entre plusieurs niveaux de réajustement a montré l'efficacité du réajustement profond ($N \geq 50$) sur la précision.

L'étude expérimentale a prouvé l'importance de la généralisation même entre des bases appartenant au même domaine, où nous avons observé l'efficacité du processus d'apprentissage transféré entre des bases histopathologiques distantes par rapport à d'autres moins distante. Nous avons remarqué que les modèles de haute qualité sur les bases sources ont subi à un sur-apprentissage sur ces bases, et cela a réduit leur efficacité en généralisation sur d'autres bases cibles. D'autre part, les modèles faibles sur la base source étaient plus prometteurs en raison de leur efficacité en généralisation sur d'autres bases cibles.

Malgré les avantages cités auparavant sur l'approche proposée, et contrairement aux modèles ImageNet, les modèles générés sont valables pour un apprentissage transféré seulement à d'autres bases histopathologiques.

Conclusion générale

Cette thèse concerne le traitement des images par les méthodes d'apprentissage profond. Le choix des méthodes d'apprentissage profond (DL) a été motivé par leurs avantages dans le traitement des objets non structurés, comme les images. Dans ce contexte, nous nous intéressons aux réseaux de neurones convolutifs (CNN) en raison de leur adaptation à la classification des images. Ces derniers sont caractérisés par des modules d'extraction de caractéristiques intégrés (couches de convolutions) par rapport aux méthodes d'apprentissage automatique classiques qui exigent un prétraitement d'extraction des handcrafted features.

La première architecture de type CNN en classification supervisée trouve ses racines dans les années 1998. La capacité limitée des ordinateurs à cette époque et le nombre réduit des images disponibles ont limité l'exploitation de ces architectures. En parallèle, les méthodes ML ont connu un grand intérêt en raison de leurs exigences réduites en termes de capacité de traitement et volumes de données. Jusqu'à 2012, où le meilleur taux d'erreur de l'architecture AlexNet sur la base d'apprentissage ImageNet, et la capacité des GPU dans la réduction du temps de calcul, ont encouragé la communauté de traitement des images à réutiliser et à optimiser ces réseaux pour les exploiter dans les applications de vision par ordinateur.

Dans l'état de l'art, plusieurs travaux étaient proposés pour la classification des images par les méthodes DL, et notamment les CNN. Le but principal de ces travaux de recherche était d'améliorer la performance de ces réseaux sur la tâche traitée. Dans ce contexte, bien que notre formation soit axée sur l'informatique, nous avons choisi de nous spécialiser à la classification des images médicales, et précisément les images histopathologiques. Nous étions intéressé par l'optimisation l'optimisation de la performance des réseaux CNN pour la classification de ces images. L'analyse des images histolopatologiques est une tâche non triviale dans le diagnostic de plusieurs types de cancers. L'examen manuel de ces images a plusieurs enjeux liés à la subjectivité des décisions des pathologistes et l'inter-variabilité des images collectées de différents laboratoires.

Le choix de ce domaine d'application était motivé par la proposition des systèmes d'aide au diagnostic basés sur les méthodes DL. L'objectif de ces systéme est d'assister les pathologistes dans leurs décisions et d'éviter donc les décisions subjectives.

L'objectif principal des approches proposées était de résoudre les différents problèmes liés à la classification des images histopathologiques par les méthodes DL.

Dans nos travaux, le domaine médical a été une nouvelle expérience puisque cette thèse se situe à l'intersection de l'informatique et du domaine médical. Pour comprendre la structure et les techniques de traite-

ment des images histopathologiques, nous avons effectué une étude approfondie sur la procédure manuelle effectuée par les pathologistes et tout le processus réalisé dans la coloration, le prétraitement et la numérisation des lames de verre. Dans ce cadre, l'un de nos travaux de recherche [Dif & Elberrichi 2020a] présente un aperçu complet sur les méthodes d'apprentissage profond pour la détection de la mitose à partir des images histopathologiques. Il désigne un bon guide pour les informaticiens pour la compréhension du background médical lié à la détection de la mitose. Il présente aussi les différents techniques DL proposées dans l'état de l'art, leurs inconvénients, et les solutions suggérés.

Dans le cadre des études expérimentales, nous avons proposé plusieurs approches [Dif & Elberrichi 2020d, Dif & Elberrichi 2020c, Dif & Elberrichi 2020b]. Les deux premiers travaux de recherche [Dif & Elberrichi 2020c, Dif & Elberrichi 2020b] exploitent les avantages des méthodes ensemblistes et la dernière contribution [Dif & Elberrichi 2020d] utilise la stratégie d'apprentissage transféré. Nous avons exploité les méthodes ensemblistes en raison de leurs avantages liés à la réduction de la variance élevée des méthodes DL et la résolution des différents problèmes de sur-apprentissage. D'autre part, l'apprentissage transféré permet de réduire les exigences matérielles et le temps de calcul des méthodes DL, ainsi que les problèmes de sur-apprentissage sur les volumes limités de données médicales.

Dans la première contribution [Dif & Elberrichi 2020b], nous avons proposé un framework basé sur deux méthodes ensemblistes statiques et l'architecture MobileNet. L'objectif principal de ce travail de recherche était d'exploiter un maximum de méthodes de régularisation afin d'éviter les problèmes de sur-apprentissage sur les volumes limités des images histopathologiques. Les deux méthodes ensemblistes combinent entre les N derniers ou les N meilleurs modèles enregistrés dans plusieurs points d'apprentissage. Le framework proposé a été évalué sur la base d'apprentissage histopathologique Lymphoma. L'étude expérimentale a montré que la combinaison de 3 à 4 modèles est suffisante pour améliorer les performances. D'autre part, les méthodes de sélection statique exploitées peuvent sélectionner des modèles qui s'approchent dans leurs décisions, et cela peut réduire la performance de l'ensemble, car la diversité est parmi les critères importants pour créer une bonne coopération entre les membres d'un ensemble.

La deuxième contribution [Dif & Elberrichi 2020c] était proposée pour résoudre les inconvénients des méthodes ensemblistes statiques. Afin d'exploiter les avantages des méthodes ensemblistes dynamiques, nous avons proposé une méthode ensembliste dynamique basée sur la métaheuristique PSO et la stratégie d'apprentissage transféré. PSO était exploité afin de sélectionner le sous ensemble de modèles approprié. Le sous ensemble est sélectionné en fonction de la performance d'un vote ou de la moyenne non pondérée des modèles appartenant à cet ensemble. L'avantage de la méthode de sélection dynamique par rapport aux méthodes statiques se résume dans sa dépendance de la performance de l'ensemble au lieu de la performance de chaque modèle appartenant à cet ensemble, et donc les modèles sont sélectionnés en fonction de leur coopération et diversité. L'étude comparative entre la méthode de sélection dynamique

et deux autres méthodes de sélection statique a prouvé l'efficacité de cette technique, où les résultats obtenus montrent qu'un sous ensemble sélectionné par PSO (composé de cinq à huit modèles) est plus performant par rapport à l'ensemble original (composé de 100 modèles).

La troisième contribution [Dif & Elberrichi 2020d] se base sur les techniques d'apprentissage transféré et de fine-tuning. Ce travail de recherche effectue un apprentissage transféré entre des bases d'apprentissage histopathologiques au lieu d'utiliser les modèles pré-entrainés sur la base ImageNet. Dans ce contexte, nous avons entrainé le réseau profond InceptionV3 sur 6 bases d'apprentissages histopathologiques sources, ensuite, nous avons exploité ces réseaux en apprentissage transféré sur d'autres bases histopathologiques cibles, où un sous ensemble des dernières couches N était réajusté. Les expérimentations effectuées ont confirmé l'efficacité de l'apprentissage transféré entre les bases non distantes par rapport à la base distante (ImageNet). D'autre part, la généralisation est importante même entre des bases appartenant au même domaine, où nous avons observé l'efficacité du processus d'apprentissage transféré entre des bases histopathologiques distantes par rapport à d'autres moins distantes.

Perspectives

À base des résultats obtenus dans les différentes approches, nous pouvons conclure que l'utilisation des méthodes ensemblistes permet de combiner la décision de plusieurs apprenants, et donc améliore la pertinence des résultats et réduit la variance élevée des réseaux DL. En plus, l'apprentissage transféré entre les bases appartenant au même domaine améliore la performance de classification, réduit les exigences des modèles utilisés en termes de temps de calcul, et diminue les différents problèmes liés au sur-apprentissage.

Malgré l'efficacité des méthodes ensemblistes, ces derniers exigent un espace de stockage élevé pour stocker les différents modèles apparentants à l'ensemble, en plus, le temps d'inférence total constitue la somme des temps d'inférence de ces modèles. D'autre part, dans les systèmes et les applications du monde réel, et notamment dans les systèmes de vision intégrés, le temps et l'espace désignent deux paramètres importants à minimiser. Ainsi, lors des futures contributions, il serait utile de s'intéresser aux techniques d'apprentissage transféré, de fine tuning, et l'extraction des caractéristiques par les réseaux CNN.

Nous sommes conscients que la troisième contribution qui montre l'efficacité de l'apprentissage transféré entre les bases histopathologiques ne constitue qu'un début aux réflexions qui devront se poursuivre dans le futur. Dans ce contexte, nous pensons à quelques issues de recherche :

1. **Perspectives à court terme**
    - La mise en place d'une approche d'apprentissage transféré à partir des bases histopathologiques classifiées par des classes synthétiques au lieu des classes réelles. Nous devons rappeler que l'apprentissage transféré s'effectue à partir d'un modèle entrainé sur de grands volumes de données et l'annota-

tion manuelle de ces volumes est une tâche fastidieuse. En plus, dans ce processus, seulement les classes réelles de la tâche cible sont intéressantes. Dans ce cadre, nous proposons une approche qui hybride entre les techniques de clustering et de classification. Le clutering permet de créer les classes synthétiques, et la classification génère les modèles de bases qui sont entrainés sur les bases classifiées dans l'étape précédente. Nous sommes entrain de concevoir une méthode qui hybride entre les algorithmes MCO-Stream (Multi-Objective Clustering) et Inception, et comme extension à ce travail de recherche, nous penserons à hybrider entre un réseau DL de clustering et un CNN.

2. **Perspectives à long terme**
   - Il serait intéressant d'exploiter le modèle Inception entrainé sur la base ICAR-2018 comme module d'extraction de caractéristiques sur les bases histopathologique au lieu d'utiliser les modèles classiques entrainés sur ImageNet. Nous devons rappeler que ce modèle a prouvé son efficacité [Dif & Elberrichi 2020d] par rapport aux modèles ImageNet en apprentissage transféré.
   - Développer une méthode dynamique de fine tuning. Cette méthode propose les couches appropriées à réajuster sur la tâche cible. Cela peut être effectué à l'aide des méthodes stochastiques comme les métaheuristiques. L'originalité de cette méthode par rapport aux autres techniques de fine tuning classiques est qu'elle propose un sous ensemble de couches à réajuster en fonction de la tâche traitée au lieu de réajuster seulement les N dernières couches.
   - Dans d'autres contributions, il serait intéressant d'exploiter les techniques de clustering pour l'extraction des patchs pertinents dans la phase de test. Les méthodes classiques utilisent la méthode de fenêtre coulissante pour l'extraction des patchs, et cela peut réduire la performance de classification à cause des patchs non discriminants.

Bibliographie

[Akkus *et al.* 2017] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin et Bradley J Erickson. *Deep learning for brain MRI segmentation : state of the art and future directions*. Journal of digital imaging, vol. 30, no. 4, pages 449–459, 2017.

[Akram *et al.* 2018] Saad Ullah Akram, Talha Qaiser, Simon Graham, Juho Kannala, Janne Heikkilä et Nasir Rajpoot. *Leveraging unlabeled whole-slide-images for mitosis detection*. In Computational Pathology and Ophthalmic Medical Image Analysis, pages 69–77. Springer, 2018.

