République Algérienne Démocratique et Populaire

UNIVERSITE KASDI MERBAH OUARGLA

Faculté des Nouvelles Technologies de l'Information et de la Communication

Département d'informatique et Technologie de l'information

Mémoire

MASTER ACADEMIQUE

Domaine : Mathématiques et Informatique

Filière : Informatique

Option: Informatique industrielle

Thème

Utilisation de la perception visuelle pour la fusion d'image multi-focus

Présenté par :

Belmesmar Bachir & Bebboukha Yacine

Devant le jury :

M. . Belhadj Mourad	**MAA**	**UKM Ouargla**	**Président**
M. ZGA ADEL	**MAA**	**UKM Ouargla**	**Examinateur**
M. ZERDOUMI OUSSAMA	**MAA**	**UKM Ouargla**	**Encadreur**

Année universitaire : 2021 /2022

REMERCIEMENTS

Nous tenons à remercier le bon dieu, le tout puissant de nous donner la patience, la santé et le courage pourfinir ce travail.

Nous remercions profondément notre

encadreur: Monsieur ZERDOUMI OUSSAMA qui nous aencouragé à faire le maximum d'efforts dans ce travail, sansses encouragements cette mémoire n'aurait sans doute pas abouti.

Et nous n'oublions pas non plus de remercier tous ceux qui nous ont aides de loin ou de près.

Dédicace

C'est avec joie que nous dédie ce modeste travail : À nos très chers parents qui nous espérons rendre fière,

pour leurs patiences et encouragements. Que dieu les protègent.

À nos amis et nos collègues pour les bons moments que nous avons passé avec eux

Belmesmar Bachir

&

Bebboukha Yacine

Résumé

Les méthodes de fusion d'images multi-focus basées sur les réseaux de neurones convolutifs (CNN) a récemment attiré une attention considérable. Ce dernier représente un type de réseau de neurones artificiels, dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux. Dans ce travail, nous avons recouru le CNN entraîné on trois dataset pour créer une application qui permet de combiner deux images partiellement focalisées en une seule image entièrement fusionnée. Pour la programmation, nous avons utilisé l'enivrement de python et la bibliothèque d'apprentissage automatique Pytorch. Les résultats obtenus montrent que les techniques de fusion des images multi-focus à la base de l'apprentissage profond donne une image fusionnée de haute précision et qui sont peu coûteuses et ne prennent pas beaucoup de temps.

Mots clés: système visuel humain, image fusion, image multi-focus, CNN.

ملخص

جذبت طرق دمج الصور متعددة البؤر القائمة على الشبكات العصبية التلافيفية (CNNs) اهتمامًا كبيرًا مؤخرًا. تمثل هذه الأخيرة نوعًا من الشبكات العصبية الاصطناعية ، حيث يكون نمط الاتصال بين الخلايا العصبية مستوحى من القشرة البصرية للحيوانات. في هذا العمل استخدمنا CNN مدربة على ثلاثة مجموعات بيانات لإنشاء تطبيق يسمح بدمج صورتين مركزة جزئيًا في صورة واحدة مدمجة بالكامل. بالنسبة للبرمجة استخدمنا بيئة Python ومكتبة Pytorch للتعلم الآلي, تظهر النتائج التي تم الحصول عليها أن تقنيات دمج الصور متعددة البؤرة القائمة على التعلم العميق تعطي صورة عالية الدقة مدمجة وغير مكلفة ولا تستغرق الكثير من الوقت.

الكلمات المفتاحية: النظام البصري البشري , صورة مدمجة , صورة متعددة التركيز , الشبكة العصبية التلافيفية

Abstract

Methods of fusion image multi-focus based on convolutional neural networks (CNNs) have recently attracted considerable attention, This last represents a type of artificial neural network, in which the connection pattern between neurons is inspired by the visual cortex of animals. In this work, we used the trained CNN on three datasets to create an application that combines two partially focused images into a single fully fused image. For programming, we use environment python and the Pytorch machine learning library, The results obtained show that multi-focus image fusion techniques based on deep learning give a high-precision fused image that are inexpensive and do not take much time.

Keywords: human visual system, fusion image, multi-focus image, CNN.

Table des matières

Table des Figures

Introduction général

Ces dernières années, la fusion d'images a été utilisée dans de nombreuses variétés d'applications telles que la télédétection, la surveillance, le diagnostic médical et les applications de photographie. En raison de la profondeur du champ limitée des lentilles optiques des appareils photo, il est difficile de capturer une seule image dont tous les composants soient évidents. Par conséquent, certaines zones des images capturées avec des capteurs de caméra sont floues. Il est possible d'enregistrer des images avec différentes profondeurs de focalisation à l'aide de plusieurs caméras. Le processus de fusion d'images est défini comme la collecte de toutes les informations importantes à partir de plusieurs images, et leur inclusion dans une seule image est plus informative et précise que n'importe quelle image source unique, et elle comprend toutes les informations nécessaires. Le but de la fusion d'images n'est pas seulement de réduire la quantité de données, mais aussi de construire des images plus appropriées et compréhensibles pour la perception humaine et machine. En vision par ordinateur, la fusion d'images multi-capteurs est le processus de combinaison d'informations pertinentes provenant de deux ou plusieurs images en une seule image. De nombreuses recherches sur la fusion d'images multi-focus ont été effectuées ces dernières années. L'apprentissage profond (DL) a prospéré dans plusieurs applications de traitement d'image et de vision par ordinateur.

Dans ce mémoire, nous allons utiliser la méthode domaine spatial basée sur un schéma de division de blocs et les réseaux de neurones convolutifs(CNN).L'objectif de notre travail est de développer une technique de fusion d'images multi-focus pour obtenir une image fusionnée de haute précision qui convient mieux à la perception humaine ou machine et à d'autres tâches de traitement d'images.

Afin d'atteindre notre objectif, nous avons suivi le plan de travail suivant :

- Le premier chapitre est axé sur une étude bibliographique sur les sujets de la perception visuelle, où nous allons présenter quelques propriétés de l'œil et du système visuelle humain, ainsi que les techniques de calcul et modélisation de l'attention visuelle.
- Le second chapitre concerne les quelques notions de base du domaine de traitement d'image telles que : les types d'images, les caractéristiques d'images, la spécificité de la fusion d'images par rapport à la fusion de données, les méthodes et les catégories de fusion d'images.
- Le troisième chapitre est consacré aux réseaux de neurones convolutifs (CNNs).Nous présenterons également les principes de l'apprentissage profond et l'architecture et les couches de CNN
- Le quatrième chapitre concerne la conception de notre application, la méthode d'implémentation de notre travail qui se base sur le langage de programmation Python, ainsi qu'une présentation des analyses et des résultats de notre travail.

Enfin, nous achèverons notre travail de recherche par une conclusion générale sur l'amélioration des travaux que nous aborderons tout au long de ce mémoire.

Chapitre 1

Perception visuelle

1.1 Introduction

La perception humaine est le pont qui relie nos sensations et notre conscience au monde réel. Ce pont est construit à travers le processus d'intégration multi-sensorielle que notre système nerveux effectue avec tous les signaux générés par nos sens.

La perception est liée à l'analyse et au traitement du signal dans de nombreux aspects. Lorsque nous essayons d'explorer ces champs, nous devons utiliser la psychophysique. Cette méthodologie scientifique a contribué à de nombreuses avancées fondamentales dans la détermination des seuils de perception et de discrimination de nos sens, en réalisant également l'évaluation de ses représentations mentales. En plus de la méthodologie psychophysique pour l'étude des perceptions, l'électronique et l'informatique ont joué un rôle important. Chacune de ces disciplines a influencé notre compréhension du système neuronal. Elles participent à la détection de l'activité électrique neuronale et à la simulation du comportement neuronal.

Dans ce chapitre, nous nous concentrons principalement sur l'étude de la perception visuelle qui est la faculté de voir, d'examiner et de donner un sens à toutes les informations visuelles proches

1.2 Le système visuel humain

Le système visuel se compose de plusieurs éléments physiologiques de la cornée au cortex visuel primaire. La vision humaine commence lorsque la lumière entre dans la cornée, se concentre à l'arrière de l'œil dans une membrane sensible qui est la rétine. Plus précisément, la lumière est concentrée sur les cellules photo réceptives (tiges et cônes) de la rétine. Les photo récepteurs transforment alors la lumière en signaux neuronaux. Ces signaux sont traités par des réactions complexes et des processus d'anticipation initiés pas différentes parties du cerveau : la rétine, les ganglions centraux jusqu'au cortex visuel. Le cortex visuel d'association permet d'interpréter toute la scène visuelle, et donc de percevoir l'objet L'œil est l'organe de la vision, il nous permet de capter la lumière de notre environnement et de la convertir en message nerveux, lequel est transmis au cerveau qui l'analyse.[01]

1.2.1. La structure de l'œil

L'œil, ou globe oculaire est l'organe de la vision. Il est de faible volume (6.5 cm3), pèse 7 grammes, et a la forme d'une sphère d'environ 24 mm de diamètre, complétée vers l'avant par une autre demi sphère de 8 mm de rayon, la cornée [02].

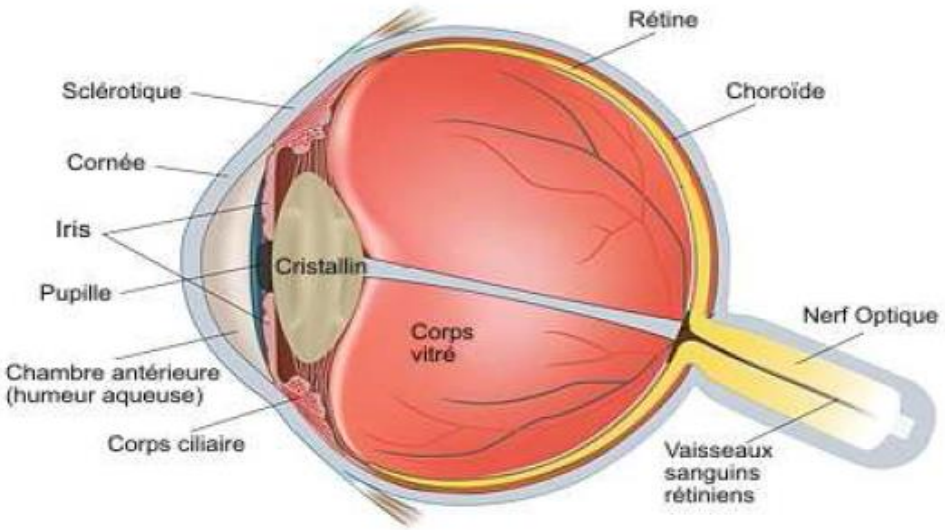

Figure 1.1: Schéma General De L'œil [02].

La cornée : qui est une membrane solide et transparente qui entoure et protège l'œil contre les micro-organismes extérieurs. Elle forme une demi-sphère d'environ 8 a 11 millimètres de diamètre. Elle joue le rôle de lentille en assurant 80% de la réfraction de la lumière. La lumière pénètre dans l'œil par celle-ci. La cornée contient 78% d'eau qui est maintenu par les larmes reparties sur la surface de l'œil par les battements des paupières.

L'humeur aqueuse :qui est un liquide transparent presque entièrement compose d'eau salée, et qui régule la pression a l'intérieur de l'œil.

La pupille : qui est un trou central de l'iris par lequel la lumière (rayons lumineux) pénètre dans l'œil (rétine). Son diamètre dépend de l'ouverture de l'iris.