[Al-Janabi *et al.* 2013] Shaimaa Al-Janabi, Henk-Jan van Slooten, Mike Visser, Tjeerd Van Der Ploeg, Paul J Van Diest et Mehdi Jiwa. *Evaluation of mitotic activity index in breast cancer using whole slide digital images*. PloS one, vol. 8, no. 12, 2013.

[Albarqouni *et al.* 2016] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci et Nassir Navab. *Aggnet : deep learning from crowds for mitosis detection in breast cancer histology images*. IEEE transactions on medical imaging, vol. 35, no. 5, pages 1313–1321, 2016.

[Albayrak & Bilgin 2016] Abdulkadir Albayrak et Gokhan Bilgin. *Mitosis detection using convolutional neural network based features*. In IEEE 17th International Symposium on Computational Intelligence and Informatics (CINTI), pages 335–340. IEEE, 2016.

[Alhindi *et al.* 2018] Taha J Alhindi, Shivam Kalra, Ka Hin Ng, Anika Afrin et Hamid R Tizhoosh. *Comparing LBP, HOG and deep features for classification of histopathology images*. In 2018 international joint conference on neural networks (IJCNN), pages 1–7. IEEE, 2018.

[Alom *et al.* 2019] Md Zahangir Alom, Chris Yakopcic, Mst Shamima Nasrin, Tarek M Taha et Vijayan K Asari. *Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network*. Journal of digital imaging, vol. 32, no. 4, pages 605–617, 2019.

[Amos *et al.* 2016] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan*et al. Openface : A general-purpose face recognition library with mobile applications*. CMU School of Computer Science, vol. 6, no. 2, 2016.

[An & Cho 2015] Jinwon An et Sungzoon Cho. *Variational autoencoder based anomaly detection using reconstruction probability*. Special Lecture on IE, vol. 2, no. 1, 2015.

[Antoniou *et al.* 2018] Antreas Antoniou, Amos J. Storkey et Harrison Edwards. *Augmenting Image Classifiers Using Data Augmentation Generative Adversarial Networks*. In Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III, volume 11141 of *Lecture Notes in Computer Science*, pages 594–603, 2018.

[Anwar *et al.* 2018] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami et Muhammad Khurram Khan. *Medical image analysis using convolutional neural networks : a review*. Journal of medical systems, vol. 42, no. 11, page 226, 2018.

[Araújo *et al.* 2017] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia et Aurélio Campilho. *Classification of breast cancer histology images using convolutional neural networks*. PloS one, vol. 12, no. 6, 2017.

[Arbelaez *et al.* 2010] Pablo Arbelaez, Michael Maire, Charless Fowlkes et Jitendra Malik. *Contour detection and hierarchical image segmentation*. IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 5, pages 898–916, 2010.

[Aresta *et al.* 2019] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan*et al. Bach : Grand challenge on breast cancer histology images*. Medical image analysis, vol. 56, pages 122–139, 2019.

[Azizpour *et al.* 2015] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki et Stefan Carlsson. *From generic to specific deep representations for visual recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 36–45, 2015.

[Badrinarayanan *et al.* 2017] Vijay Badrinarayanan, Alex Kendall et Roberto Cipolla. *Segnet : A deep convolutional encoder-decoder architecture for image segmentation*. IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pages 2481–2495, 2017.

[Bai *et al.* 2019] Jie Bai, Huiyan Jiang, Siqi Li et Xiaoqi Ma. *NHL Pathological Image Classification Based on Hierarchical Local Information and GoogLeNet-Based Representations*. BioMed research international, 2019.

[Beevi *et al.* 2019] K Sabeena Beevi, Madhu S Nair et GR Bindu. *Automatic mitosis detection in breast histopathology images using Convolutional Neural Network based deep transfer learning*. Biocybernetics and Biomedical Engineering, vol. 39, no. 1, pages 214–223, 2019.

[Beikman *et al.* 2013] Susan Beikman, Patricia Gordon, Shannon Ferrari, Monica Siegel, Mary Ann Zalewski et Margaret Q Rosenzweig. *Understanding the Implications of the Breast Cancer Pathology Report : A*

Case Study. Journal of the advanced practitioner in oncology, vol. 4, no. 3, pages 176–181, 2013.

[Bejnordi *et al.* 2015] Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer et Jeroen AWM van der Laak. *Stain specific standardization of whole-slide histopathological images*. IEEE transactions on medical imaging, vol. 35, no. 2, pages 404–415, 2015.

[Bejnordi *et al.* 2017] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol*et al.* *Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer*. Jama, vol. 318, no. 22, pages 2199–2210, 2017.

[Bello *et al.* 2019] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens et Quoc V Le. *Attention augmented convolutional networks*. In Proceedings of the IEEE International Conference on Computer Vision, pages 3286–3295, 2019.

[Beresford *et al.* 2006] Mark J Beresford, George D Wilson et Andreas Makris. *Measuring proliferation in breast cancer : practicalities and applications*. Breast Cancer Research, vol. 8, no. 6, pages 1–11, 2006.

[Berg *et al.* 2010] A Berg, J Deng et L Fei-Fei. *Large scale visual recognition challenge 2010*. http://www.image-net.org/challenges/LSVRC/2010/index, 2010.

[Bian *et al.* 2016] Xiao Bian, Ser Nam Lim et Ning Zhou. *Multiscale fully convolutional network with application to industrial inspection*. In 2016 IEEE winter conference on applications of computer vision (WACV), pages 1–8. IEEE, 2016.

[Bianconi *et al.* 2019] Francesco Bianconi, Jakob N Kather et Constantino C Reyes-Aldasoro. *Evaluation of Colour Pre-processing on Patch-Based Classification of H&E-Stained Images*. In European Congress on Digital Pathology, pages 56–64. Springer, 2019.

[Bishop *et al.* 1995] Christopher M Bishop*et al.* Neural networks for pattern recognition. Oxford university press, 1995.

[Bishop 2006] Christopher M Bishop. Pattern recognition and machine learning. Springer, 2006.

[Bonert & Tate 2017] Michael Bonert et Angela J Tate. *Mitotic counts in breast cancer should be standardized with a uniform sample area*. Biomedical engineering online, vol. 16, no. 1, pages 1–8, 2017.

[Bray *et al.* 2018] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre et Ahmedin Jemal. *Global cancer statistics 2018 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA : a cancer journal for clinicians, vol. 68, no. 6, pages 394–424, 2018.

[Butepage *et al.* 2017] Judith Butepage, Michael J Black, Danica Kragic et Hedvig Kjellstrom. *Deep representation learning for human motion prediction and classification*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6158–6166, 2017.

[Byra *et al.* 2018] Michał Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, ukasz Michałowski, Rafał Paluszkiewicz, Bogna Ziarkiewicz-Wróblewska, Krzysztof Zieniewicz, Piotr Sobieraj et Andrzej Nowicki. *Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images*. International journal of computer assisted radiology and surgery, vol. 13, no. 12, pages 1895–1903, 2018.

[Canziani *et al.* 2016] Alfredo Canziani, Adam Paszke et Eugenio Culurciello. *An Analysis of Deep Neural Network Models for Practical Applications*. CoRR. abs/1605.07678, 2016.

[Cascianelli *et al.* 2018] Silvia Cascianelli, Raquel Bello-Cerezo, Francesco Bianconi, Mario L Fravolini, Mehdi Belal, Barbara Palumbo et Jakob N Kather. *Dimensionality reduction strategies for cnn-based classification of histopathological images*. In International Conference on Intelligent Interactive Multimedia Systems and Services, pages 21–30. Springer, 2018.

[Chellapilla *et al.* 2006] Kumar Chellapilla, Sidd Puri et Patrice Simard. *High performance convolutional neural networks for document processing*. In International workshop on Frontiers in handwriting recognition, 2006.

[Chen & Yuille 2014] Xianjie Chen et Alan L Yuille. *Articulated pose estimation by a graphical model with image dependent pairwise relations*. In Advances in neural information processing systems, pages 1736–1744, 2014.

[Chen *et al.* 2014] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun et Alan Yuille. *Detect what you can : Detecting and representing objects using holistic models and body parts*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1971–1978, 2014.

[Chen *et al.* 2015] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy et Alan L. Yuille. *Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs*. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[Chen *et al.* 2016a] Hao Chen, Qi Dou, Xi Wang, Jing Qin et Pheng Ann Heng. *Mitosis detection in breast cancer histology images via deep cascaded networks*. In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[Chen *et al.* 2016b] Hao Chen, Xi Wang et Pheng Ann Heng. *Automated mitosis detection with deep regression networks*. In 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pages 1204–1207. IEEE, 2016.

[Chen *et al.* 2017a] Hugh Chen, Scott Lundberg et Su-In Lee. *Checkpoint ensembles : Ensemble methods from a single training process. CoRR abs/1710.03282*, 2017.

[Chen *et al.* 2017b] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy et Alan L Yuille. *Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*. IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pages 834–848, 2017.

[Chen *et al.* 2017c] Tao Chen, Shijian Lu et Jiayuan Fan. *S-CNN : Subcategory-aware convolutional networks for object detection*. IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 10, pages 2522–2528, 2017.

[Ching *et al.* 2018] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman*et al. Opportunities and obstacles for deep learning in biology and medicine*. Journal of The Royal Society Interface, vol. 15, no. 141, 2018.

[Chollet 2017] François Chollet. *Xception : Deep learning with depthwise separable convolutions*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017.

[Chopra *et al.* 2013] Sumit Chopra, Suhrid Balakrishnan et Raghuraman Gopalan. *Dlid : Deep learning for domain adaptation by interpolating between domains*. In ICML workshop on challenges in representation learning, volume 2, 2013.

[Christoph & Pinz 2016] RPW Christoph et Feichtenhofer Axel Pinz. *Spatiotemporal residual networks for video action recognition*. Advances in Neural Information Processing Systems, pages 3468–3476, 2016.

[Ciompi *et al.* 2017] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva De Souza, Alexi Baidoshvili, Geert Litjens, Bram Van Ginneken, Iris Nagtegaal et Jeroen Van Der Laak. *The importance of stain normalization in colorectal tissue classification with convolutional networks*. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pages 160–163. IEEE, 2017.

[Ciregan *et al.* 2012] Dan Ciregan, Ueli Meier et Jürgen Schmidhuber. *Multi-column deep neural networks for image classification*. In 2012 IEEE conference on computer vision and pattern recognition, pages 3642–3649. IEEE, 2012.

[Ciresan *et al.* 2011] Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella et Jurgen Schmidhuber. *Convolutional neural network committees for handwritten character classification*. In 2011 International Conference on Document Analysis and Recognition, pages 1135–1139. IEEE, 2011.

[Ciresan *et al.* 2012] Dan Ciresan, Alessandro Giusti, Luca M Gambardella et Jürgen Schmidhuber. *Deep neural networks segment neuronal membranes in electron microscopy images*. In Advances in neural information processing systems, pages 2843–2851, 2012.

[Cireşan *et al.* 2013] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella et Jürgen Schmidhuber. *Mitosis detection in breast cancer histology images with deep neural networks*. In International conference on medical image computing and computer-assisted intervention, pages 411–418. Springer, 2013.

[Codella *et al.* 2016] Noel Codella, Mehdi Moradi, Matt Matasar, Tanveer Sveda-Mahmood et John R Smith. *Lymphoma diagnosis in histopathology using a multi-stage visual learning approach*. In Medical Imaging 2016 : Digital Pathology, volume 9791, pages 131 – 137. International Society for Optics and Photonics, SPIE, 2016.