L'iris : diaphragme perce par la pupille. C'est un tissu musculaire en forme d'anneau colorée (bleu, marron, vert…) dont son ouverture varie, entre 2,5 et 7 millimètres, afin de modifier la quantité de lumière pénétrante dans l'œil pour éviter l'aveuglement. Lors de la présence de lumière vive (en plein soleil), ou pour capter le plus de rayons, lors d'une obscurité (la nuit).

Le cristallin : Le cristallin est une lentille transparente déformable qui est responsable de la convergence et de la divergence de la lumière. Celui-ci se déforme et devient plus convergent de manière a ce que l'image se forme toujours nette sur la rétine lorsque l'objet se rapproche de l'œil (on dit qu'il accommode ou fait la mise au point de l'image). Il possède une forme biconvexe. Lorsqu'il ne peut s'accommoder, nous avons des problèmes de vision tels que la myopie (œil trop long), l'hypermétropie (œil trop court), la presbytie (cristallin impuissant -> vieillesse).

L'humeur vitrée : qui est un liquide situe derrière le cristallin et qui occupe 90% du volume de l'œil. Il maintient la rétine en place et la protège en amortissant les chocs et en garantissant la rigidité de l'œil.

Le nerf optique :qui transmet les informations de l'œil au cerveau. Il y a également d'autres composantes dans l'œil qui n'interviennent pas forcement dans la traversée de l'œil par la lumière, mais qui sont très utiles :

La choroïde, qui est une couche vasculaire de couleur noire qui nourrit les photorécepteurs de la rétine.

La sclérotique, qui est une enveloppe de protection recouvrant 5/6 de la surface de l'oeil. L'oeil lui doit sa blancheur et sa rigidité. **[02]**

1.2.2 Structure de la rétine

La rétine est l'organe le plus important de l'œil. Elle mesure environ 0,5 mm d'épaisseur, et recouvre les trois quarts de l'intérieur du globe oculaire.

La rétine est un tissu nerveux constitue de 3 couches de neurones. Certains d'entre eux, les photorécepteurs sont des neurones qui détectent la lumière grâce a des molécules appelées pigments la nature et la quantité de pigment influent sur la perception des images. Un pigment correspond a une longueur d'onde. Il existe 2 types de photorécepteur :

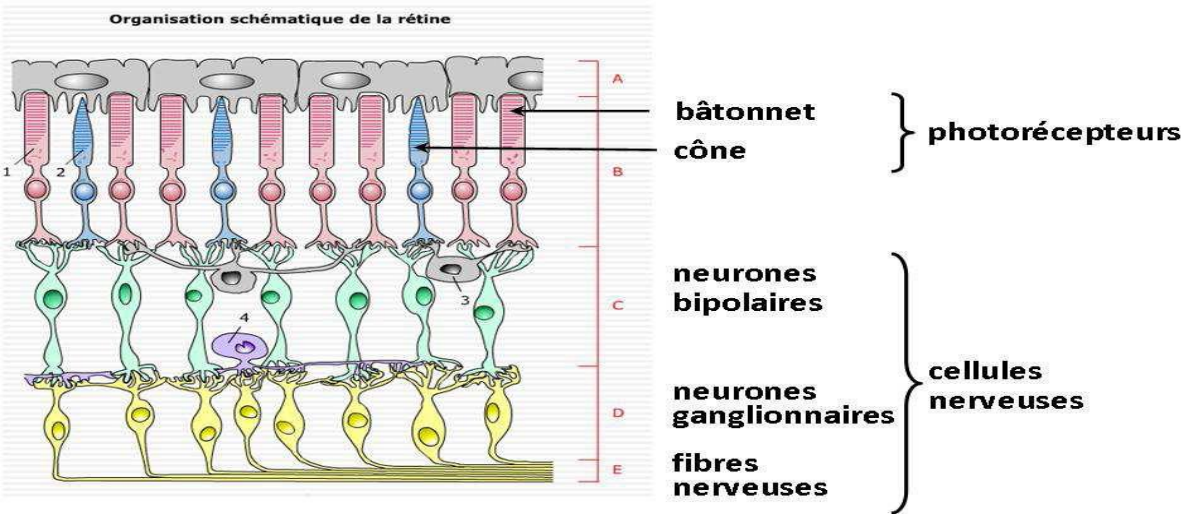

Figure 1.2: Organisation Des Couches Cellulaire De La Rétine [02].

Les cônes

Les cônes, eux, représentent seulement 5% des photorécepteurs (5 millions) et permettent la vision des couleurs ainsi que la perception des images détaillées. On peut constater la présence de trois types de cônes en fonction des pigments qu'ils détiennent l'opsine S, l'opsine M et l'opsine L. Chaque cône présente un spectre d'absorption de la lumière spécifique, défini par des longueurs d'ondes allant a peu prés de 400 a 800 nanomètres. Ces spectres ont une absorption maximale soit:

- dans le bleu, entre 420 et 500 nm pour les cônes S
- dans le vert, entre 500 et 560 nm pour les cônes M
- dans le rouge, entre 600 et 750 nm pour les cônes L

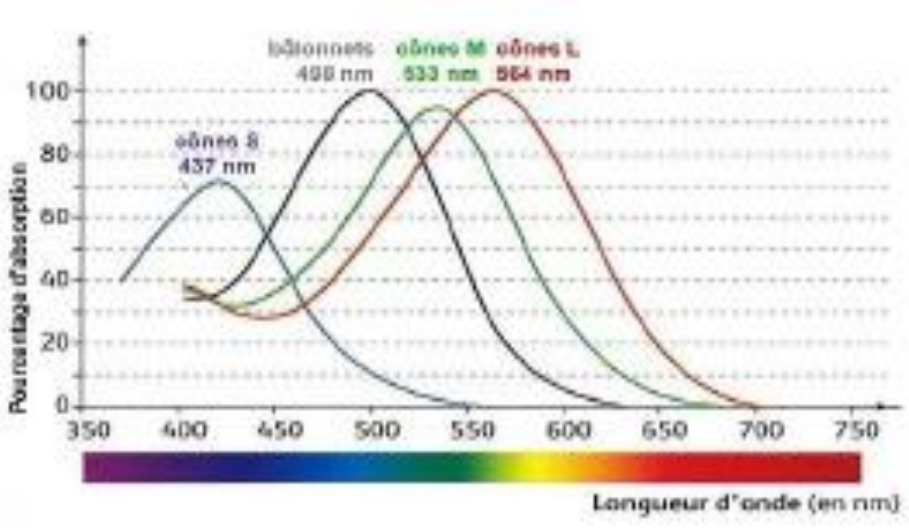

Figure 1.3: Sensibilité Des Différents Types De Photorécepteur A La Longueur D'onde De La Lumière [02]

Chaque couleur, c'est-a-dire chaque longueur d'onde, est perçue grâce a l'activité de trois types de cônes (cônes S, M et L). C'est ce qu'on appelle la trichromatie. En additionnant les trois couleurs dites primaires (rouge, vert, bleu), on obtient toutes les autres couleurs du spectre. Il s'agit alors d'une synthèse additive. Grace a ce fonctionnement, on peut distinguer 2,3 millions de couleurs différentes. Le daltonisme est l'absence d'un cône, ce qui empêche la distinction de deux couleurs qui paraissent alors identiques.

Les bâtonnets

Les bâtonnets, qui représentent 95% des photorécepteurs, peuvent réagir a des éclairements très faibles et sont donc utilises pour distinguer différents niveaux de clarté. Ils ne possèdent qu'un type de pigment : la rhodopsine. Ils ne sont pas dans la fovéa (ou sont les cônes) mais ils sont repartis dans la rétine (120 millions). Ils perçoivent mal les couleurs car ils ont peu de liaisons directes avec le nerf optique, contrairement aux cônes. **[02]**

1.3 L'attention visuelle

1.3.1 Qu'est-ce que l'attention visuelle ?

L'attention est définie comme la capacité de concentrer son activité mentale sur un objet déterminé. Mais elle signifie aussi une marque d'intérêt ou d'affection : « des petites attentions », ou un danger imminent : « Attention ! », ou au fait d'être conscient de quelque chose et d'y prendre garde :
« faire attention à ». [03]

L'attention visuelle est liée au changement d'état attentionnel d'un observateur tout en gardant constante l'image rétinienne, affectant ainsi les performances de perception et l'activité des neurones sensoriels.

L'attention est un processus cognitif permettant de sélectionner certains aspects de l'environnement. Cela permet de recueillir des informations propres à un objet ou à une partie de l'espace (attention spatiale) an de l'analyser. La notion d'attention est étroitement liée à la notion de ressources attentionnelles. [04]

1.3.2 Quel processus attentionnel pour la sélection visuelle ?

1.3.2.1 Bottom-up versus top-down : Les processus visuels bottom-up renvoient à la saillance physique d'une scène visuelle, et ils dépendent directement des entrées sensorielles. Ils débutent dans les aires cérébrales de bas niveaux qui vont ensuite transmettre les informations vers les aires cérébrales supérieures. Les processus top-down renvoient quant à eux fonctions des objets présents dans la scène visuelle et à leur localisation. Ils sont associés au but suivi par l'individu, à son système de récompense, ou encore à un danger potentiel. Ils sont initiés dans les aires cérébrales

supérieures, puis les informations issues de ces processus sont transmis vers les aires de bas niveau a n de moduler le traitement des entrées sensorielles. Ils ne dépendent donc pas directement de ces dernières.

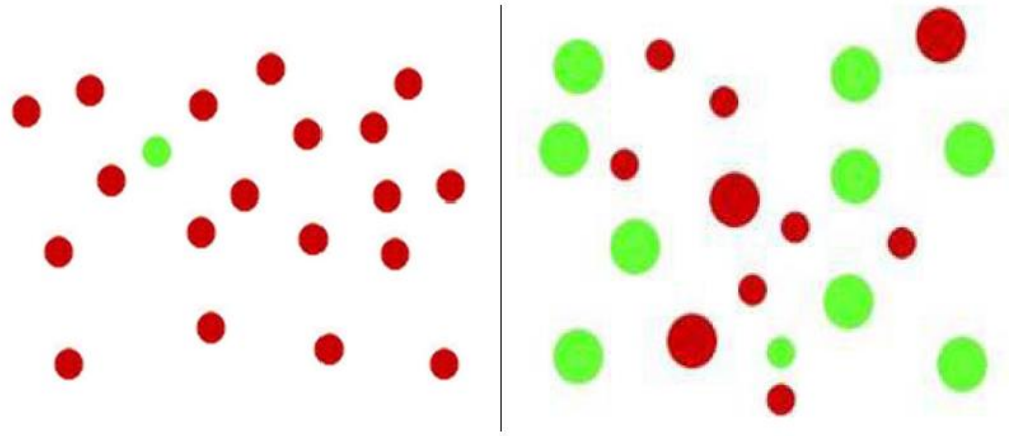

Figure 1.4.: la tache dans cet exemple est de trouver le petit rond vert[04]

Deux types de traitement top-down peuvent être distingués (Baluch & Itti, 2011)[05] Chun & Jiang, 1998; Gilbert & Sigman, 2007) [06] . Les processus top-down volontaires_ (volitional top-down procès) permettent par exemple, lors d'une illusion visuelle, de basculer entre deux interprétations (voir Figure 1.4, à gauche). A l'inverse, les processus top-down involontaires_ (mandatory top-down procès) sont automatiques et persistants. Ils contraignent, par exemples lors d'un autre type d'illusion visuelle, à une seule interprétation, et cela alors même que la nature de l'illusion est consciente chez le participant (voir Figure 1.5, à droite).[04]

Contrairement à la distinction entre processus pré-attentif et attentif vu précédemment, cette taxonomie s'attache d'avantage à ce qui dirige l'attention qu'à ce qui caractérise les processus. Néanmoins, il est possible de rapprocher ces deux taxonomies. Dans les faits, les prédictions posées par les modèles sous-jacents à ces deux distinctions se recouvrent suffisamment pour que cela ne gêne pas la généralisation de ces modèles utilisant les deux nomenclatures. En particulier, les phénomènes pop-out et set-size sont également utilisés pour définir les processus bottom-up et top-down. Par exemple, si un T rouge est présenté au milieu de T noirs, le regard se portera immédiatement dessus, puisque ce T est saillant, et la localisation de ce T sera connue immédiatement, puisque une seule feature le distingue des autres.