[Courjon 1990] Daniel Courjon. *Scanning tunneling optical microscopy*. In Scanning Tunneling Microscopy and Related Methods, pages 497–505. Springer, 1990.

[Dai *et al.* 2015] Jifeng Dai, Kaiming He et Jian Sun. *Boxsup : Exploiting bounding boxes to supervise convolutional networks for semantic segmentation*. In Proceedings of the IEEE International Conference on Computer Vision, pages 1635–1643, 2015.

[Dalmış *et al.* 2017] Mehmet Ufuk Dalmış, Geert Litjens, Katharina Holland, Arnaud Setio, Ritse Mann, Nico Karssemeijer et Albert Gubern-Mérida. *Using deep learning to segment breast and fibroglandular tissue in MRI volumes*. Medical physics, vol. 44, no. 2, pages 533–546, 2017.

[Das & Dutta 2019] Dev Kumar Das et Pranab Kumar Dutta. *Efficient automated detection of mitotic cells from breast histological images using deep convolution neutral network with wavelet decomposed patches*. Computers in biology and medicine, vol. 104, pages 29–42, 2019.

[Deng *et al.* 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei. *Imagenet : A large-scale hierarchical image database*. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. IEEE, 2009.

[Di Ruberto *et al.* 2015] Cecilia Di Ruberto, Giuseppe Fodde et Lorenzo Putzu. *On different colour spaces for medical colour image classification*. In International Conference on Computer Analysis of Images and Patterns, pages 477–488. Springer, 2015.

[Diba *et al.* 2017] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash et Luc Van Gool. *Weakly supervised cascaded convolutional networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 914–922, 2017.

[Dif & Elberrichi 2020a] Nassima Dif et Zakaria Elberrichi. *Deep Learning Methods for Mitosis Detection in Breast Cancer Histopathological Images : A Comprehensive Review*. In Artificial Intelligence and Machine Learning for Digital Pathology, pages 279–306. Springer, 2020.

[Dif & Elberrichi 2020b] Nassima Dif et Zakaria Elberrichi. *Efficient Regularization Framework for Histopathological Image Classification Using Convolutional Neural Networks.* International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), vol. 14, no. 4, pages 62–81, 2020.

[Dif & Elberrichi 2020c] Nassima Dif et Zakaria Elberrichi. *A New Deep Learning Model Selection Method for Colorectal Cancer Classification*. International Journal of Swarm Intelligence Research (IJSIR), vol. 11, no. 3, pages 72–88, 2020.

[Dif & Elberrichi 2020d] Nassima Dif et Zakaria Elberrichi. *A New Intra Fine-Tuning Method Between Histopathological Datasets in Deep Learning*. International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), vol. 11, no. 2, pages 16–40, 2020.

[Dong *et al.* 2018] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang et Yi-Hsuan Yang. *MuseGAN : Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment*. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[Dorigo *et al.* 1996] Marco Dorigo, Vittorio Maniezzo et Alberto Colorni. *Ant system : optimization by a colony of cooperating agents*. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 26, no. 1, pages 29–41, 1996.

[Dou *et al.* 2016] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, Defeng Wang, Vincent CT Mok, Lin Shi et Pheng-Ann Heng. *Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks*. IEEE transactions on medical imaging, vol. 35, no. 5, pages 1182–1195, 2016.

[Du *et al.* 2015] Yong Du, Wei Wang et Liang Wang. *Hierarchical recurrent neural network for skeleton based action recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1110–1118, 2015.

[Duchi *et al.* 2011] John Duchi, Elad Hazan et Yoram Singer. *Adaptive subgradient methods for online learning and stochastic optimization.* Journal of machine learning research, vol. 12, no. 7, pages 2121–2159, 2011.

[Eberhart & Kennedy 1995] Russell Eberhart et James Kennedy. *A new optimizer using particle swarm theory*. In MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, pages 39–43. IEEE, 1995.

[Engstrom *et al.* 2017] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt et Aleksander Madry. *A Rotation and a Translation Suffice : Fooling CNNs with Simple Transformations*. CoRR. abs/1712.02779, 2017.

[Esteva *et al.* 2017] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau et Sebastian Thrun. *Dermatologist-level classification of skin cancer with deep neural networks*. Nature, vol. 542, no. 7639, pages 115–118, 2017.

[Everingham *et al.*] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn et Andrew Zisserman. *The PASCAL visual object classes challenge 2007 (VOC2007) results*. http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop.

[Everingham *et al.* 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn et Andrew Zisserman. *The pascal visual object classes (voc) challenge*. International journal of computer vision, vol. 88, no. 2, pages 303–338, 2010.

[Everingham *et al.* 2015] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn et Andrew Zisserman. *The pascal visual object classes challenge : A retrospective*. International journal of computer vision, vol. 111, no. 1, pages 98–136, 2015.

[Eyben *et al.* 2013] Florian Eyben, Felix Weninger, Stefano Squartini et Björn Schuller. *Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies*. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 483–487. IEEE, 2013.

[Fan *et al.* 2018] Weikang Fan, Huiqin Jiang, Ling Ma, Jianbo Gao et Haojin Yang. *A modified faster R-CNN method to improve the performance of the pulmonary nodule detection*. In Tenth International Conference on Digital Image Processing (ICDIP 2018), volume 10806, pages 1469 – 1476. International Society for Optics and Photonics, SPIE, 2018.

[Farahani *et al.* 2015] Navid Farahani, Anil V Parwani et Liron Pantanowitz. *Whole slide imaging in pathology : advantages, limitations, and emerging perspectives*. Pathol Lab Med Int, vol. 7, pages 23–33, 2015.

[Ferreira *et al.* 2018] Carlos A Ferreira, Tânia Melo, Patrick Sousa, Maria Inês Meyer, Elham Shakibapour, Pedro Costa et Aurélio Campilho. *Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2*. In International Conference Image Analysis and Recognition, pages 763–770. Springer, 2018.

[Fischer *et al.* 2008] Andrew H Fischer, Kenneth A Jacobson, Jack Rose et Rolf Zeller. *Hematoxylin and eosin staining of tissue and cell sections*. Cold spring harbor protocols, vol. 6, 2008.

[Fishman *et al.* 2003] Joel E Fishman, Clara Milikowski, Rajeev Ramsinghani, M Victoria Velasquez et Galit Aviram. *US-guided core-needle biopsy of the breast : how many specimens are necessary ?* Radiology, vol. 226, no. 3, pages 779–782, 2003.

[Fok *et al.* 2018] Wilson Fok, Kevin Jamart, Jichao Zhao et Justin Fernandez. *Ensemble of Convolutional Neural Networks for Heart Segmentation*. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 282–291. Springer, 2018.

[Frierson Jr *et al.* 1995] Henry F Frierson Jr, Robert A Wolber, Kenneth W Berean, Douglas W Franquemont, Michael J Gaffey, James C Boyd et David C Wilbur. *Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma*. American journal of clinical pathology, vol. 103, no. 2, pages 195–198, 1995.

[Fukushima 1980] Kunihiko Fukushima. *Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*. Biological cybernetics, vol. 36, no. 4, pages 193–202, 1980.

[Gal *et al.* 2005] Rivka Gal, Lea Rath-Wolfson, Yevgenia Rosenblatt, Marisa Halpern, Ariel Schwartz et Rumelia Koren. *An improved technique for mitosis counting*. International journal of surgical pathology, vol. 13, no. 2, pages 161–165, 2005.

[Gan *et al.* 2015] Zhe Gan, Chunyuan Li, Ricardo Henao, David E Carlson et Lawrence Carin. *Deep temporal sigmoid belief networks for sequence modeling*. In Advances in Neural Information Processing Systems, pages 2467–2475, 2015.

[Gao *et al.* 2019] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang et Philip HS Torr. *Res2net : A new multi-scale backbone architecture*. IEEE transactions on pattern analysis and machine intelligence, 2019.

[Garcia-Gonzalo & Fernandez-Martinez 2012] Esperanza Garcia-Gonzalo et Juan Luis Fernandez-Martinez. *A brief historical review of particle swarm optimization (PSO).* Journal of Bioinformatics and Intelligent Control, vol. 1, no. 1, pages 3–16, 2012.

[Geessink *et al.* 2017] Oscar Geessink, Péter Bándi, Geert Litjens et Jeroen van der Laak. *Camelyon17 : Grand challenge on cancer metastasis detection and classification in lymph nodes*, 2017.

[Girshick *et al.* 2014] Ross Girshick, Jeff Donahue, Trevor Darrell et Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[Girshick 2015] Ross Girshick. *Fast r-cnn*. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.

[Goodfellow *et al.* 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio. *Generative adversarial nets*. In Advances in neural information processing systems, pages 2672–2680, 2014.

[Goodhew & Humphreys 2000] Peter J Goodhew et John Humphreys. Electron microscopy and analysis. CRC Press, 2000.

[Hamidinekoo *et al.* 2018] Azam Hamidinekoo, Erika Denton, Andrik Rampun, Kate Honnor et Reyer Zwiggelaar. *Deep learning in mammography and breast histology, an overview and future trends*. Medical image analysis, vol. 47, pages 45–67, 2018.

[Hammoudi *et al.* 2018] Karim Hammoudi, Mahmoud Melkemi, Fadi Dornaika, Halim Benhabiles, Feryal Windal et Oussama Taoufik. *A comparative study of 2 resolution-level LBP descriptors and compact versions for visual analysis*. In Advanced Multimedia and Ubiquitous Engineering, pages 221–227. Springer, 2018.

[Han *et al.* 2018] Yamin Han, Peng Zhang, Tao Zhuo, Wei Huang et Yanning Zhang. *Going deeper with two-stream ConvNets for action recognition in video surveillance*. Pattern Recognition Letters, vol. 107, pages 83–90, 2018.

[Hao *et al.* 1999] Jin-Kao Hao, Philippe Galinier et Michel Habib. *Métaheuristiques pour l'optimisation combinatoire et l'affectation sous contraintes*. Revue d'intelligence artificielle, vol. 13, no. 2, pages 283–324, 1999.

[He *et al.* 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[He *et al.* 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár et Ross Girshick. *Mask r-cnn*. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[Hinton *et al.* 2006] Geoffrey E Hinton, Simon Osindero et Yee-Whye Teh. *A fast learning algorithm for deep belief nets*. Neural computation, vol. 18, no. 7, pages 1527–1554, 2006.

[Hochreiter & Schmidhuber 1997] Sepp Hochreiter et Jürgen Schmidhuber. *Long short-term memory*. Neural computation, vol. 9, no. 8, pages 1735–1780, 1997.

[Howard *et al.* 2017] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto et Hartwig Adam. *MobileNets : Efficient Convolutional Neural Networks for Mobile Vision Applications*. CoRR. abs/1704.04861, 2017.

[Hu *et al.* 2015] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang et Hengchao Li. *Deep convolutional neural networks for hyperspectral image classification*. Journal of Sensors, 2015.

[Huang & Chung 2018] Yongxiang Huang et Albert Chi-shing Chung. *Improving high resolution histology image classification with deep spatial fusion network*. In Computational Pathology and Ophthalmic Medical Image Analysis, pages 19–26. Springer, 2018.

[Huang *et al.* 2014] Kai-Qi Huang, Wei-Qiang Ren et TN Tan. *A review on image object classification and detection*. Chinese Journal of Computers, vol. 37, no. 6, pages 1225–1240, 2014.

[Huang *et al.* 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten et Kilian Q Weinberger. *Densely connected convolutional networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.

[Hubel & Wiesel 1962] David H Hubel et Torsten N Wiesel. *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex*. The Journal of physiology, vol. 160, no. 1, pages 106–154, 1962.

[Iandola *et al.* 2016] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally et Kurt Keutzer. *SqueezeNet : AlexNet-level accuracy with 50x fewer parameters and <1MB model size*. CoRR. abs/1602.07360, 2016.