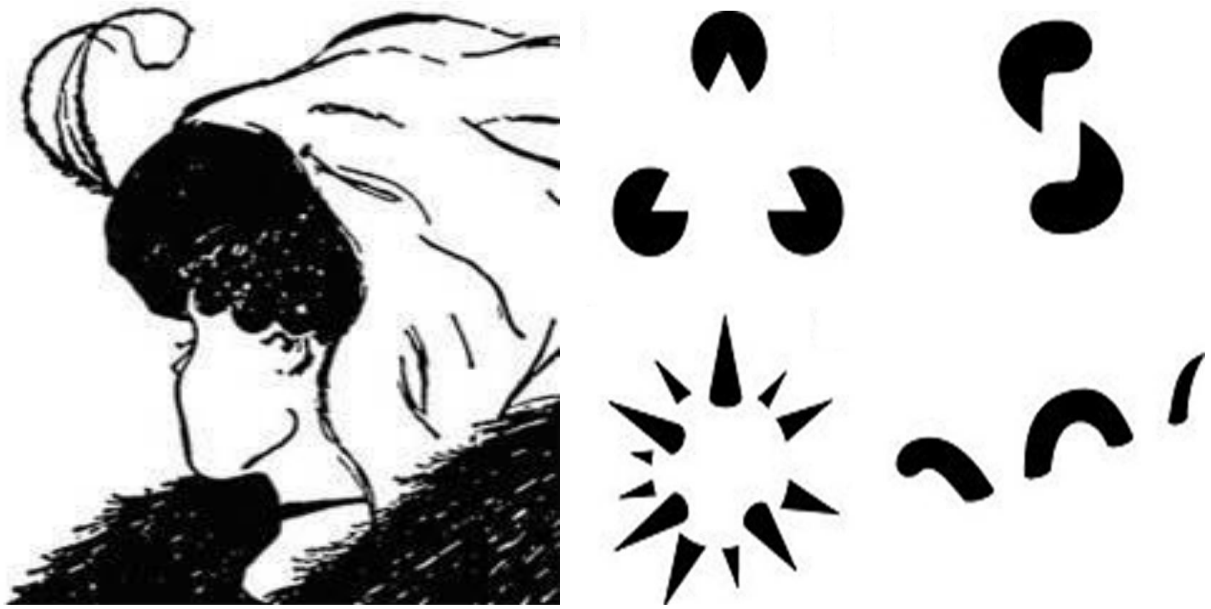

Figure 1.5.: a gauche, l'illusion de boring (1930) mettant en évidence un exemple de Processus top-down volontaire[04]

D'autre part, le modèle de Treisman & Gelade (1980) [07], utilisant les notions de processus attentif et pré-attentif, a donné lieu a des modèles tel que celui de Koch & Ullman (1985)
puis au modèle computationnel de Itti et al. (1998), basés tout deux sur les notions de processus bottom-up et top-down. De plus, Treisman & Gelade (1980) [07] précisent que pour les processus attentifs, la localisation d'une conjonction doit précéder l'identification de celle-ci, tandis que pour une feature, sa localisation et son identification peuvent être indépendantes. Cela repose sur le fait que des connaissances au préalable portent sur la localisation d'une conjonction concernant l'objet de la recherche visuelle (Posner & Cohen, 1984), rejoignant ainsi la définition des processus top-down qui nécessitent des connaissances et des attentes vis à vis de l'objet cherché, dont sa localisation.

1.3.2.2 Exogène versus endogène Une nomenclature proche de la notion de processus bottom-up et top-down introduit celle de processus endogène et exogène. Cette taxonomie met l'accent sur la source de l'information (dedans / dehors), tandis que la précédente met l'accent sur la circulation de l'information (de l'extérieur vers l'intérieur ou inversement). Les processus endogènes dépendent des connaissances préalables qu'une personne a de l'environnement et de ce que cette personne veut accomplir. Les processus exogènes sont des réponses proches du réflexe à l'environnement extérieur. Ils seront notamment dominants lorsque la personne a très peu de connaissance a priori dans une situation donnée, par exemple dans un lieu inconnu, ou lorsqu'aucun but n'est défini, par exemple lors d'une tâche d'exploration libre.

1.3.3 Des modèles différents pour des processus différents

1.3.3. 1 Distinctions entre les modèles théoriques et computationnels

Parmi les modèles proposés dans l'attention visuelle, il est possible de distinguer deux types de modèles : ceux dont l'objectif principal sera de valider un cadre théorique et ceux dont l'objectif sera d'émettre des prédictions quantitatives.

Cette distinction que l'on propose ici n'est pas absolue, et la majorité des modèles actuels tendent à inclure ces deux facettes. En, les modèles plus théoriques permettent d'émettre des prédictions qualitatives mais également parfois quantitatives, et les modèles plus quantitatifs se basent néanmoins sur un cadre théorique. Cependant, pour la majorité des modèles, une de ces deux facettes est prédominante. Ce point est important car il permet de mieux appréhender la suite de ce travail. Comme nous allons le voir, l'objectif actuel de la plupart des modèles bottom-up, le plus souvent bio-inspirés, est de pouvoir prédire la position du regard, tandis que celui des modèles top-down est principalement d'agrandir et d'élargir nos connaissances des processus attentionnels. De ce fait, un des enjeux majeurs de la recherche dans ce domaine serait de pouvoir coupler ces deux approches.

Ces deux objectifs théorique et prédictif sont en partie liés à un angle d'approche différent modulé en partie par la discipline concernée : les sciences cognitives et les sciences pour l'ingénieur. Les limites et les avantages sont donc très différents selon l'approche, comme nous le verrons par la suite pour les modèles bottom-up et top-down. Nous verrons également que l'idéal serait d'arriver à trouver un modèle mixte, tant sur le plan théorique en incluant les processus bottom-up et top-down, que sur le plan opérationnel en regroupant les deux types d'approches : cadre théorique et prédiction quantitative.[04]

La frontière entre ces deux approches et entre les deux disciplines est de plus en plus poreuse et des modèles mixtes sont proposés. Toutefois, en fonction de l'orientation académique des chercheurs, cela donne le plus souvent non pas des modèles parfaitement mixtes, mais des modèles soit top-down auxquels une composante bottom-up a été ajoutée, mais qui ne sont pas implémentables, soit bottom-up auxquels une composante top-down a été ajoutée, mais qui ne sont pas satisfaisants d'un point de vue théorique quant à l'implication des processus top-down.

1.3.3.2. Les modèles bottom-up : modèles computationnels

Un modèle de l'attention basé sur la saillance visuelle

A la suite des travaux de Treisman & Gelade (1980) [07] et de leur (Feature Intégration Theory), un modèle théorique de carte de saillance a été proposé par Koch & Ullman (1985). Ce modèle est doté d'une dynamique interne qui génère des shifts attentionnels. C'est le premier modèle dans la

littérature de contrôle de l'attention visuelle qui a une architecture computationnelle basée sur les neurosciences. Il s'articule autour de la notion de carte de saillance, qui reçoit une entrée sensorielle et fournit une stratégie de l'allocation de l'attention : le centre de l'attention balaye la carte de saillance par ordre décroissant de saillance.

Ce modèle théorique a ensuite servi de base au modèle proposé par Itti et al. (1998), qui est, parmi les modèles quantitatifs, le modèle bottom-up sans doute le plus connu. Ces auteurs ont adapté le modèle de Koch & Ullman (1985) afin de pouvoir l'implémenter et ainsi proposer des prédictions quantitatives. C'est donc un modèle quantitatif de la saillance visuelle d'une scène. Plus précisément, le modèle de Itti et al. (1998) permet de calculer quelles sont les zones les plus saillantes. Ce modèle, et d'autres de la même classe, sont des modèles prédictifs de l'allocation de l'attention manifeste qui reposent sur l'hypothèse que l'observateur fixe en priorité les zones les plus saillantes.

Pour Itti & Koch (2001), les modèles prédisant l'allocation de l'attention en se basant sur les processus bottom-up reposent sur cinq points théoriques : la saillance perceptive du stimulus, qui dépend du contexte ; la carte de saillance, qui est une représentation unique et topographique de la scène visuelle ; l'inhibition du retour, qui empêche temporairement de fixer à nouveau une zone déjà fixée ; l'interaction entre les mouvements des yeux et l'allocation de l'attention ; l'interprétation de la scène et la reconnaissance des objets, qui vont modifier le choix des zones fixées (composante top-down).

Les différents modèles basés sur la saillance visuelle se distinguent généralement par la stratégie utilisée afin de trier les entrées sensorielles (choix des Feature) et d'extraire la carte de saillance (comparaisons et compétition des Feature). Nous allons détailler dans ce qui suit celui proposé par Itti et al. (1998) et certaines de ses évolutions, afin de mieux appréhender l'essence d'un modèle construit initialement autour des processus bottom-up.

Les entrées sensorielles Le modèle initial de Itti et al. (1998) permet de prédire l'attention portée sur une scène visuelle en fonction de certaines de ses propriétés (Feature). Ces Feature ont été définies comme étant celles provoquant un effet de pop-out. Il s'agit de la couleur, l'orientation, et l'intensité.

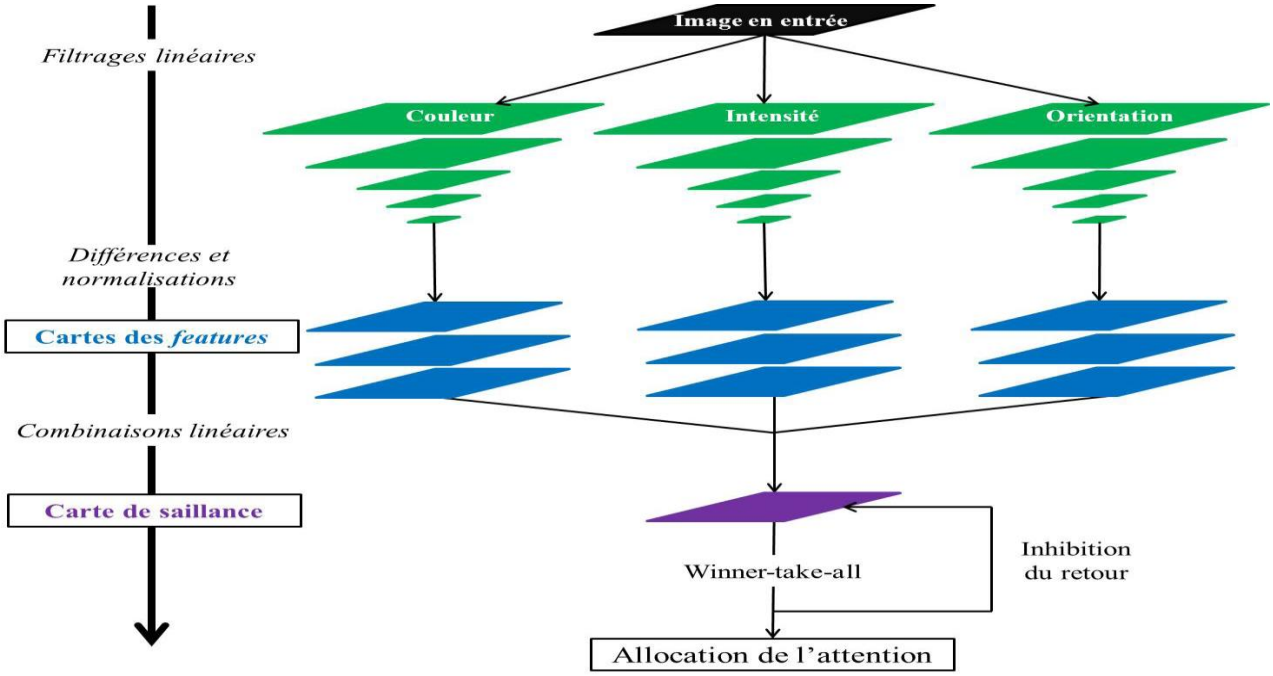

Figure 1.6.: le modèle décrit par Itti et al. (1998) puis par Itti & koch (2001) [04].

Par la suite, ce modèle reposant sur trois Feature s'est vu compléter avec le mouvement en tant que tel, permettant de poser des prédictions dans un environnement dynamique . L'hétérogénéité de la rétine a également été implémentée afin de rendre compte de l'anatomie de l'œil, permettant une perception différente des Feature en fonction de leur localisation par rapport à la position de l'œil de l'observateur , Target Acquisition Model). De même, d'autres évolutions permettent de prendre en compte certains effets, tel que le biais de centralité et le masquage visuel (Le Meur et al., 2006).
De même, la cohérence de ces modèles avec les données de la physiologie ne cesse d'augmenter, et des applications de plus en plus nombreuses voient le jour, notamment en robotique et pour le traitement d'image. Par ailleurs, de nombreuses autres Feature ont été mises en évidence (e.g. la résolution spatiale, le flux optique, la symétrie, la profondeur), mais des lacunes persistent dans les connaissances fondamentales (Wolfe & Horowitz, 2004), or les implémentations nécessitent une base solide de connaissances. Ce sont donc les trois Feature introduites précédemment qui restent les plus répandues[08].

La carte de saillance Le modèle quantitatif de Itti et al. (1998) s'est construit sur la base de l'intensité, de la couleur et de l'orientation.
Dans le modèle théorique, les trois Feature sont initialement traitées par trois canaux distincts. Ainsi, des Feature de la même modalité sont en compétition pour l'accès à la carte de saillance, puis

pour être sélectionnées (avec le mode winner-takesall), tandis que les différentes modalités contribuent indépendamment à la carte de saillance.

Une carte de saillance est réalisée pour chaque caractéristique issue de l'information bottom-up, à plusieurs échelles, puis ces cartes sont combinées an de donner une seule carte de saillance (Koch & Ullman, 1985). Ceci implique la présence de mécanismes de compétition locale entre les différentes cartes (Itti et al., 1998; Itti & Koch, 2000). Le modèle est ensuite doté d'une dynamique interne, avec l'inhibition du retour, permettant de générer des shifts attentionnels.

Limites des modèles bottom-up : leur validité en situation naturelle

D'un point de vue théorique, les modèles bottom-up sont par définition dépendants des entrées sensorielles uniquement, et sont donc in_exibles dans le sens où les prédictions seront toujours strictement identiques du moment que la scène visuelle ne varie pas. Or, les mouvements des yeux peuvent varier, alors même que l'environnement visuel est le même. C'est un des résultats principaux des travaux de Yarbus. Dans une de ses études (Yarbus, 1967), des photographies ont été présentées plusieurs fois aux participants. Les questions posées par rapport à ces photographies variaient, modulant ainsi la tâche du participant (e.g. questions : « Donner l'âge des personnes » ou « Se rappeler des positions des personnes et des objets dans la chambre »). Alors même que les photographies étaient strictement identiques (les mêmes entrées sensorielles),des patterns oculomoteurs différents ont été observés.

Cette étude, et celles qui ont suivi, tendent donc à mettre en évidence la nécessité de prendre en compte les processus top-down. Lors de situations naturelles, il semblerait même que ceux-ci soient dominants par rapport aux entrées sensorielles.

L'importance du but de l'observateur a donc été mise en évidence en démontrant que les mouvements des yeux dépendent des instructions données, et donc de la tâche (Howard et al., 2011; Yarbus, 1967). A cela s'ajoute le fait que plus la tâche se complexifie, et moins le mouvement des yeux dépend des propriétés de la scène visuelle.

La plupart de ces modèles bottom-up sont pensés, testés et validés en référence à des situations spécifiques, la majorité du temps l'exploration d'une image avec un délai court et en (free viewing) observation libre, donc sans objectif spécifiquement défini (Hwang et al., 2009). Or, de nombreux travaux de la littérature tendent à montrer que lorsqu'il s'agit d'une scène naturelle (e.g. une photographie), l'orientation du mouvement des yeux par la saillance visuelle (bottom-up) joue un rôle mineur, tandis que la sémantique des objets, la tâche et les connaissances a priori sont de meilleurs prédicateurs de la direction du regard. En effet, les modèles bottom-up rendent très peu

compte des patterns de fixations dans ces situations écologiques, les prédictions des modèles basés sur les processus top-down sont alors bien meilleurs.

Lors d'une expérience en environnement virtuel, Rothkopf et al. (2007) ont fait varier la saillance visuelle de différents éléments de l'environnement, ainsi que la tâche exécutée par les participants. Ils observent que la proportion des fixations, leur durée et les zones fixées dépendent de la tâche, avec un rôle minime de la saillance visuelle (bottom-up). Le contexte sémantique semble également prédictif. Puis ils ont implémenté le modèle bottom-up de Itti & Koch (2000), qu'ils ont comparé à leurs données comportementales, et les prédictions quantitatives se sont montrées non satisfaisantes. Ces auteurs ont donc pu conclure que les traitements bottom-up ne sont pas de bons prédicateurs de la direction du regard humain dans un environnement naturel.

Des recherches sur les mécanismes neuraux valident également le rôle joué par les traitements top-down. Nous pouvons citer l'étude de Li et al. (2004). Ces auteurs ont réalisé une expérience sur des singes qui devaient effectuer sur les mêmes stimuli soit une bissection de ligne, soit une tâche d'acuité visuelle. Alors que le stimulus visuel est strictement identique, ils ont pu observer une activation différente au sein de V1 selon la tâche exécutée. V1 étant la première aire corticale recevant l'information visuelle, il est intéressant de noter que l'influence de la tâche y joue « déjà » un rôle dans le traitement de l'information. Cette observation expérimentale est renforcée par des études en IRMf qui ont montré l'existence d'activation top-down au niveau du système visuel.

Les modèles strictement bottom-up sont donc mal adaptés aux tâches naturelles. C'est pourquoi de plus en plus d'auteurs vont introduire une composante top-down dans ces modèles afin de produire de meilleures prédictions.

1.3.3.3. Les modèles top-down : modèles théoriques

Un modèle de l'attention basé sur les attentes

Summer_eld & Egner (2009) [08] proposent un modèle basé sur les attentes (expectations). Ces attentes induisent des processus top-down « involontaires » et reflètent les informations a priori en fonction de leur possibilité ou de leur probabilité d'occurrences dans un environnement sensoriel futur. Il existe deux hypothèses concernant l'utilisation de ces attentes : les attentes peuvent guider la prise d'information, ou les attentes peuvent faciliter l'interprétation des entrées visuelles. En effet, il a été observé qu'un objet est identité plus rapidement lorsqu'il est présenté dans un contexte congruent que s'il l'est dans un contexte non-congruent (Bar, 2004).

Ces mécanismes top-down peuvent faciliter le traitement de l'information à différentes étapes. Pendant la phase d'anticipation, ils permettent d'anticiper les caractéristiques qui vont être présentes et leur position. Pendant la phase d'acquisition sensorielle, ils permettent l'accumulation

de preuves, et augmentent le ratio signal sur bruit afin de permettre aux caractéristiques visuelles attendues d'émerger. En présence de bruit, ils augmentent la trace du signal et du bruit, diminuent le temps de détection, mais augmentent les fausses alarmes. Pour finir, ils diminuent le seuil de décision.

Ce modèle repose sur la détection des erreurs de prédiction. Il produit des prédictions concernant la scène visuelle, puis ces prédictions vont être comparées aux entrées sensorielles. Lorsqu'il n'y a pas d'écart, donc que les entrées sensorielles sont en accord avec les prédictions, la réponse visuelle est très faible, signalant seulement la validation des prédictions. A l'inverse, lorsqu'il y a un écart entre les entrées sensorielles et les prédictions, la réponse visuelle augmente proportionnellement à cet écart, telle une alarme, ce qui correspond à ce que Itti & Baldi (2009) ont appelé “surprise“.

Limites des modèles top-down : l'implémentation

Les modèles de l'allocation de l'attention permettant d'émettre des prédictions quantitatives sont historiquement l'apanage de ceux basés sur les processus bottom-up. Les processus top-down sont beaucoup plus complexes à implémenter du fait qu'ils dépendent des buts de l'observateur et des connaissances préalables, ils sont difficiles à quantifier et même à décrire de manière objective. En effet, afin de prendre en compte l'ensemble des éléments qui contribuent à l'attention top-down, il faudrait pouvoir modéliser toute l'expérience de la personne, tout ce qu'elle a stocké dans sa mémoire à long terme, mais également le cheminement de ses processus cognitifs qui aboutissent à une action telle que le regard.

Et même sans aller aussi loin, le simple fait de connaitre précisément le but suivi par la personne est un travail ardu, alors même que ce but lui a été donné par consigne. En effet, cela nécessite d'être certain de sa compréhension de la consigne, de son expérience à ce sujet, et de ses objectifs personnels. Par exemple, le seul fait d'être en train de passer une expérience, et a fortiori d'être en présence d'un expérimentateur ou de savoir que celui-ci enregistre son comportement, peut induire des objectifs supplémentaires qui peuvent influencer le comportement de cette personne (e.g. les travaux en psychologie sociale sur la soumission à l'autorité, dans la continuité des travaux de Milgram (1963)).

Les processus bottom-up, quant à eux, nécessitent principalement la prise en compte des caractéristiques de la scène visuelle et les caractéristiques de l'œil qui sont bien connues, identiques pour tous, et plus facilement quantifiables, même s'il reste de nombreuses choses à comprendre.

Un certain nombre de modèles ont été proposés, partant de modèles bottom-up, afin de profiter de leur capacité prédictive, en y ajoutant des modulations top-down, afin de se rapprocher de situation naturelle.