[Ilea & Whelan 2011] Dana E Ilea et Paul F Whelan. *Image segmentation based on the integration of colour–texture descriptors—A review*. Pattern Recognition, vol. 44, no. 10, pages 2479–2501, 2011.

[Ishii *et al.* 2013] Takaaki Ishii, Hiroki Komiyama, Takahiro Shinozaki, Yasuo Horiuchi et Shingo Kuroiwa. *Reverberant speech recognition based on denoising autoencoder.* In Interspeech, pages 3512–3516, 2013.

[Isola *et al.* 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou et Alexei A Efros. *Image-to-image translation with conditional adversarial networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1125–1134, 2017.

[Izmailov *et al.* 2018] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov et Andrew Gordon Wilson. *Averaging Weights Leads to Wider Optima and Better Generalization*. In Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018, pages 876–885. AUAI Press, 2018.

[Jain *et al.* 2014] Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W Taylor et Christoph Bregler. *Learning Human Pose Estimation Features with Convolutional Networks*. CoRR. abs/1312.7302, 2014.

[Janowczyk & Madabhushi 2016] Andrew Janowczyk et Anant Madabhushi. *Deep learning for digital pathology image analysis : A comprehensive tutorial with selected use cases*. Journal of pathology informatics, vol. 7, 2016.

[Ji *et al.* 2012] Shuiwang Ji, Wei Xu, Ming Yang et Kai Yu. *3D convolutional neural networks for human action recognition*. IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pages 221–231, 2012.

[Jimenez-del Toro *et al.* 2017] Oscar Jimenez-del Toro, Sebastian Otálora, Mats Andersson, Kristian Eurén, Martin Hedlund, Mikael Rousson, Henning Müller et Manfredo Atzori. *Analysis of histopathology images : From traditional machine learning to deep learning*. In Biomedical Texture Analysis, pages 281–314. Elsevier, 2017.

[Ju *et al.* 2018] Cheng Ju, Aurélien Bibaut et Mark van der Laan. *The relative performance of ensemble methods with deep convolutional neural networks for image classification*. Journal of Applied Statistics, vol. 45, no. 15, pages 2800–2818, 2018.

[Jung *et al.* 2018] Hwejin Jung, Bumsoo Kim, Inyeop Lee, Junhyun Lee et Jaewoo Kang. *Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method*. BMC medical imaging, vol. 18, no. 1, 2018.

[Kainz *et al.* 2015] Philipp Kainz, Michael Pfeiffer et Martin Urschler. *Semantic Segmentation of Colon Glands with Deep Convolutional Neural Networks and Total Variation Segmentation*. CoRR. abs/1511.06919, 2015.

[Kallenberg *et al.* 2016] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y Ng, Pengfei Diao, Christian Igel, Celine M Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer*et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring*. IEEE transactions on medical imaging, vol. 35, no. 5, pages 1322–1331, 2016.

[Kaman *et al.* 1984] EJ Kaman, AWM Smeulders, PW Verbeek, IT Young et JPA Baak. *Image processing for mitoses in sections of breast cancer : A feasibility study*. Cytometry : The Journal of the International Society for Analytical Cytology, vol. 5, no. 3, pages 244–249, 1984.

[Kampffmeyer *et al.* 2016] Michael Kampffmeyer, Arnt-Borre Salberg et Robert Jenssen. *Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 1–9, 2016.

[Kang *et al.* 2017] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang*et al. T-cnn : Tubelets with convolutional neural networks for object detection from videos*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pages 2896–2907, 2017.

[Karlik & Olgac 2011] Bekir Karlik et A Vehbi Olgac. *Performance analysis of various activation functions in generalized MLP architectures of neural networks*. International Journal of Artificial Intelligence and Expert Systems, vol. 1, no. 4, pages 111–122, 2011.

[Karpathy *et al.* 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar et Li Fei-Fei. *Large-scale video classification with convolutional neural networks*. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[Kather *et al.* 2016] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx et Frank Gerrit Zöllner. *Multi-class texture analysis in colorectal cancer histology*. Scientific reports, vol. 6, 2016.

[Kather *et al.* 2019] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber*et al. Predicting survival from colorectal cancer histology slides using deep learning : A retrospective multicenter study*. PLoS medicine, vol. 16, no. 1, 2019.

[Kausar *et al.* 2018] Tasleem Kausar, MingJiang Wang, Boqian Wu, Muhammad Idrees et Benish Kanwal. *Multi-Scale Deep Neural Network for Mitosis Detection in Histological Images*. In 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), volume 3, pages 47–51. IEEE, 2018.

[Ker *et al.* 2017] Justin Ker, Lipo Wang, Jai Rao et Tchoyoson Lim. *Deep learning applications in medical image analysis*. IEEE Access, vol. 6, pages 9375–9389, 2017.

[Keshari *et al.* 2018] Rohit Keshari, Mayank Vatsa, Richa Singh et Afzel Noore. *Learning structure and strength of CNN filters for small sample size training*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9349–9358, 2018.

[Khan *et al.* 2014] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor et Derek Magee. *A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution*. IEEE Transactions on Biomedical Engineering, vol. 61, no. 6, pages 1729–1738, 2014.

[Khan *et al.* 2019] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din et Joel JP C Rodrigues. *A novel deep learning based framework for the detection and classification of breast cancer using transfer learning*. Pattern Recognition Letters, vol. 125, pages 1–6, 2019.

[Kim 2014] Yoon Kim. *Convolutional Neural Networks for Sentence Classification*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1746–1751. ACL, 2014.

[Kingma & Ba 2015] Diederik P Kingma et Jimmy Ba. *Adam : A Method for Stochastic Optimization*. In Yoshua Bengio et Yann LeCun, editeurs, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[Kobayashi 2018] Hayato Kobayashi. *Frustratingly Easy Model Ensemble for Abstractive Summarization*. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4165–4176, 2018.

[Krähenbühl & Koltun 2011] Philipp Krähenbühl et Vladlen Koltun. *Efficient inference in fully connected crfs with gaussian edge potentials*. In Advances in neural information processing systems, pages 109–117, 2011.

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever et Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks*. In Advances in neural information processing systems, pages 1097–1105, 2012.

[Kügler *et al.* 2018] David Kügler, Andrei Stefanov et Anirban Mukhopadhyay. *i3PosNet : Instrument pose estimation from X-ray*. arXiv preprint arXiv :1802.09575, 2018.

[Kumar *et al.* 2017] Meghana Dinesh Kumar, Morteza Babaie, Shujin Zhu, Shivam Kalra et Hamid R Tizhoosh. *A comparative study of CNN, BoVW and LBP for classification of histopathological images*. In 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pages 1–7. IEEE, 2017.

[Lai *et al.* 2015] Siwei Lai, Liheng Xu, Kang Liu et Jun Zhao. *Recurrent convolutional neural networks for text classification*. In Twenty-ninth AAAI conference on artificial intelligence, 2015.

[Lakhani & Sundaram 2017] Paras Lakhani et Baskaran Sundaram. *Deep learning at chest radiography : automated classification of pulmonary tuberculosis by using convolutional neural networks*. Radiology, vol. 284, no. 2, pages 574–582, 2017.

[Lawrence *et al.* 1997] Steve Lawrence, C Lee Giles, Ah Chung Tsoi et Andrew D Back. *Face recognition : A convolutional neural-network approach*. IEEE transactions on neural networks, vol. 8, no. 1, pages 98–113, 1997.

[Lea *et al.* 2016] Colin Lea, Rene Vidal, Austin Reiter et Gregory D Hager. *Temporal convolutional networks : A unified approach to action segmentation*. In European Conference on Computer Vision, pages 47–54. Springer, 2016.

[LeCun *et al.* 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner*et al. Gradient-based learning applied to document recognition*. Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, 1998.

[Lee *et al.* 2015] Sue Han Lee, Chee Seng Chan, Paul Wilkin et Paolo Remagnino. *Deep-plant : Plant identification with convolutional neural networks*. In 2015 IEEE international conference on image processing (ICIP), pages 452–456. IEEE, 2015.

[Lee *et al.* 2019] HyunJae Lee, Hyo-Eun Kim et Hyeonseob Nam. *Srm : A style-based recalibration module for convolutional neural networks*. In Proceedings of the IEEE International Conference on Computer Vision, pages 1854–1862, 2019.

[Levi *et al.* 2015] Dan Levi, Noa Garnett, Ethan Fetaya et Israel Herzlyia. *StixelNet : A Deep Convolutional Network for Obstacle Detection and Road Segmentation.* In BMVC, 2015.

[Li *et al.* 2014a] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng et Mei Chen. *Medical image classification with convolutional neural network*. In 2014 13th international conference on control automation robotics & vision (ICARCV), pages 844–848. IEEE, 2014.

[Li *et al.* 2014b] Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen et Shuiwang Ji. *Deep learning based imaging data completion for improved brain disease diagnosis*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 305–312. Springer, 2014.

[Li *et al.* 2018a] Chao Li, Xinggang Wang, Wenyu Liu et Longin Jan Latecki. *DeepMitosis : Mitosis detection via deep detection, verification and segmentation networks*. Medical image analysis, vol. 45, pages 121–133, 2018.

[Li *et al.* 2018b] Yuguang Li, Ezgi Mercan, Stevan Knezevitch, Joann G. Elmore et Linda G. Shapiro. *Efficient and Accurate Mitosis Detection-A Lightweight RCNN Approach*. In ICPRAM, 2018.

[Li *et al.* 2019] Xiang Li, Xiaolin Hu et Jian Yang. *Spatial group-wise enhance : Improving semantic feature learning in convolutional networks*. arXiv preprint arXiv :1905.09646, 2019.

[Lin *et al.* 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C Lawrence Zitnick. *Microsoft coco : Common objects in context*. In European conference on computer vision, pages 740–755. Springer, 2014.

[Litjens *et al.* 2017] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken et Clara I Sánchez. *A survey on deep learning in medical image analysis.* Medical image analysis, vol. 42, pages 60–88, 2017.

[Liu & Wang 2018] Hongyi Liu et Lihui Wang. *Gesture recognition for human-robot collaboration : A review*. International Journal of Industrial Ergonomics, vol. 68, pages 355–367, 2018.

[Liu *et al.* 2015] Wei Liu, Andrew Rabinovich et Alexander C Berg. *Parsenet : Looking wider to see better*. arXiv preprint arXiv :1506.04579, 2015.

[Lo *et al.* 1995] S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien et Seong Ki Mun. *Artificial convolution neural network techniques and applications for lung nodule detection*. IEEE transactions on medical imaging, vol. 14, no. 4, pages 711–718, 1995.

[Long *et al.* 2015] Jonathan Long, Evan Shelhamer et Trevor Darrell. *Fully convolutional networks for semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[Long *et al.* 2017] Yang Long, Yiping Gong, Zhifeng Xiao et Qing Liu. *Accurate object localization in remote sensing images based on convolutional neural networks*. IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 5, pages 2486–2498, 2017.

[Luc *et al.* 2016] Pauline Luc, Camille Couprie, Soumith Chintala et Jakob Verbeek. *Semantic segmentation using adversarial networks*. arXiv preprint arXiv :1611.08408, 2016.

[Luvizon *et al.* 2018] Diogo C Luvizon, David Picard et Hedi Tabia. *2d/3d pose estimation and action recognition using multitask deep learning*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5137–5146, 2018.

[Ly *et al.* 2017] Tiffany Ly, Rituparna Sarkar, Kevin Skadron et Scott T Acton. *Classifying images in a histopathological dataset using the cumulative distribution transform on an automata architecture*. In 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pages 730–734. IEEE, 2017.