Borji & Itti (2013) ont proposé une revue des modèles prédisant l'allocation de l'attention, et ont recensé un grand nombre des modèles quantitatifs existants. Leur article montre le classement de ces modèles en fonction de différentes caractéristiques, et nous pouvons voir 52 modèles basés sur les processus bottom-up versus 10 basés sur les processus top-down, top-down étant ici généralisable aux effets de contexte et correspond à tout ce qui n'est pas directement associé à une entrée sensorielle. Ainsi sont classifiés les modèles suivants comme étant des modèles généraux de l'attention : les modèles de McCallum (1996) et de Jodogne & Piater (2007) fondés sur l'apprentissage par renforcement, plus précisément pour le second, sur la récursivité spatiale à partir de l'extraction de Feature ; le modèle de Rao et al. (2002) inspiré de la psychophysique et de la notion d'excentricité ; le modèle de Ramström & Christensen (2002) fondé sur la théorie des jeux ; le modèle de Navalpakkam & Itti (2005) qui ajoute un biais top-down afin d'augmenter la saillance d'un objet en fonction de ses caractéristiques visuelles ; le modèle de Paletta et al. (2005) qui permet la reconnaissance d'objet à partir d'information visuelle locale ; le modèle de Walker & Malik (2002) qui catégorise les objets à partir de la texture ; et le modèle de Borji et al. (2011) qui ajoute au modèle de Itti et al. (1998) une influence inhibitrice entre chaque étape de construction de la carte de saillance à l'aide de poids qui évoluent avec l'apprentissage. Deux problèmes se posent concernant ces huit modèles : i/ ils sont très fortement inspirés de modèles bottom-up de l'attention visuelle auxquels des influences top-down ont été ajoutées, ce qui limite par définition la possibilité de faire évoluer ces composantes top-down vers des modèles théoriques explicatifs de l'attention visuelle ; ii/ ils ne sont pas applicables à une situation dynamique. Les deux modèles restants sont les seuls modèles quantitatifs de l'attention visuelle applicables à une situation dynamique et avec une composante top-down d'après Borji & Itti (2013) : le modèle de Butko & Movellan (2009) et le modèle de Sprague & Ballard (2003). Le modèle de Butko & Movellan (2009) a pour objectif la détection d'objets. Il se fonde sur les travaux de Najemnik & Geisler (2005) qui ont proposé un modèle optimal du mouvement des yeux. Ce modèle n'est donc pas à proprement parler un modèle de l'attention visuelle top-down.

Ainsi, aucun n'est réellement satisfaisant à l'heure actuelle : Actuellement, aucun modèle du contrôle du regard ne peut rendre compte du contrôle du regard dans des environnements naturels et dynamiques. [...] la classe des modèles basés sur des propriétés du stimulus (i.e., les modèles de saillance) ne peut pas expliquer les patterns oculomoteurs que nous observons. Bien qu'il y ait eu plusieurs tentatives pour modéliser les effets top-down pour le contrôle du regard, de telles

tentatives prennent généralement la forme d'une pondération top-down d'une carte de saillance issue d'un filtrage bottom-up de l'image et ne s'attachent pas aux mécanismes qui déterminent si, et quand, un objet ou un emplacement est choisi par l'observateur comme une cible dans le cadre d'une séquence comportementale en cours.

1.4 Modèles à base d'apprentissage

L'adoption récente des méthodes d'apprentissage machine, approche permettant d'atteindre l'intelligence artificielle (IA), dans l'analyse d'images gagne rapidement du terrain dans de nombreux domaines de recherche. L'apprentissage profond fait partie d'un algorithme d'apprentissage machine basé sur un réseau neuronal artificiel. L'architecture de l'apprentissage profond a résolu avec succès des problèmes d'analyse complexes dans des applications d'imagerie médicale, de pathologie et de biologie

« Deep learning » correspond à des modèles d'IA (Intelligence Artificielle) très spécifiques qui utilisent l'apprentissage par couches successives. Ainsi, on entend parler de réseau de neurones convolutifs, de réseau convolutionnel, de DNN (pour « deep neural network ») ou encore de CNN (pour « convolutional neural network »).

Un modèle basé sur CNN utilise généralement un regroupement de couches convolutives, de mise en commun et entièrement connectées. Les régions les plus saillantes d'une image sont fournies par les neurones avec de grands champs récepteurs, tandis que les neurones avec de petits champs récepteurs génèrent des informations locales qui peuvent être exploitées pour affiner les cartes de saillance produites par les couches supérieures.

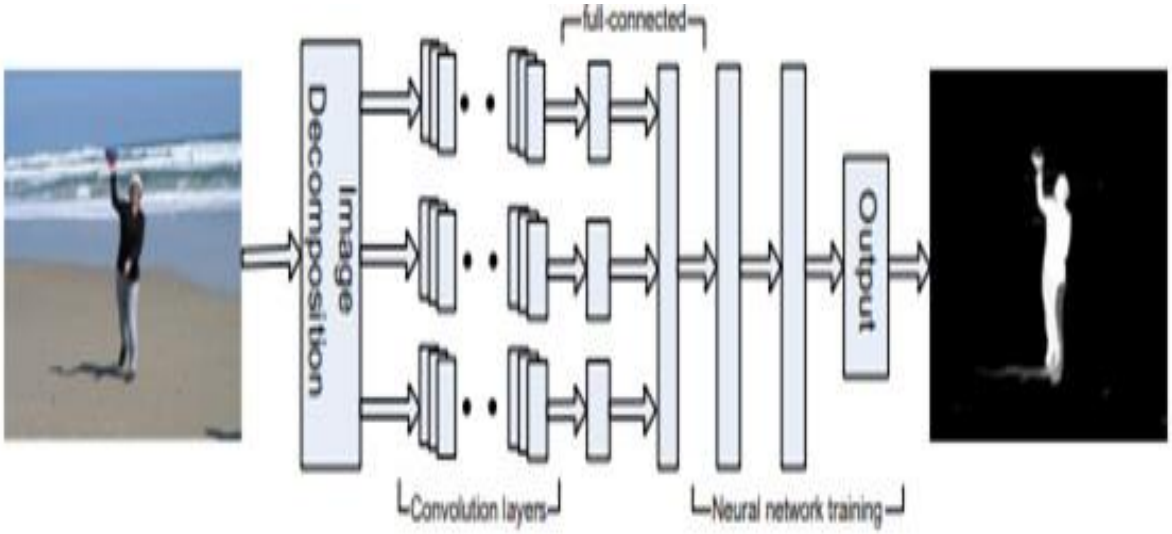

Figure 1.7: L'architecture globale de l'extraction de saillance visuelle à l'aide de CNN [09].

Les modèles de détection d'objets en saillie basés sur l'apprentissage en profondeur peuvent être divisés en deux catégories principales. La première catégorie comprend des modèles qui utilisent des perceptrons multicouches (MLP) pour prédire le score de saillance des caractéristiques profondes extraites de chaque unité d'imagerie. La deuxième catégorie comprend des modèles basés sur des réseaux entièrement convolutifs (basés sur FCN). Les modèles utilisent en général un algorithme d'encodeur automatique basé sur un réseau de neurones convolutifs (CNN) pré-entraîné sur une tâche de classification d'images à grande échelle, et combinent les représentations résultantes avec des informations de scène globales. [09]

1.5 Conclusion

L'attention visuelle est nécessaire car la quantité d'informations nous arrivons est trop grande pour être traitée dans sa globalité en effet l'énergie disponible pour le traitement d'informations est insuffisante. L'attention sert donc à sélectionner ce qui semble important de traiter en priorité

Dans ce chapitre, nous avons défini la perception visuelle puis nous avons discuté des avancées récentes dans la modélisation de l'attention visuelle en mettant l'accent sur les modèles de saillance ascendants, les progrès dans ce domaine pourraient grandement aider à résoudre d'autres problèmes de vision difficiles tels que l'interprétation de scènes encombrées et la reconnaissance d'objets.et la fusion des images multi-focus, le deep learning a été introduit dans le domaine de l'attention visuelle et diverses méthodes ont été proposées.

Chapitre 2

Fusion d'images

2.1 Introduction

Ces dernières années, la fusion d'images a été utilisée dans un large éventail d'applications, notamment la détection, la surveillance, les diagnostics médicaux et les applications photographiques. Le but de la fusion est de créer une nouvelle image qui conserve certaines des informations contenues dans chacune des photos originales.

Un bon processus de fusion devrait inclure l'information redondante mais en même temps ne devrait pas surcharger l'image fusionnée. Il ne doit pas non plus introduire d'artefacts ou de bruit dans l'image. Avec des avancées technologiques rapides, il est désormais possible d'obtenir des informations provenant d'images multi-sources pour produire une image fusionnée de haute qualité avec des informations spatiales et spectrales.

2.2 Les bases de les images

2.2.1 Définition d'une image

Une image est une représentation planaire d'une scène ou d'un objet situé en général dans un espace tridimensionnel, elle est issue du contact des rayons lumineux provenant des objets formants la scène avec un capteur (caméra, scanner, rayons X, ...). Il ne s'agit en réalité que d'une représentation spatiale de la lumière.

L'image est considérée comme un ensemble de points auquel est affectée une grandeur physique (luminance, couleur). Ces grandeurs peuvent être continues (image analogique) ou bien discrètes (images digitales). Mathématiquement, l'image représente une fonction continue IF, appelée fonction image, de deux variables spatiales représentée par IF(x, y) mesurant la nuance du niveau de gris de l'image aux coordonnées (x, y). [10]

2.2.2 Image numérique

L'image numérique est constituée d'un ensemble de points appelés pixels. Un pixel (abréviation de **PIC**ture **El**ément) est défini comme le plus petit élément constitutif d'une image numérique matricielle. Pour une image à deux tons, noir et blanc, le pixel peut être codé par un seul bit codant (0 pour noir ou 1 pour blanc). Pour des images en nuances de gris ou en couleurs le pixel peut être codé par 2, 4, 8, 16, 24 ou 32 bits.[11]

La numérisation d'une image est la conversion de celle-ci de son état analogique en une image numérique représentée par une matrice bidimensionnelle de valeurs numériques f(x,y), comme la montre la figure 2.1 où : x,y: coordonnées cartésiennes d'un point de l'image. F(x,y) : niveau d'intensité

2.2.3 les types de format d'image

2.2.3.1 Image couleur RVB : Une image couleur est en réalité composée de trois images, afin de représenter le rouge, le vert, et le bleu. Chacune de ces trois images s'appelle un canal. Cette représentation en rouge, vert et bleu mime le fonctionnement du système visuel humain.

L'œil humain analyse la couleur à l'aide de trois types de cellules photo 'les cônes'. Ces cellules sont sensibles aux basses, moyennes, ou hautes fréquences (rouge, vert, bleu). Pour représenter la couleur d'un pixel, il faut donc donner trois nombres, qui correspondent au dosage de trois couleurs de base.[10]

2.2.3.2 Images à niveaux de gris (monochromes): Le niveau de gris est la valeur de l'intensité lumineuse en un point. La couleur du pixel peut prendre des valeurs allant du noir au blanc en passant par un nombre fini de niveaux intermédiaires. Donc pour représenter les images à niveaux de gris, on peut attribuer à chaque pixel de l'image une valeur correspondant à la quantité de lumière renvoyée. Cette valeur peut être comprise par exemple entre 0 et 255. Chaque pixel n'est donc plus représenté par 1 bit, mais par 1 octet. Pour cela, il faut que le matériel utilisé pour afficher l'image, soit capable de produire les différents niveaux de gris correspondant.[12]

2.2.3.3 Image binaire : Une image binaire est une matrice rectangulaire dans l'élément valent 0 ou 1. Lorsque l'on visualise une telle image, les 0 sont représentés par du noir et les 1 par du blanc. [10]

2.2.4 Caractéristiques de l'image

L'image est un ensemble structuré d'information caractérisé par les paramètres suivants :

2.2.4.1 Pixel

Contraction de l'expression anglaise " picture eléments ": éléments d'image, le pixel est le plus petit point de l'image, c'est une valeur numérique représentative des intensités lumineuses. Si le bit est la plus petite unité d'information que peut traiter un ordinateur, le pixel est le plus petit élément que peuvent manipuler les matériels et logiciels sur l'image. La lettre a, par exemple, peut être affichée comme un groupe de pixels dans la figure ci-dessous.[8]

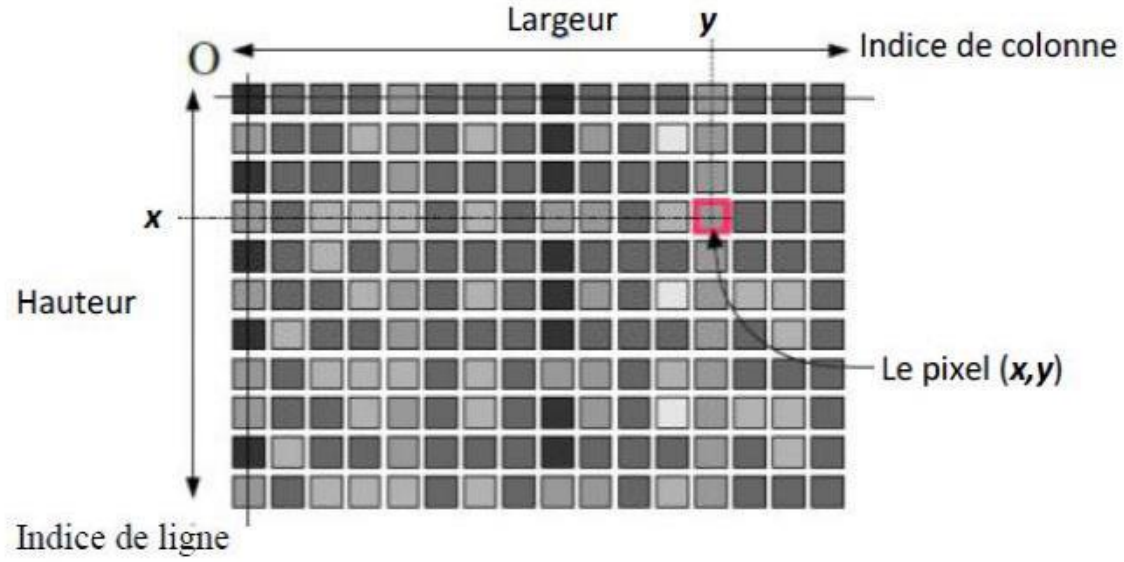

Figure 2.1 : représentation le pixel in l'image numérique.[8]

2.2.4.2 Dimension & Résolution

La dimension est la taille de l'image. Elle se présente sous forme d'une matrice dont les éléments sont des valeurs numériques représentatives des intensités lumineuses (pixels). Le nombre de lignes de cette matrice multiplié par le nombre de colonnes nous donne le nombre total de pixels dans une image.

Par contre, la résolution est la clarté ou la finesse de détails atteinte par un moniteur ou une imprimante dans la production d'images. Sur les moniteurs d'ordinateur, la résolution est exprimée en nombre de pixels par unité de mesure (pouce ou centimètre). On utilise aussi le mot résolution pour désigner le nombre total de pixels horizontaux et verticaux sur un moniteur. Plus ce nombre est grand, plus la résolution est meilleure.[10]

2.2.4.3 Contours et textures

Une texture est une région dans une image numérique qui a des caractéristiques homogènes. Ces caractéristiques sont par exemple un motif basique qui se répète .la texture est composée de Texel, l'équivalent des pixels.[13]

Les contours représentent la frontière entre les objets de l'image, ou la limite entre deux pixels dont les niveaux de gris représentent une différence significative.

2.2.4.4 La taille d'une image

Pour connaître la taille d'une image, il est nécessaire de compter le nombre de pixels que contient l'image, cela revient à calculer le nombre des cases du tableau, soit la hauteur de celui-ci que multiplie sa largeur. La taille de l'image est alors le nombre des pixels que multiplie la taille (en octet) de chacun de ces éléments.[13]

2.2.4.5 Luminance

C'est le degré de luminosité des points de l'image. Elle est définie aussi comme étant le quotient de l'intensité lumineuse d'une surface par l'aire apparente de cette surface, pour un observateur lointain, le mot luminance est substitué au mot brillance, qui correspond à l'éclat d'un objet. [14]

2.2.4.6 Bruit

Un bruit (parasite) dans une image est considéré comme un phénomène de brusque variation de l'intensité d'un pixel par rapport à ses voisins, il provient de l'éclairage des dispositifs optiques et électroniques du capteur.[10]

2.2.4.7 Histogramme

L'histogramme des niveaux de gris ou des couleurs d'une image est une fonction qui donne la fréquence d'apparition de chaque niveau de gris (couleur) dans l'image. Pour diminuer l'erreur de quantification, pour comparer deux images obtenues sous des éclairages différents, ou encore pour mesurer certaines propriétés sur une image.[10]

2.2.5 La profondeur de champ

La profondeur de champ est la zone de l'image dans laquelle les objets paraissent nets à l'œil nu. Cette profondeur de champ se répartit à l'avant et à l'arrière du plan où est faite la mise au point (le plan focal). Dans une photo à faible profondeur de champ, seule une petite zone de l'image est nette (ex: portraits avec arrière-plan flou), tandis que dans une photo à grande profondeur de champ, la majeure partie de l'image est nette (ex : paysages).

2.3 Fusion des données :

La fusion de données est analogue aux capacités cognitives de l'être humain à intégrer divers stimuli à partir de différents sens (vue, ouïe, odeur, goût, toucher) pour l'inférence de connaissances sur le monde externe. Il est la discipline qui cherche à combiner des informations obtenues de différents systèmes dans le but d'effectuer des inférences à partir de ces observations.

La fusion de données est un domaine en émergence avec des applications variées. Historiquement, la plupart des techniques de fusion de données ont été développées pour des applications militaires : la surveillance des océans, la défense aérienne, le renseignement sur les champs de bataille et l'identification de cibles. La technologie est maintenant largement utilisée dans des applications non militaires telles que l'acquisition de données, le contrôle automatisé, la surveillance d'équipements complexes et la robotique.

La fusion peut se pratiquer à trois moments différents du travail (Fig. 2). Soit au niveau du pixel (fusion des pixels), soit au niveau des caractéristiques après une segmentation (fusion des objets extraits de l'image), soit au niveau décisionnel lors de la phase finale de la segmentation (fusion des objets extraits et identifiés) [Pohl et Van Genderen, 1998].

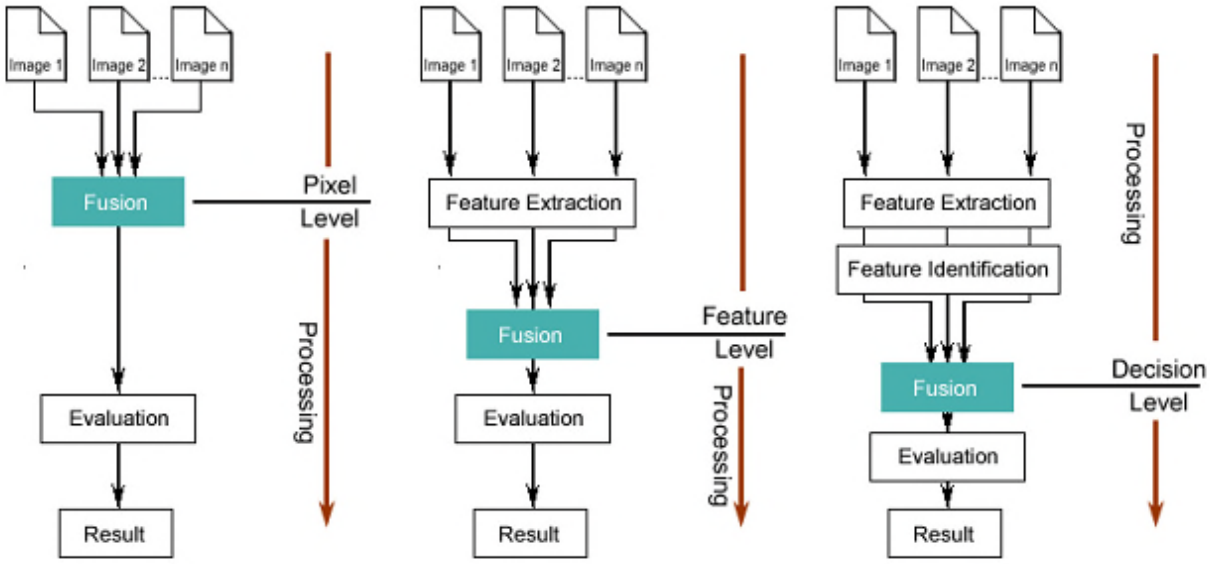

Figure. 2.2 : Niveaux de traitements de la fusion d'images [Pohl et Van Genderen, 1998]

Pour effectuer la fusion au niveau du pixel, les capteurs doivent être identiques (par ex., plusieurs caméras infrarouges) ou commensurables (par ex., images infrarouges et images radar).

Pour la fusion au niveau des caractéristiques, un vecteur d'attributs est extrait à partir de la sortie de chaque capteur. Ces vecteurs d'attributs sont ensuite combinés (fusionnés) et une déclaration d'identification est ensuite effectuée sur la base de ce vecteur conjoint. Les outils utilisés pour la

déclaration d'identification comprennent les techniques statistiques (par ex., analyse de regroupements), les réseaux de neurones, les techniques structurelles et à base de connaissances.

Dans la fusion au niveau des décisions (ou niveau des déclarations), chaque capteur effectue de façon indépendante un estimé ou une déclaration de la scène observée. Ces estimations sont ensuite combinées via un procédé de fusion.

Dans notre cas, la fusion aura pour but d'affiner la précision spatiale de l'image multi spectrale, de préciser et de faciliter la photo-interprétation des images, il est donc plus approprié de la réaliser au niveau du pixel.

2.4 fusion d'image

On définit généralement la fusion d'images comme la combinaison de deux ou de plusieurs images différentes pour former à l'aide d'un algorithme une nouvelle image [Pohl et Van Genderen, 1998].

Le processus de fusion d'images est défini comme la collecte de toutes les informations importantes à partir de plusieurs images et leur inclusion dans moins d'images, généralement une seule. Cette image unique est plus informative et précise que n'importe quelle image, et elle comprend toutes les informations nécessaire. Le but de la fusion d'images n'est pas seulement de réduire la quantité de données mais aussi de construire des images plus appropriées et compréhensibles pour la perception humaine et machine.

Les objectifs les plus communs de la fusion d'images sont :

- Netteté de l'image
- Amélioration de la précision radiométrique
- Création d'ensembles stéréo de données
- Augmentation des caractéristiques
- Amélioration de la classification
- Détection de changement dans le temps
- Franchissement des écarts

2.5 Les enjeux de la qualité des images fusionnées

Les logiciels commerciaux proposent de nombreuses méthodes pour la fusion d'images et il n'est pas évident pour des non-spécialistes de sélectionner une méthode plutôt qu'une autre pour un cas donné. L'utilisateur veut naturellement produire une image de bonne qualité. Cependant, la notion de qualité des produits de fusion est difficile à appréhender puisque chaque acteur possède sa propre vision de la définition d'une image fusionnée de bonne qualité.