[Ma *et al.* 2018] Minglin Ma, Yinghuan Shi, Wenbin Li, Yang Gao et Jun Xu. *A Novel Two-Stage Deep Method for Mitosis Detection in Breast Cancer Histology Images*. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 3892–3897. IEEE, 2018.

[Macenko *et al.* 2009] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt et Nancy E Thomas. *A method for normalizing histology slides for quantitative analysis*. In 2009 IEEE International Symposium on Biomedical Imaging : From Nano to Macro, pages 1107–1110. IEEE, 2009.

[Mahajan *et al.* 2018] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe et Laurens van der Maaten. *Exploring the limits of weakly supervised pretraining*. In Proceedings of the European Conference on Computer Vision (ECCV), pages 181–196, 2018.

[Makki 2015] Jaafar Makki. *Diversity of breast carcinoma : histological subtypes and clinical relevance*. Clinical Medicine Insights : Pathology, vol. 8, pages 23–31, 2015.

[Malik *et al.* 2019] Junaid Malik, Serkan Kiranyaz, Suchitra Kunhoth, Turker Ince, Somaya Al-Maadeed, Ridha Hamila et Moncef Gabbouj. *Colorectal cancer diagnosis from histology images : A comparative study*. arXiv preprint arXiv :1903.11210, 2019.

[Malon & Cosatto 2013] Christopher D Malon et Eric Cosatto. *Classification of mitotic figures with convolutional neural networks and seeded blob features*. Journal of pathology informatics, vol. 4, 2013.

[Malon *et al.* 2008] Christopher Malon, Matthew Miller, Harold Christopher Burger, Eric Cosatto et Hans Peter Graf. *Identifying histological elements with convolutional neural networks*. In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, pages 450–456, 2008.

[McCulloch & Pitts 1943] Warren S McCulloch et Walter Pitts. *A logical calculus of the ideas immanent in nervous activity*. The bulletin of mathematical biophysics, vol. 5, no. 4, pages 115–133, 1943.

[Mehra *et al.* 2018] Rajesh Mehra*et al. Breast cancer histology images classification : Training from scratch or transfer learning ?* ICT Express, vol. 4, no. 4, pages 247–254, 2018.

[Meng *et al.* 2010] Tao Meng, Lin Lin, Mei-Ling Shyu et Shu-Ching Chen. *Histology image classification using supervised classification and multimodal fusion*. In 2010 IEEE international symposium on multimedia, pages 145–152. IEEE, 2010.

[Mercadier *et al.* 2019] Deniz Sayin Mercadier, Beril Besbinar et Pascal Frossard. *Automatic Segmentation of Nuclei in Histopathology Images Using Encoding-decoding Convolutional Neural Networks*. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1020–1024. IEEE, 2019.

[Mescher 2013] Anthony L Mescher. Junqueira's basic histology : text and atlas. LANGE Series, McGraw-Hill Medical, 2013.

[Milletari *et al.* 2016] Fausto Milletari, Nassir Navab et Seyed-Ahmad Ahmadi. *V-net : Fully convolutional neural networks for volumetric medical image segmentation*. In 2016 Fourth International Conference on 3D Vision (3DV), pages 565–571. IEEE, 2016.

[Minsky & Papert 2017] Marvin Minsky et Seymour A Papert. Perceptrons : An introduction to computational geometry. MIT press, 2017.

[Mosca & Magoulas 2017] Alan Mosca et George D Magoulas. *Deep incremental boosting*. arXiv preprint arXiv :1708.03704, 2017.

[Mottaghi *et al.* 2014] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun et Alan Yuille. *The role of context for object detection and semantic segmentation in the wild*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 891–898, 2014.

[Muhammad *et al.* 2018] Khan Muhammad, Jamil Ahmad et Sung Wook Baik. *Early fire detection using convolutional neural networks during surveillance for effective disaster management*. Neurocomputing, vol. 288, pages 30–42, 2018.

[Nanni *et al.* 2018] Loris Nanni, Stefano Ghidoni et Sheryl Brahnam. *Ensemble of convolutional neural networks for bioimage classification*. Applied Computing and Informatics, 2018.

[Nanni *et al.* 2019a] Loris Nanni, Sheryl Brahnam, Stefano Ghidoni et Gianluca Maguolo. *General Purpose (GenP) Bioimage Ensemble of Handcrafted and Learned Features with Data Augmentation*. arXiv preprint arXiv :1904.08084, 2019.

[Nanni *et al.* 2019b] Loris Nanni, Sheryl Brahnam et Gianluca Maguolo. *Data Augmentation for Building an Ensemble of Convolutional Neural Networks*. In Innovation in Medicine and Healthcare Systems, and Multimedia, pages 61–69. Springer, 2019.

[Nava *et al.* 2016] Rodrigo Nava, German González, Jan Kybic et Boris Escalante-Ramírez. *Characterization of hematologic malignancies based on discrete orthogonal moments*. In 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6. IEEE, 2016.

[Nesterov 1983] Y Nesterov. *A method of solving a convex programming problem with convergence rate*. In Soviet Math. Dokl, volume 27, pages 543–547, 1983.

[Nogueira *et al.* 2016] Rodrigo Frassetto Nogueira, Roberto de Alencar Lotufo et Rubens Campos Machado. *Fingerprint liveness detection using convolutional neural networks*. IEEE transactions on information forensics and security, vol. 11, no. 6, pages 1206–1213, 2016.

[Orchid & Puthanpurayil 2016] Navya Narayanan Orchid et Sathi Puthanpurayil. *Factors affecting the assessment of mitotic count in histopathological sections of tumors : a study of interobserver and intraobserver variability*. International Journal of Research in Medical Sciences, vol. 4, no. 3, pages 762–765, 2016.

[Orlov *et al.* 2010] Nikita V Orlov, Wayne W Chen, David Mark Eckley, Tomasz J Macura, Lior Shamir, Elaine S Jaffe et Ilya G Goldberg. *Automatic classification of lymphoma images with transform-based global features*. IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pages 1003–1013, 2010.

[Oyama *et al.* 2004] Tetsunari Oyama, Yukio Koibuchi et Grace McKee. *Core needle biopsy (CNB) as a diagnostic method for breast lesions : comparison with fine needle aspiration cytology (FNA)*. Breast Cancer, vol. 11, no. 4, pages 339–342, 2004.

[Paeng *et al.* 2017] Kyunghyun Paeng, Sangheum Hwang, Sunggyun Park et Minsoo Kim. *A unified framework for tumor proliferation score prediction in breast histopathology*. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 231–239. Springer, 2017.

[Pantanowitz *et al.* 2011] Liron Pantanowitz, Paul N Valenstein, Andrew J Evans, Keith J Kaplan, John D Pfeifer, David C Wilbur, Laura C Collins et Terence J Colgan. *Review of the current state of whole slide imaging in pathology*. Journal of pathology informatics, vol. 2, 2011.

[Papandreou *et al.*] G Papandreou, L-Ch Chen, K Murphy et AL Yuille. *Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. arXiv, 2015*. arXiv preprint arXiv :1502.02734.

[Parkhi *et al.* 2015] Omkar M Parkhi, Andrea Vedaldi et Andrew Zisserman. *Deep face recognition*. Proceedings of the British Machine Vision, 2015.

[Peng *et al.* 2018] Binbin Peng, Lin Chen, Mingsheng Shang et Jianjun Xu. *Fully Convolutional Neural Networks for Tissue Histopathology Image Classification and Segmentation*. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 1403–1407. IEEE, 2018.

[Pereira *et al.* 2016] Sérgio Pereira, Adriano Pinto, Victor Alves et Carlos A Silva. *Brain tumor segmentation using convolutional neural networks in MRI images*. IEEE transactions on medical imaging, vol. 35, no. 5, pages 1240–1251, 2016.

[Pezzotti *et al.* 2017] Nicola Pezzotti, Thomas Höllt, Jan Van Gemert, Boudewijn PF Lelieveldt, Elmar Eisemann et Anna Vilanova. *Deepeyes : Progressive visual analytics for designing deep neural networks*. IEEE transactions on visualization and computer graphics, vol. 24, no. 1, pages 98–108, 2017.

[Pham 2017] Tuan D Pham. *Scaling of texture in training autoencoders for classification of histological images of colorectal cancer*. In International Symposium on Neural Networks, pages 524–532. Springer, 2017.

[Pineda 1987] Fernando J Pineda. *Generalization of back-propagation to recurrent neural networks*. Physical review letters, vol. 59, no. 19, pages 2229–2232, 1987.

[Popa *et al.* 2017] Alin-Ionut Popa, Mihai Zanfir et Cristian Sminchisescu. *Deep multitask architecture for integrated 2d and 3d human sensing*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6289–6298, 2017.

[Qian 1999] Ning Qian. *On the momentum term in gradient descent learning algorithms*. Neural networks, vol. 12, no. 1, pages 145–151, 1999.

[Rączkowski *et al.* 2019] ukasz Rączkowski, Marcin Możejko, Joanna Zambonelli et Ewa Szczurek. *ARA : accurate, reliable and active histopathological image classification framework with Bayesian deep learning*. Scientific reports, vol. 9, no. 1, pages 1–12, 2019.

[Radford *et al.* 2015] Alec Radford, Luke Metz et Soumith Chintala. *Unsupervised representation learning with deep convolutional generative adversarial networks*. arXiv preprint arXiv :1511.06434, 2015.

[Rajkomar *et al.* 2017] Alvin Rajkomar, Sneha Lingam, Andrew G Taylor, Michael Blum et John Mongan. *High-throughput classification of radiographs using deep convolutional neural networks*. Journal of digital imaging, vol. 30, no. 1, pages 95–101, 2017.

[Rajpurkar *et al.* 2017] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya*et al. Chexnet : Radiologist-level pneumonia detection on chest x-rays with deep learning*. arXiv preprint arXiv :1711.05225, 2017.

[Ramachandran *et al.* 2017] Prajit Ramachandran, Barret Zoph et Quoc V Le. *Searching for activation functions*. arXiv preprint arXiv :1710.05941, 2017.

[Rao 2018] Siddhant Rao. *Mitos-rcnn : A novel approach to mitotic figure detection in breast cancer histopathology images using region based convolutional neural networks*. arXiv preprint arXiv :1807.01788, 2018.

[Redmon & Farhadi 2017] Joseph Redmon et Ali Farhadi. *YOLO9000 : better, faster, stronger*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7263–7271, 2017.

[Redmon & Farhadi 2018] Joseph Redmon et Ali Farhadi. *Yolov3 : An incremental improvement*. arXiv preprint arXiv :1804.02767, 2018.

[Redmon *et al.* 2016] Joseph Redmon, Santosh Divvala, Ross Girshick et Ali Farhadi. *You only look once : Unified, real-time object detection*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[Reed *et al.* 2014] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan et Andrew Rabinovich. *Training deep neural networks on noisy labels with bootstrapping*. arXiv preprint arXiv :1412.6596, 2014.

[Reinhard *et al.* 2001] Erik Reinhard, Michael Adhikhmin, Bruce Gooch et Peter Shirley. *Color transfer between images*. IEEE Computer graphics and applications, vol. 21, no. 5, pages 34–41, 2001.

[Ren *et al.* 2015] Shaoqing Ren, Kaiming He, Ross Girshick et Jian Sun. *Faster r-cnn : Towards real-time object detection with region proposal networks*. In Advances in neural information processing systems, pages 91–99, 2015.

[Ribli *et al.* 2018] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner et István Csabai. *Detecting and classifying lesions in mammograms with deep learning*. Scientific reports, vol. 8, no. 1, pages 1–7, 2018.

[Riedmiller & Braun 1993] Martin Riedmiller et Heinrich Braun. *A direct adaptive method for faster backpropagation learning : The RPROP algorithm*. In Proceedings of the IEEE international conference on neural networks, volume 1993, pages 586–591. San Francisco, 1993.