La connaissance des points forts et points faibles d'une méthode ne peut être fiable que si un cadre formel d'évaluation de la qualité est établi et respecté. De la même manière, ce cadre est nécessaire une fois le produit de fusion synthétisé. Les utilisateurs doivent avoir les moyens de juger la qualité effectivement obtenue par une liste limitée d'indices pertinents. Si cette liste est trop longue, ils risquent de s'égarer dans leur démarche d'estimation de la qualité. Pour le moment, seuls quelques auteurs ont entrepris la démarche d'établir un protocole normalisé d'évaluation de la qualité.[9]

2.6 Applications de la fusion d'images :

Le champ d'application de la fusion d'images est potentiellement vaste. Dans le contexte du rehaussement d'images, elle a pour tâche principale de réunir au sein d'une image, l'information provenant de différentes sources imagées (images multi sources). On retrouve parmi les domaines d'applications:

2.6.1 Photos prises en dehors du focal (out-of-focus) :

En raison de la faible profondeur focale des lentilles optiques (particulièrement ceux avec de longues profondeurs focales) il n'est souvent pas possible d'obtenir une image qui contient tous les objets appropriés. Une possibilité pour surmonter ce problème est de prendre plusieurs photos avec différents points focales et de les fusionner ensemble dans une seule image qui finalement contient les régions focalisées de toutes les images d'entrée.

2.6.2 L'aide à la navigation :

Permettre aux pilotes par exemple de voler dans de mauvaises conditions de visibilité (telles que le brouillard ou la forte pluie). Les hélicoptères sont équipés de plusieurs capteurs d'imagerie, qui peuvent être consultés par le pilote. Une suite typique de capteurs inclut un capteur de lumière basse et un capteur infrarouge pour les images thermiques. Dans la configuration actuelle, le pilote peut choisir un des deux capteurs à observer dans son affichage. Une amélioration possible est de combiner les deux sources d'images dans une seule image fusionnée qui contient l'information appropriée des deux dispositifs imageurs.

2.6.3 L'imagerie Médical :

Avec le développement de nouvelles méthodes d'imagerie dans le diagnostic médical, se pose la nécessité d'une véritable combinaison de tous les ensembles de données d'images disponibles. Exemples de dispositifs d'imagerie, la tomographie par ordinateur (CT) ou l'imagerie par résonance magnétique (IRM).

2.6.4 télédétection :

La télédétection est une application typique de la fusion d'image : Les modules à balayage spectraux modernes recueillent jusqu'à plusieurs centaines de bandes spectrales qui peuvent être visualisées et traitées individuellement, ou qui peuvent être fusionnées en une seule image, en fonction de la tâche d'analyse d'image.

2.7 Catégories de la fusion d'images

Les méthodes de fusion d'images classent en fonction des données d'entrée Fusionner, selon le but de la fusion. Les images d'entrée peuvent provenir de l'une des catégories suivantes :

2.7.1 Images multimodales :

La fusion multimodale d'images est appliquée à des images provenant de différentes modalités comme le visible et l'infrarouge, la tomodensitométrie et la RMN, ou les images satellitaires panchromatiques et multi spectrales. Le but du système de fusion d'image multimodal est de diminuer la quantité de données et de mettre l'accent sur les informations spécifiques à la bande.[15]

2.7.2 Images multi-focus :

Dans les applications d'appareils photo numériques, lorsqu'un objectif se concentre sur un sujet à une certaine distance, tous les sujets situés à cette distance ne sont pas très nets. Un moyen possible de résoudre ce problème est la fusion d'images, dans laquelle on peut acquérir une série d'images avec différents réglages de mise au point et les fusionner pour produire une image unique avec une profondeur de champ étendue. Le but de ce type de fusion est d'obtenir une seule image tout en focus.

2.7.3 Images multi-vues :

Dans la fusion d'images multi-vues, un ensemble d'images de la même scène est pris par le même capteur mais à partir de différents points de vue ou plusieurs ions d'acquisition 3D du même spécimen sont fusionnés pour obtenir une image avec une résolution plus élevée. Le but de ce type de la fusion est de fournir des informations complémentaires à partir de différents points de vue.

2.7.4 Images multi-temporelles :

Dans la fusion d'images multi-temporelles, les images prises à différents moments (secondes à années) afin de détecter les changements entre elles sont fusionnées pour obtenir une seule image. .[15]

2.8 méthodes de fusion d'images multi-focus

De nombreuses méthodes de fusion d'images ont été proposées dans ces dernières années. En fonction de leur représentation, ils sont classés en deux groupes principaux : domaine spatial et domaine de transformation.

2.8.1 Méthodes du domaine spatial

La technique de fusion du domaine spatial utilise des caractéristiques spatiales locales telles que le gradient, la fréquence spatiale et l'écart type local. Les valeurs des pixels de deux images ou plus sont rassemblées et manipulées pour obtenir les résultats souhaités par la technique de fusion de domaine spatial.

Dans cette catégorie de méthodes, les images sources sont fusionnées dans le domaine spatial, c'est-à-dire en utilisant certaines caractéristiques spatiales des images. Par rapport aux méthodes de domaine de transformation, la caractéristique la plus importante des méthodes de domaine spatial est qu'elles ne contiennent pas l'étape de transformation inverse pour reconstruire l'image fusionnée. Selon la manière de traitement de pixels adoptée, les méthodes de domaine spatial peuvent être regroupées en méthodes suivantes. [16]

2.8.1.1 Méthodes basées pixel

Les méthodes du domaine spatial qui fonctionne au niveau des pixels traite directement de la position des pixels de l'image d'entrée. Ce sont les valeurs des pixels qui sont manipulées directement pour obtenir le résultat souhaité.

Dans ces méthodes, une mesure du niveau d'activité, également connue sous le nom de mesure de mise au point dans la fusion d'images multi-focus, est d'abord appliquée pour évaluer la saillance des pixels dans les images sources. Ensuite, les mesures de mise au point obtenues à partir de différentes images sources sont comparées pour générer une carte de poids au niveau des pixels.

Les méthodes de fusion de domaine spatial basées sur les pixels peuvent être classées sous trois aspects : la mesure du niveau d'activité, la règle de fusion et le raffinement de la carte poids/décision [15]:

Les mesures conventionnelles du niveau d'activité telles que la variance, SF (spatial frequency), EOG (energy of gradient), EOL(energy of Laplacian), SML (sum-modified-laplacian) sont également fréquemment utilisées dans les méthodes de domaine spatial basées sur les pixels.

Pour la règle de fusion, la sélection maximale et la moyenne pondérée restent les règles les plus largement utilisées dans les méthodes de fusion basées sur les pixels.

L'hypothèse de base de ces méthodes est que les régions avec des attributs différents doivent être fusionnées par des règles différentes. Une méthode courante consiste à diviser les images sources en

régions focalisées, en régions dé focalisées et en régions limites, et la fusion des régions limites nécessite généralement des schémas plus complexes pour améliorer la qualité visuelle des images fusionnées.

Il existe également des méthodes basées sur les pixels qui convertissent la tâche d'estimation de la carte de poids en résolution d'un problème d'optimisation, telles que la méthode basée sur un modèle variation , les méthodes basées sur les marches aléatoires (RW), le la méthode basée sur le champ aléatoire conditionnel (CRF), la méthode basée sur le modèle multi-matting, etc.

Les approches de raffinement de la carte poids/décision sont basées sur des techniques de filtrage d'images, qui incluent le filtrage morphologique et le filtrage statistique. L'objectif étant d'éliminer les régions isolées susceptibles d'être mal classées dans la carte de décision initiale, tandis que les filtres préservant les contours tels que le filtre guidé et le filtre bilatéral sont principalement conçus pour rendre les poids en régions frontalières plus lisses et naturelles. Il existe d'autres approches de raffinement de carte de poids/décision qui incluent celles basées sur le champ aléatoire de Markov (MRF), celles basées sur RW, celles basées sur la coupe normalisée, celles basées sur l'appariement de caractéristiques locales, basé sur une coupe de graphe, basé sur un modèle de contour actif, etc.

2.8.1.2 Méthodes basées blocs

En 2001, Li et al. ont introduit une méthode de fusion d'images multi-focus dans le domaine spatial basée sur un schéma de division de blocs, dans lequel chaque image source est divisée en un certain nombre de blocs de taille fixe. La fréquence spatiale est utilisée comme mesure du niveau d'activité de chaque bloc et une règle de fusion adaptative basée sur un seuil est utilisée pour obtenir le bloc fusionné. L'image fusionnée est finalement construite après l'application d'une approche de vérification de cohérence basée sur un filtrage majeur [16]

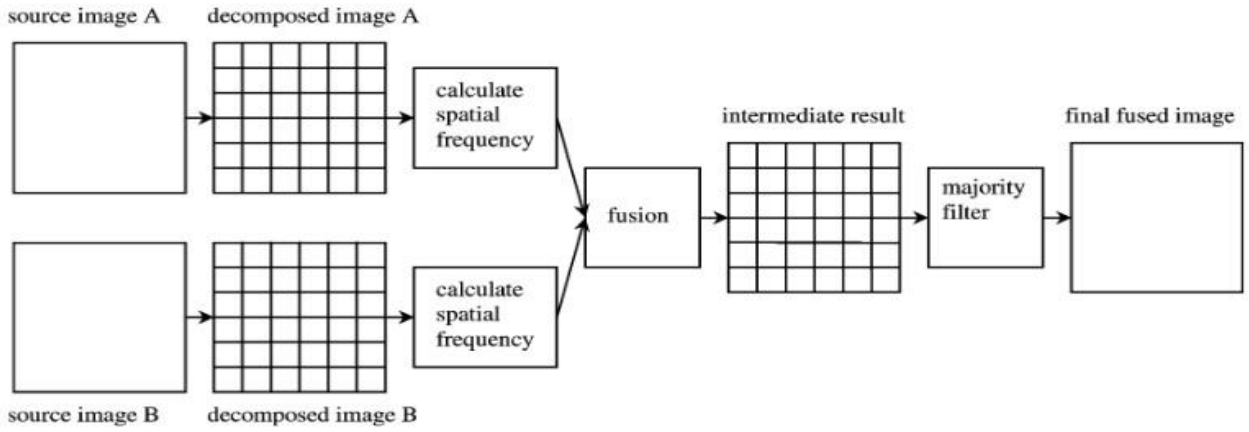

Figure 2.3: Schéma De Principe Pour La Fusion D'images Multi-focus: Méthode De Li Et Al Basée Blocs [16]

Depuis lors, les méthodes basées sur les blocs ont émergé comme une direction active dans la fusion d'images multi-focus et diverses améliorations ont été apportées à la mesure du niveau d'activité, à la règle de fusion, à la stratégie de division des blocs, etc.