[Robbins & Monro 1951] Herbert Robbins et Sutton Monro. *A stochastic approximation method*. The annals of mathematical statistics, pages 400–407, 1951.

[Ronneberger *et al.* 2015] Olaf Ronneberger, Philipp Fischer et Thomas Brox. *U-net : Convolutional networks for biomedical image segmentation*. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[Rosenblatt 1957] Frank Rosenblatt. The perceptron, a perceiving and recognizing automaton project para. Cornell Aeronautical Laboratory, 1957.

[Roth *et al.* 2018] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa et Kensaku Mori. *An application of cascaded 3D fully convolutional networks for medical image segmentation*. Computerized Medical Imaging and Graphics, vol. 66, pages 90–99, 2018.

[Roux *et al.* 2013] Ludovic Roux, Daniel Racoceanu, Nicolas Loménie, Maria Kulikova, Humayun Irshad, Jacques Klossa, Frédérique Capron, Catherine Genestie, Gilles Le Naour et Metin N Gurcan. *Mitosis detection in breast cancer histological images An ICPR 2012 contest*. Journal of pathology informatics, vol. 4, 2013.

[Roux *et al.* 2014] Ludovic Roux, Daniel Racoceanu, Frédérique Capron, Jessica Calvo, Elham Attieh, Gilles Le Naour et Anne Gloaguen. *Mitos & atypia*. Image Pervasive Access Lab (IPAL), Agency Sci., Technol. & Res. Inst. Infocom Res., Singapore, Tech. Rep, vol. 1, pages 1–8, 2014.

[Rumelhart *et al.* 1986] David E Rumelhart, Geoffrey E Hinton et Ronald J Williams. *Learning representations by back-propagating errors*. nature, vol. 323, pages 533–536, 1986.

[Sacco 2005] Maddalena Sacco. *Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images*. PhD thesis, Seconda Università degli Studi di Napoli, 2005.

[Saha *et al.* 2018] Monjoy Saha, Chandan Chakraborty et Daniel Racoceanu. *Efficient deep learning model for mitosis detection using breast histopathology images*. Computerized Medical Imaging and Graphics, vol. 64, pages 29–40, 2018.

[Samala *et al.* 2016] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Jun Wei et Kenny Cha. *Mass detection in digital breast tomosynthesis : Deep convolutional neural network with transfer learning from mammography*. Medical physics, vol. 43, no. 12, pages 6654–6666, 2016.

[Sánchez & Perronnin 2011] Jorge Sánchez et Florent Perronnin. *High-dimensional signature compression for large-scale image classification*. In CVPR 2011, pages 1665–1672. IEEE, 2011.

[Sandler *et al.* 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov et Liang-Chieh Chen. *Mobilenetv2 : Inverted residuals and linear bottlenecks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018.

[Sang *et al.* 2018a] Dinh Viet Sang, Dang Manh Cuong et Le Tran Bao Cuong. *An Effective Ensemble Deep Learning Framework for Malware Detection*. In Proceedings of the Ninth International Symposium on Information and Communication Technology, pages 192–199, 2018.

[Sang *et al.* 2018b] Dinh Viet Sang, Pham Thai Ha*et al. Discriminative deep feature learning for facial emotion recognition*. In 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR), pages 1–6. IEEE, 2018.

[Schroff *et al.* 2015] Florian Schroff, Dmitry Kalenichenko et James Philbin. *Facenet : A unified embedding for face recognition and clustering*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[Sennrich *et al.* 2016] Rico Sennrich, Barry Haddow et Alexandra Birch. *Edinburgh neural machine translation systems for WMT 16*. arXiv preprint arXiv :1606.02891, 2016.

[Shaban *et al.* 2019] M Tarek Shaban, Christoph Baur, Nassir Navab et Shadi Albarqouni. *Staingan : Stain style transfer for digital histological images*. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pages 953–956. IEEE, 2019.

[Shah *et al.* 2017] Manan Shah, Dayong Wang, Christopher Rubadue, David Suster et Andrew Beck. *Deep learning assessment of tumor proliferation in breast cancer histological images*. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 600–603. IEEE, 2017.

[Shamir *et al.* 2008] Lior Shamir, Nikita Orlov, David Mark Eckley, Tomasz J Macura et Ilya G Goldberg. *IICBU 2008 : a proposed benchmark suite for biological image analysis*. Medical & biological engineering & computing, vol. 46, no. 9, pages 943–947, 2008.

[Shen *et al.* 2015] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang et Jie Tian. *Multi-scale convolutional neural networks for lung nodule classification*. In International Conference on Information Processing in Medical Imaging, pages 588–599. Springer, 2015.

[Shin *et al.* 2017] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura et Ronald M Summers. *Three Aspects on Using Convolutional Neural Networks for Computer-Aided Detection in Medical Imaging*. In Deep Learning and Convolutional Neural Networks for Medical Image Computing, pages 113–136. Springer, 2017.

[Sifre 2014] Laurent Sifre. *Rigid-motion scattering for image classification*. Ph. D. thesis, 2014.

[Simonyan & Zisserman 2014a] Karen Simonyan et Andrew Zisserman. *Two-stream convolutional networks for action recognition in videos*. In Advances in neural information processing systems, pages 568–576, 2014.

[Simonyan & Zisserman 2014b] Karen Simonyan et Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv :1409.1556, 2014.

[Sinha *et al.* 2018] Rajat Kumar Sinha, Ruchi Pandey et Rohan Pattnaik. *Deep learning for computer vision tasks : a review*. arXiv preprint arXiv :1804.03928, 2018.

[Sirinukunwattana *et al.* 2016] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree et Nasir M Rajpoot. *Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images*. IEEE transactions on medical imaging, vol. 35, no. 5, pages 1196–1206, 2016.

[Skalic *et al.* 2017] Miha Skalic, Marcin Pekalski et Xingguo E Pan. *Deep learning methods for efficient large scale video labeling*. arXiv preprint arXiv :1706.04572, 2017.

[Song *et al.* 2016] Yang Song, Weidong Cai, Heng Huang, Dagan Feng, Yue Wang et Mei Chen. *Bioimage classification with subcategory discriminant transform of high dimensional visual descriptors*. BMC bioinformatics, vol. 17, no. 1, pages 465–479, 2016.

[Song *et al.* 2017a] Yang Song, Hang Chang, Heng Huang et Weidong Cai. *Supervised intra-embedding of fisher vectors for histopathology image classification*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 99–106. Springer, 2017.

[Song *et al.* 2017b] Yang Song, Qing Li, Heng Huang, Dagan Feng, Mei Chen et Weidong Cai. *Low dimensional representation of fisher vectors for microscopy image classification*. IEEE transactions on medical imaging, vol. 36, no. 8, pages 1636–1649, 2017.

[Spanhol *et al.* 2015] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean et Laurent Heutte. *A dataset for breast cancer histopathological image classification*. IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pages 1455–1462, 2015.

[Spanhol *et al.* 2016] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean et Laurent Heutte. *Breast cancer histopathological image classification using convolutional neural networks*. In 2016 international joint conference on neural networks (IJCNN), pages 2560–2567. IEEE, 2016.

[Srivastava *et al.* 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever et Ruslan Salakhutdinov. *Dropout : a simple way to prevent neural networks from overfitting*. The journal of machine learning research, vol. 15, no. 1, pages 1929–1958, 2014.

[Srivastava *et al.* 2015] Rupesh Kumar Srivastava, Klaus Greff et Jürgen Schmidhuber. *Highway networks*. arXiv preprint arXiv :1505.00387, 2015.

[Su *et al.* 2017] Yu-Ting Su, Yao Lu, Mei Chen et An-An Liu. *Spatiotemporal joint mitosis detection using CNN-LSTM network in time-lapse phase contrast microscopy images*. IEEE Access, vol. 5, pages 18033–18041, 2017.

[Sun *et al.* 2015] Lin Sun, Kui Jia, Dit-Yan Yeung et Bertram E Shi. *Human action recognition using factorized spatio-temporal convolutional networks*. In Proceedings of the IEEE international conference on computer vision, pages 4597–4605, 2015.

[Szegedy *et al.* 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke et Andrew Rabinovich. *Going deeper with convolutions*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[Szegedy *et al.* 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens et Zbigniew Wojna. *Rethinking the inception architecture for computer vision*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[Szegedy *et al.* 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke et Alexander A Alemi. *Inception-v4, inception-resnet and the impact of residual connections on learning*. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[Taigman *et al.* 2014] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato et Lior Wolf. *Deepface : Closing the gap to human-level performance in face verification*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1701–1708, 2014.

[Tajbakhsh *et al.* 2016] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway et Jianming Liang. *Convolutional neural networks for medical image analysis : Full training or fine tuning ?* IEEE transactions on medical imaging, vol. 35, no. 5, pages 1299–1312, 2016.

[Tan & Eswaran 2011] Chun Chet Tan et Chikkannan Eswaran. *Using autoencoders for mammogram compression*. Journal of medical systems, vol. 35, no. 1, pages 49–58, 2011.

[Tang *et al.* 2015] Duyu Tang, Bing Qin et Ting Liu. *Document modeling with gated recurrent neural network for sentiment classification*. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 1422–1432, 2015.

[Tellez *et al.* 2018] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer*et al.* *Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks*. IEEE transactions on medical imaging, vol. 37, no. 9, pages 2126–2136, 2018.

[Ten Kate *et al.* 1993] TK Ten Kate, JAM Belien, AWM Smeulders et JPA Baak. *Method for counting mitoses by image processing in Feulgen stained breast cancer sections*. Cytometry : The Journal of the International Society for Analytical Cytology, vol. 14, no. 3, pages 241–250, 1993.

[Thomas *et al.* 2015] Anish Thomas, Stephen V Liu, Deepa S Subramaniam et Giuseppe Giaccone. *Refining the treatment of NSCLC according to histological and molecular subtypes*. Nature reviews Clinical oncology, vol. 12, no. 9, pages 511–526, 2015.

[Tian *et al.* 2018] Yuchi Tian, Kexin Pei, Suman Jana et Baishakhi Ray. *Deeptest : Automated testing of deep-neural-network-driven autonomous cars*. In Proceedings of the 40th international conference on software engineering, pages 303–314, 2018.

[Tieleman & Hinton 2012] Tijmen Tieleman et Geoffrey Hinton. *Lecture 6.5-rmsprop : Divide the gradient by a running average of its recent magnitude*. COURSERA : Neural networks for machine learning, vol. 4, no. 2, pages 26–31, 2012.

[Tolba *et al.* 2006] AS Tolba, AH El-Baz et AA El-Harby. *Face recognition : A literature review*. International Journal of Signal Processing, vol. 2, no. 2, pages 88–103, 2006.

[Tompson *et al.* 2014] Jonathan J Tompson, Arjun Jain, Yann LeCun et Christoph Bregler. *Joint training of a convolutional network and a graphical model for human pose estimation*. In Advances in neural information processing systems, pages 1799–1807, 2014.

[Torre *et al.* 2015] Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent et Ahmedin Jemal. *Global cancer statistics, 2012*. CA : a cancer journal for clinicians, vol. 65, no. 2, pages 87–108, 2015.

[Toshev & Szegedy 2014] Alexander Toshev et Christian Szegedy. *Deeppose : Human pose estimation via deep neural networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1653–1660, 2014.

[Tosta *et al.* 2017] Thaína A Azevedo Tosta, Leandro A Neves et Marcelo Z do Nascimento. *Segmentation methods of H&E-stained histological*

images of lymphoma : a review. Informatics in Medicine Unlocked, vol. 9, pages 35–43, 2017.

[Tran *et al.* 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani et Manohar Paluri. *Learning spatiotemporal features with 3d convolutional networks*. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015.

[Turan *et al.* 2018] Mehmet Turan, Yasin Almalioglu, Helder Araujo, Ender Konukoglu et Metin Sitti. *Deep endovo : A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots*. Neurocomputing, vol. 275, pages 1861–1870, 2018.

[Uijlings *et al.* 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers et Arnold WM Smeulders. *Selective search for object recognition*. International journal of computer vision, vol. 104, no. 2, pages 154–171, 2013.

[Vahadane *et al.* 2016] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito et Nassir Navab. *Structure-preserving color normalization and sparse stain separation for histological images*. IEEE transactions on medical imaging, vol. 35, no. 8, pages 1962–1971, 2016.

[Vaswani *et al.* 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser et Illia Polosukhin. *Attention is all you need*. In Advances in neural information processing systems, pages 5998–6008, 2017.

[Veeling *et al.* 2018] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen et Max Welling. *Rotation equivariant CNNs for digital pathology*. In International Conference on Medical image computing and computer-assisted intervention, pages 210–218. Springer, 2018.

[Vesal *et al.* 2018] Sulaiman Vesal, Nishant Ravikumar, AmirAbbas Davari, Stephan Ellmann et Andreas Maier. *Classification of breast cancer histology images using transfer learning*. In International conference image analysis and recognition, pages 812–819. Springer, 2018.

[Veta *et al.* 2014] Mitko Veta, Josien PW Pluim, Paul J Van Diest et Max A Viergever. *Breast cancer histopathology image analysis : A review*. IEEE Transactions on Biomedical Engineering, vol. 61, no. 5, pages 1400–1411, 2014.

[Veta *et al.* 2015] Mitko Veta, Paul J Van Diest, Stefan M Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, Anders BL Larsen, Jacob S Vestergaard, Anders B Dahl*et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images*. Medical image analysis, vol. 20, no. 1, pages 237–248, 2015.

[Veta *et al.* 2016] Mitko Veta, Paul J Van Diest, Mehdi Jiwa, Shaimaa Al-Janabi et Josien PW Pluim. *Mitosis counting in breast cancer : Object-level interobserver agreement and comparison to an automatic method*. PloS one, vol. 11, no. 8, 2016.

[Veta *et al.* 2019] Mitko Veta, Yujing J Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A Shah, Dayong Wang, Mikael Rousson*et al. Predicting breast tumor proliferation from whole-slide images : the TUPAC16 challenge*. Medical image analysis, vol. 54, pages 111–121, 2019.

[Vinyals *et al.* 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio et Dumitru Erhan. *Show and tell : A neural image caption generator*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164, 2015.

[Vo *et al.* 2019] Duc My Vo, Ngoc-Quang Nguyen et Sang-Woong Lee. *Classification of breast cancer histology images using incremental boosting convolution networks*. Information Sciences, vol. 482, pages 123–138, 2019.

[Voulodimos *et al.* 2018] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis et Eftychios Protopapadakis. *Deep learning for computer vision : A brief review*. Computational intelligence and neuroscience, 2018.

[Wahab *et al.* 2017] Noorul Wahab, Asifullah Khan et Yeon Soo Lee. *Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection*. Computers in biology and medicine, vol. 85, pages 86–97, 2017.

[Wang *et al.* 2014] Haibo Wang, Angel Cruz Roa, Ajay N Basavanhally, Hannah L Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez et Anant Madabhushi. *Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features*. Journal of Medical Imaging, vol. 1, no. 3, pages 1 – 8, 2014.

[Wang *et al.* 2017a] Chaofeng Wang, Jun Shi, Qi Zhang et Shihui Ying. *Histopathological image classification with bilinear convolutional neural networks*. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 4050–4053. IEEE, 2017.

[Wang *et al.* 2017b] Xiang Wang, Huimin Ma, Xiaozhi Chen et Shaodi You. *Edge preserving and multi-scale contextual neural network for salient object detection*. IEEE Transactions on Image Processing, vol. 27, no. 1, pages 121–134, 2017.

[Wang *et al.* 2018] Kang Wang, Rui Zhao et Qiang Ji. *Human computer interaction with head pose, eye gaze and body gestures*. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 789–789. IEEE, 2018.

[Ward *et al.* 2015] Elizabeth M Ward, Carol E DeSantis, Chun Chieh Lin, Joan L Kramer, Ahmedin Jemal, Betsy Kohler, Otis W Brawley et Ted Gansler. *Cancer statistics : breast cancer in situ*. CA : a cancer journal for clinicians, vol. 65, no. 6, pages 481–495, 2015.

[Weinland *et al.* 2011] Daniel Weinland, Remi Ronfard et Edmond Boyer. *A survey of vision-based methods for action representation, segmentation and recognition*. Computer vision and image understanding, vol. 115, no. 2, pages 224–241, 2011.

[Weninger *et al.* 2015] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey et Björn Schuller. *Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR*. In International Conference on Latent Variable Analysis and Signal Separation, pages 91–99. Springer, 2015.

[Willems *et al.* 2012] Stefan Martin Willems, CHM Van Deurzen et PJ Van Diest. *Diagnosis of breast lesions : fine-needle aspiration cytology or core needle biopsy ? A review*. Journal of clinical pathology, vol. 65, no. 4, pages 287–292, 2012.

[Wollmann & Rohr 2017a] Thomas Wollmann et Karl Rohr. *Automatic grading of breast cancer whole-slide histopathology images*. In Bildverarbeitung für die Medizin 2017, pages 249–253. Springer, 2017.

[Wollmann & Rohr 2017b] Thomas Wollmann et Karl Rohr. *Deep residual Hough voting for mitotic cell detection in histopathology images*. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pages 341–344. IEEE, 2017.

[Wolterink *et al.* 2016] Jelmer M Wolterink, Tim Leiner, Max A Viergever et Ivana Išgum. *Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease*. In Reconstruction, segmentation, and analysis of medical images, pages 95–102. Springer, 2016.

[Wu & Chen 2015] Meiyin Wu et Li Chen. *Image recognition based on deep learning*. In 2015 Chinese Automation Congress (CAC), pages 542–546. IEEE, 2015.

[Wu *et al.* 2017] Boqian Wu, Tasleem Kausar, Qiao Xiao, Mingjiang Wang, Wenfeng Wang, Binwen Fan et Dandan Sun. *FF-CNN : an efficient deep neural network for mitosis detection in breast cancer histological images*. In Annual Conference on Medical Image Understanding and Analysis, pages 249–260. Springer, 2017.

[Wu *et al.* 2018] Xiang Wu, Ran He, Zhenan Sun et Tieniu Tan. *A light cnn for deep face representation with noisy labels*. IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pages 2884–2896, 2018.

[Xie *et al.* 2016] Yuanpu Xie, Zizhao Zhang, Manish Sapkota et Lin Yang. *Spatial clockwork recurrent neural network for muscle perimysium segmentation*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 185–193. Springer, 2016.

[Xu *et al.* 2015] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang et Anant Madabhushi. *Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images*. IEEE transactions on medical imaging, vol. 35, no. 1, pages 119–130, 2015.

[Xu *et al.* 2016] Jun Xu, Xiaofei Luo, Guanhao Wang, Hannah Gilmore et Anant Madabhushi. *A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images*. Neurocomputing, vol. 191, pages 214–223, 2016.

[Xu *et al.* 2017] Jun Xu, Chao Zhou, Bing Lang et Qingshan Liu. *Deep learning for histopathological image analysis : Towards computerized diagnosis on cancers*. In Deep Learning and Convolutional Neural Networks for Medical Image Computing, pages 73–95. Springer, 2017.

[Yang & Deb 2009] Xin-She Yang et Suash Deb. *Cuckoo search via Lévy flights*. In 2009 World congress on nature & biologically inspired computing (NaBIC), pages 210–214. IEEE, 2009.

[Yang *et al.* 2015] Dong Yang, Shaoting Zhang, Zhennan Yan, Chaowei Tan, Kang Li et Dimitris Metaxas. *Automated anatomical landmark detection ondistal femur surface using convolutional neural network*. In 2015 IEEE 12th international symposium on biomedical imaging (ISBI), pages 17–21. IEEE, 2015.

[Yang *et al.* 2016] Wei Yang, Wanli Ouyang, Hongsheng Li et Xiaogang Wang. *End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3073–3082, 2016.

[Yang 2009] Xin-She Yang. *Firefly algorithms for multimodal optimization*. In International symposium on stochastic algorithms, pages 169–178. Springer, 2009.

[Yang 2010] Xin-She Yang. *A new metaheuristic bat-inspired algorithm*. In Nature inspired cooperative strategies for optimization (NICSO 2010), pages 65–74. Springer, 2010.

[Yao *et al.* 2017] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu et Tao Mei. *Boosting image captioning with attributes*. In Proceedings of the IEEE International Conference on Computer Vision, pages 4894–4902, 2017.

[Yao *et al.* 2019] Guangle Yao, Tao Lei et Jiandan Zhong. *A review of Convolutional-Neural-Network-based action recognition*. Pattern Recognition Letters, vol. 118, pages 14–22, 2019.

[Ying *et al.* 2018] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton et Jure Leskovec. *Graph convolutional neural networks for web-scale recommender systems*. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 974–983. ACM, 2018.

[Yosinski *et al.* 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio et Hod Lipson. *How transferable are features in deep neural networks ?* In Advances in neural information processing systems, pages 3320–3328, 2014.

[Yun *et al.* 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe et Youngjoon Yoo. *Cutmix : Regularization strategy to train strong classifiers with localizable features*. In Proceedings of the IEEE International Conference on Computer Vision, pages 6023–6032, 2019.

[Zabalza *et al.* 2016] Jaime Zabalza, Jinchang Ren, Jiangbin Zheng, Huimin Zhao, Chunmei Qing, Zhijing Yang, Peijun Du et Stephen Marshall. *Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging*. Neurocomputing, vol. 185, pages 1–10, 2016.

[Zagoruyko & Komodakis 2016] Sergey Zagoruyko et Nikos Komodakis. *Wide residual networks*. arXiv preprint arXiv :1605.07146, 2016.

[Zanjani *et al.*] Farhad G Zanjani, Svitlana Zinger, Babak E Bejnordi, Jeroen AWM van der Laak*et al. Histopathology stain-color normalization using deep generative models*.

[Zeiler & Fergus 2014] Matthew D Zeiler et Rob Fergus. *Visualizing and understanding convolutional networks*. In European conference on computer vision, pages 818–833. Springer, 2014.

[Zeiler *et al.* 2011] Matthew D Zeiler, Graham W Taylor et Rob Fergus. *Adaptive deconvolutional networks for mid and high level feature learning*. In 2011 International Conference on Computer Vision, pages 2018–2025. IEEE, 2011.

[Zeiler 2012] Matthew D Zeiler. *ADADELTA : an adaptive learning rate method*. arXiv preprint arXiv :1212.5701, 2012.

[Zerhouni *et al.* 2017] Erwan Zerhouni, Dávid Lányi, Matheus Viana et Maria Gabrani. *Wide residual networks for mitosis detection*. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pages 924–928. IEEE, 2017.

[Zhang *et al.* 2013] Xin Zhang, Yee-Hong Yang, Zhiguang Han, Hui Wang et Chao Gao. *Object class detection : A survey*. ACM Computing Surveys (CSUR), vol. 46, no. 1, pages 1–53, 2013.

[Zhang *et al.* 2016] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht et Oriol Vinyals. *Understanding deep learning requires rethinking generalization*. arXiv preprint arXiv :1611.03530, 2016.

[Zhang *et al.* 2018] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin et Jian Sun. *Shufflenet : An extremely efficient convolutional neural network for mobile devices*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6848–6856, 2018.

[Zhi *et al.* 2007] Hui Zhi, Bing Ou, Bao-Ming Luo, Xia Feng, Yan-Ling Wen et Hai-Yun Yang. *Comparison of ultrasound elastography, mammography, and sonography in the diagnosis of solid breast lesions*. Journal of ultrasound in medicine, vol. 26, no. 6, pages 807–815, 2007.

[Zhi *et al.* 2017] Weiming Zhi, Henry Wing Fung Yueng, Zhenghao Chen, Seid Miad Zandavi, Zhicheng Lu et Yuk Ying Chung. *Using transfer learning with convolutional neural networks to diagnose breast cancer from histopathological images*. In International Conference on Neural Information Processing, pages 669–676. Springer, 2017.

[Zhu *et al.* 2016] Fan Zhu, Ling Shao, Jin Xie et Yi Fang. *From handcrafted to learned representations for human action recognition : A survey*. Image and Vision Computing, vol. 55, pages 42–52, 2016.

[Zoph *et al.* 2018] Barret Zoph, Vijay Vasudevan, Jonathon Shlens et Quoc V Le. *Learning transferable architectures for scalable image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8697–8710, 2018.

Notations

(A)	Adénose
ACO	Optimisation par colonies de fourmis
AMIDA13	Assessment of mitosis detection algorithms 2012
ANN	Réseaux de neurones artificiels
ARA-CNN	Accurate, reliable and active CNN
BAT	Algorithme des chauves-souris
(BC)	cancer du sein
BCNN	Bilinear convolutional neural networks
BGD	Descente de gradient à mini-lots
BoVW	Bag-of-visual words
CAD	Systèmes d'aide au diagnostic
CAMELYON16	Challenge on cancer metastasis detection in lymph node
CasNN	Deep cascaded neural network
CBFS	Correlationbased feature selection
CD	Carcinome canalaire
CDT	Transformation de distribution cumulative
CE	Entropie croisée
CENTRIST	Census transform histogram
C-FV	Convnet-based FV
CFV	Compressing fisher vector
CLL	leucémie lymphocytaire chronique
CNB	Core needle biopsy
CNN	Réseau de neurones convolutif
COCO	Common Objects in Context
CRF	Conditional random field
C-RSPM	Collateral representative subspace projection modeling
CS	Recherche coucou
CT	Tomodensitométrie
DBN	Réseaux de croyances profondes
(DC)	Carcinome canalaire
DCNN	Deep Convolutional Neural Network
DC	depthwise convolution
DL	Apprentissage profond
DLBCL	Lymphome diffus à grandes cellules B
DNN	Réseaux d'apprentissage profond
DOM	Moments orthogonaux discrets
DRN	Deep regression networks

DSC	Convolutions séparables en profondeur
DT	Arbres de décision
EB	Excision biopsy
EP	Tissus épithélial
(F)	Fibroadénome
FA	Algorithme des lucioles
FC	Couche entiérement connectée
F/C	Fisher/Correlation algorithm
FCN	Fully convolutional network
FF-CNN	Deep fused fully convolutional neural network
FL	lymphome folliculaire
FLD	Fisher linear discriminant
FN	Faux négatifs
FNA	Fine-needle aspiration
FP	Faux positifs
GA	Algorithmes génétiques
GAN	Réseau contradictoire génératif
GD	Descente de gradient
GLCM	Grey level co-occurrence matrix
GOFAI	Good old fashioned artificial intelligence
GPU	Processeur graphique
GRP	Gaussian random projection
GRU	Gated recurrent units
HC + CNN	Hybrid handcrafted features and CNN features
H &	E hématoxyline et éosine
HL	Lymphomes hodgkin
HOG	Histograms of oriented gradients
HOF	Histograms of optic flow
HPF	High power field
IA	Intelligence artificielle
ICAR18	International conference on arabidopsis research 2018
ICBN	Incremental boosting convolution networks
ICPR12	International Conference on Pattern Recognition 2012
IFV	Improved fisher vector
ILSVRC	ImageNet large scale visual recognition competition
IRLB	Inverted residual with linear bottleneck
KFDA	Kernel fisher discriminant analysis
KHA	Krill held algorithm
KNN	Plus proches voisins
LBP	Local binary patterns
(LC)	Carcinome lobulaire
LR	Learning rate (taux d'appentissage)
LRN	Local response normalization
LSTM	Mémoire à long-court terme
L-view	Large-view

MAP	Mean average precision
MargHists	Marginal colour histograms
MBH	Motion Boundary Histograms
(MC)	Carcinome mucineux
MCL	lymphome à cellules du manteau
MFF−CNN	Deep multi−scale fused fully convolutional neural network
ML	Apprentissage automatique
MLP	Perceptron multicouche
MoNN	Mixture of neural−network experts
MR−CN-PV	Multi−resolution convolutional network
MRI	Imagerie par résonance
mRMR	Minimum redundancy maximum relevance
MSE	Erreur moyenne quadratique
MSSN	Multi−scale and similarity learning convnets
NAG	Descente de gradient accélérée de Nesterov
NAS	Score de l'atypie nucléaire
NGS	Système de gradation de Nottingham
NHL	Lymphomes non hodgkiniens
NIN	Network in network
PASCAL VOC	PASCAL visual object classification
PET	Tomographie par émission de positrons
PC	Pointwise convolution
RGB	Red green blue
(PC)	C papillaire
PCA	Principal component analysis
PSO	Optimisation par essaim de particules
(PT)	Tumeur phyllode
R−CNN	Régions avec réseaux de neurones convolutif
RELU	Unité linéaire rectifiée
RNN	Réseau de neurones récurrent
ROI	Région d'intérêt
RPN	Region proposal network
SAE	Auto−encodeurs empilés
SDR	Deparation−guided dimension reduction
SDT	Subcategory discriminant transform
SGD	Descente de gradient stochastique
SSAE	Stacked sparse autoencoder
SST	Stain−style transfer
ST	Tissus stroma
SVHN	Street View House Numbers
SVM	Machines à vecteur de support
(TA)	Adénome tubulaire
TN	Vrais négatifs
TP	Vrais positifs
TUPAC16	Tumor Proliferation Assessment Challenge 2016

US	Ultrason
VGG−VD	Pretrained VGGNet16 on ImageNet
WMVA	Weighted majority voting alogrithm
WND	Weighted neighbor distances
WND−CHARM	Weighted neighbor distances using a compound hierarchy of algorithms representing morphology
WSD	Whole slide digital scanners
WSI	Whole slide images
YOLO	You only look once

الملخص

الرؤية الحاسوبية هي أحدى مجالات علم الحاسوب التي تمكن الأنظمة الأوتوماتيكية بالتعرف على المعطيات المرئية (الصورة والفيديو). تستخدم هذه الأنظمة عادة لاداء مهام التوصية. في السنوات الأخيرة، تزايد كمية البيانات الرقمية ساهم بشكل كبير في الإهتمام المتزايد بأنظمة الرؤية الحاسوبية وذلك لمعالجة هذه الكمية المعتبرة من المعلومات وتسهيل إستخراج المعارف المهمة منها.

تعتمد أنظمة الرؤية الحاسوبية بشكل أساسي على طرق تعلم الآلة وطرق التعلم العميق. في السنوات الأخيرة، ساهمت الكميات المعتبرة من المعلومات ووحدات معالجة الرسوميات القوية في تشجيع الباحثين على إستغلال طرق التعلم العميق. تتميز هذه التقنيات بأدائها الجيد على الكميات المعتبرة من البيانات، إضافة إلى ذلك، تتميز أيضا بقدرتها على الإستخلاص الأوتوماتيكي للميزات من البيانات الغير المنظمة، مثل الصور. أستخدمت طرق التعلم العميق في العديد من التطبيقات في مجال الرؤية الحاسوبية وذلك من أجل أداء مهام مختلفة مثل التصنيف، الكشف، وتقسيم الصور الرقمية.

في هذه الأطروحة، وجهنا إهتمامنا بشكل خاص لإستخدام صنف خاص من خوارزميات التعلم العميق من أجل تصنيف الخصائص النسيجية للصور. في هذا السياق، إقترحنا العديد من الطرق من أجل معالجة مختلف المشاكل المتعلقة بتطبيق طرق التعلم العميق لمعالجة هذه الصور. تعتمد التقنيات المقترحة بشكل رئيسي على تقنيات التنظيم ، التعلم الجماعي، ونقل التعلم. طرق التعلم الجماعي تساعد على حل مختلف المشاكل المتعلقة بالتباين المرتفع، التحيز، والتأثر المعتبر لطرق التعلم العميق بتغير البيانات. من ناحية أخرى، طرق نقل التعلم تستخدم من أجل حل مشاكل طرق التعلم العميق على الكميات المحدودة من البيانات.

Résumé

La vision par ordinateur est un champ d'étude qui permet aux systèmes automatiques à reconnaitre les entrées visuelles pour les exploiter dans des tâches de recommandation. Dans ces dernières années, la quantité des images et des vidéos a largement augmenté. L'exploitation des systèmes de vision par ordinateur pour l'analyse de cette quantité d'informations devient importante afin d'extraire de l'information pertinente. Les systèmes de vision par ordinateur sont basés essentiellement sur les méthodes d'apprentissage automatique (ML) et d'apprentissage profond (DL). Avec l'augmentation de la quantité de données et la disponibilité du matériel puissant, les méthodes DL ont connu un grand intérêt en raison de leur bonne performance sur les grands volumes de données et leur capacité d'extraction de caractéristique dans le cadre des données non structurées. Ces techniques étaient exploitées dans différents sous domaines

en vision par ordinateur pour effectuer plusieurs tâches : classification, localisation, détection, et segmentation.

Dans le contexte de la présente étude, nous nous intéressons à la classification des images histopathologiques par les méthodes DL, précisément par les réseaux de neurones convolutifs (CNN). Dans ce cadre, nous avons proposé plusieurs approches pour répondre aux différents problèmes liés à l'application des techniques DL en classification de ce type d'images.

Les approches proposées sont basées essentiellement sur les techniques de régularisation, les méthodes ensemblistes, et les stratégies d'apprentissage transféré et de fine tuning. Il est intéressant de noter que les méthodes ensemblistes sont exploitées afin de résoudre les différents problèmes liés à la variance élevée, le sur-apprentissage, et la sensibilité des réseaux DL au changement de données. En plus, elles permettent de combiner les prédictions de plusieurs modèles, et cela génère des décisions plus robustes et stables au changement de données. D'autre part, les techniques d'apprentissage transféré et de fine tuning sont utilisés afin de résoudre le problème de sur-apprentissage sur les volumes limités de données.

Abstract

Computer vision is defined as a field of computer science that enable automatic systems to identify visual inputs. These systems are usually used to perform recommendation tasks. In recent years, the amount of digital data, such as images and videos, have largely increased. In this regard, the exploitation of computer vision systems became essential to maintain these volumes and also to extract relevant information.

Computer vision systems are based on machine learning (ML) and deep learning (DL) methods. Many factors, such as the growing volumes of data and the availability of powerful graphical processing units (GPU) have encouraged the computer vision community to exploit DL methods. These techniques are characterized by their efficiency on large volumes of data and also by their capacity to extract features from non-structured data. DL methods have been exploited in different applications in computer vision to perform several tasks: classification, localization, detection, and segmentation.

In this study, we are particularly interested in the classification of histopathological images by convolutional neural networks (CNN). In this context, we have proposed several pipelines to solve the different issues related to the application of DL methods on these types of images. The proposed frameworks are based mainly on regularization methods, ensemble learning techniques, and transfer learning and fine-tuning strategies. We should note that ensemble learning techniques are used to solve the different issues related to the high variance, overfitting, and the sensevity of neural networks to data changes. On the other hand, transfer learning and fine-tuning strategies are used to solve the overfitting problem on limited volumes of data.