Nous allons, dans ce qui suit donné un aperçu de quelques techniques :

Huang et Jing ont présenté une évaluation d'un ensemble de mesures de mise au point fréquemment utilisées dans la fusion d'images multi-focus, notamment la variance, l'énergie de gradient (EOG), l'énergie du Laplacian (EOL), le Laplacian à Somme Modifiée (SML), la fréquence spatiale. (SF), etc. Ils ont également proposé une mesure du niveau d'activité combiné EOL et PCNN pour la fusion basée sur des blocs, conduisant à un schéma populaire (c'est-à-dire combinant une mesure de mise au point traditionnelle avec un modèle PCNN) dans la fusion d'images multi-focus. Ils ont aussi présenté un schéma générique pour les techniques de fusion d'images multi-focus basée sur la sélection de blocs [17].

Zhan et al. ont proposé une mesure de mise au point basée sur la congruence de phase (PC) pour fusionner des images multi-focus. Les mesures du niveau d'activité basées sur la DCT ont également été appliquées à la fusion d'images multi-focus par blocs. En plus de développer des mesures d'activité plus efficaces, certains chercheurs ont tenté d'appliquer des mesures d'activité multiples pour remplacer la manière ci-dessus basée sur une mesure unique et concevoir des modèles de classification correspondants comme règle de fusion pour combiner les blocs d'images sources.

Kausar et Majid ont introduit la forêt aléatoire comme classificateur pour la détermination de la propriété de mise au point sur la base de neuf caractéristiques locales couramment utilisées telles que la visibilité, la SF, la variance, l'EOG, les caractéristiques basées sur DWT et les caractéristiques basées sur DCT. Un schéma de classification basé sur le vote majoritaire a également été adopté dans cette catégorie de méthodes de fusion. Toutes les méthodes de fusion par blocs mentionnées ci-dessus sont basées sur une taille de bloc fixe qui est définie empiriquement. De toute évidence, la taille du bloc a un impact crucial sur les résultats de fusion finaux. La manière basée sur une taille fixe est très susceptible d'introduire des effets de blocage indésirables dans l'image fusionnée. Pour résoudre ce problème, des stratégies améliorées de division de blocs ont été proposées par les chercheurs.[8]

2.8.2 Domaine de transformation

les images d'entrée sont décomposées sur la base de coefficients de transformation. Après cela, la technique de fusion est appliquée et la carte de décision de fusion est obtenue. Enfin, la transformation inverse s'applique à cette carte de décision qui produit une image fusionnée.

Les méthodes du domaine de transformation se déroulent en trois étapes principales comme illustré sur la figure 2.4 Dans la première étape les images sources sont converties en un domaine de transformation en appliquant une approche de décomposition/représentation d'images.

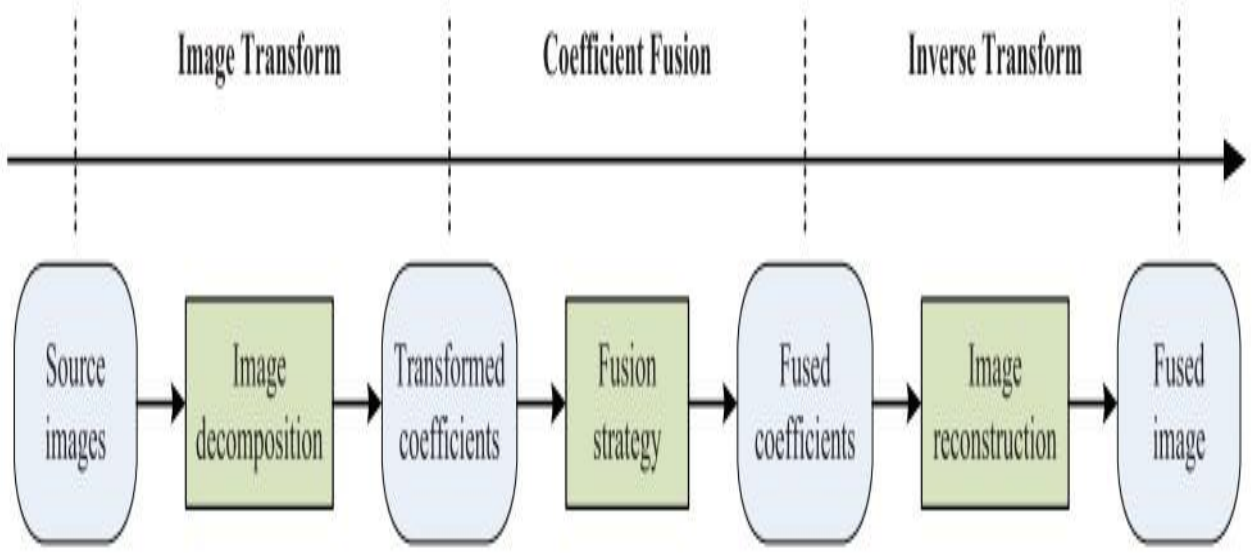

Figure 2.4: Schéma général des méthodes de domaine de transformation.[09]

2.8.2.1 TCD (transformée en cosinus discrète)

Dans cette méthode, les images d'entrée sont divisées en blocs non chevauchants ayant la taille NxN. Les coefficients DCT sont calculés pour chaque bloc et des règles de fusion sont appliquées pour obtenir des coefficients DCT fusionnés. Enfin, IDCT a appliqué les coefficients fusionnés pour produire l'image fusionnée finale.[18]

2.8.2.2 Fusion Transformée de Contourlet

« Contourlet Transform » apporte la douceur dans une image fusionnée avec deux modalités différentes d'images. Cette transformation basée sur une région est mise en œuvre en deux étapes. Dans la première étape, un schéma de banque à double filtre est appliqué pour la transformation et dans l'étape suivante, la décomposition est effectuée avec des règles de fusion. Enfin, l'image fusionnée est récupérée en utilisant la procédure de reconstruction.[19]

2.8.2.3 Transformée en ondelettes stationnaire

La transformée en ondelettes stationnaire (SWT) est similaire à la transformée en ondelettes discrètes (DWT), mais le seul processus de sous-échantillonnage est supprimé, ce qui signifie que la SWT est invariante par translation.

2.8.2.4 Méthode pyramidale

La méthode pyramide consiste en un ensemble de copie passe-bas ou passe-bande d'une image. Chaque copie d'une image représente les informations de modèle d'une échelle différente. Cependant, dans la méthode pyramidale, chaque niveau est un facteur de deux plus petit que son prédécesseur, et un niveau supérieur se concentre sur les fréquences partielles inférieures. Cette pyramide ne contient pas toutes les informations concernant la reconstruction de l'image originale [20] .

2.9 Conclusion

Dans ce chapitre, nous illustrons les différentes méthodes d'application de la fusion d'images. L'objectif de notre travail était de comparer des méthodes de fusion basées sur l'analyse multi résolution , pour obtenir des images de synthèse à haute résolution spatiale. En effet, la fusion au niveau pixel des images panchromatique et multi spectrales a permis de synthétiser des images de la même résolution spatiale que l'image panchromatique tout en conservant le contenu spectral des images multi spectrales originales.

Chapitre 3

Les réseaux de neurones convolutifs (CNN)

3.1 Introduction

Depuis quelques années, un nouveau lexique lié à l'émergence de l'intelligence artificielle dans notre société inonde les articles scientifiques, et il est parfois difficile de comprendre de quoi il s'agit. Lorsqu'on parle d'intelligence artificielle, on fait très souvent l'allusion aux technologies associées comme le Machine learning ou le Deep learning. Deux termes extrêmement utilisés avec des applications toujours plus nombreuses.

Dans ce chapitre nous allons présenter L'apprentissage automatique et les notions fondamentales de l'apprentissage en profondeur, Il existe un grand nombre de variables d'architectures profondes. La plupart d'entre eux sont dérivés de certaines architectures originales. Nous allons choisir les réseaux de neurones convolutif (CNN).

3.2 L'intelligence artificielle (IA)

L'intelligence artificielle, souvent appelée « IA », est déjà tout autour de nous : dans notre téléphone, dans le moteur de recherche de notre ordinateur, dans les voitures et dans les maisons.

L'Intelligence Artificielle, branche de l'Informatique fondamentale s'est développée avec pour objectif la simulation des comportements du cerveau humain. Les premières tentatives de modélisation du cerveau sont anciennes et précèdent même l'informatique.

Ces dernières années, l'intelligence artificielle a beaucoup progressé. Elle est même devenue experte dans certains domaines. Elle est maintenant capable d'exécuter des tâches sophistiquées grâce aux instructions que les ingénieurs ont programmées. Ces instructions sont appelées des algorithmes.

L'intelligence artificielle a plusieurs buts, parmi lesquels l'apprentissage, le raisonnement et la perception. Elle est utilisée dans toutes les industries, à tel point que les applications sont infinies et impossibles à énumérer de façon exhaustive.

- Dans le domaine de la santé, elle est utilisée pour développer des traitements personnalisés, découvrir de nouveaux médicaments, ou encore pour analyser les imageries médicales telles que les rayons X et les IRM.
- Le secteur du commerce de détail utilise l'IA pour proposer des recommandations et des publicités personnalisées aux clients. Elle permet aussi d'optimiser la disposition des produits ou de mieux gérer les inventaires.
- Dans les usines, l'intelligence artificielle analyse les données des équipements IT pour prédire la charge et la demande grâce au Deep Learning. Elle permet aussi d'anticiper un éventuel dysfonctionnement pour intervenir de manière précoce.[21]

3.3 Apprentissage automatique

L'apprentissage automatique (en anglais : machine learning) est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes.

Les algorithmes sont les moteurs du machine learning. En général, deux principaux types d'algorithmes de machine learning sont utilisés aujourd'hui : l'apprentissage supervisé et l'apprentissage non supervisé.[22]

3.3.1 l'apprentissage supervisé

Les algorithmes de machine learning supervisé sont les plus couramment utilisés. Avec ce modèle, un data scientiste sert de guide et enseigne à l'algorithme les conclusions qu'il doit tirer. Tout comme un enfant apprend à identifier les fruits en les mémorisant dans un imagier, en apprentissage supervisé, l'algorithme apprend grâce à un jeu de données déjà étiqueté et dont le résultat est prédéfini.[22]

Comme exemples de machine learning supervisé, on peut citer des algorithmes tels que la régression linéaire et logistique, la classification en plusieurs catégories et les machines à vecteurs de support.

3.3.2 l'apprentissage non supervisé

Le machine learning non supervisé utilise une approche plus indépendante dans laquelle un ordinateur apprend à identifier des processus et des schémas complexes sans un quelconque guidage humain constant et rigoureux. Le machine learning non supervisé implique une formation basée sur des données sans étiquette ni résultat spécifique défini.

Pour continuer avec l'analogie de l'enseignement scolaire, le machine learning non supervisé s'apparente à un enfant qui apprend à identifier un fruit en observant des couleurs et des motifs, plutôt qu'en mémorisant les noms avec l'aide d'un enseignant. L'enfant cherche des similitudes entre les images et les sépare en groupes, en attribuant à chaque groupe sa propre étiquette. Comme exemples d'algorithmes de machine learning non supervisé, on peut citer la mise en cluster de k-moyennes, l'analyse de composants principaux et indépendants, et les règles d'association.[22]

3.4 Apprentissage profond

apprentissage profond (en anglais : deep learning) est un type d'intelligence artificielle dérivé du machine learning (apprentissage automatique) où la machine est capable d'apprendre par elle-même, contrairement à la programmation où elle se contente d'exécuter à la lettre des règles prédéterminées.

une technique de machine learning reposant sur le modèle des réseaux neurones: des dizaines voire des centaines de couches de neurones sont empilées pour apporter une plus grande complexité à l'établissement des règles.[23]

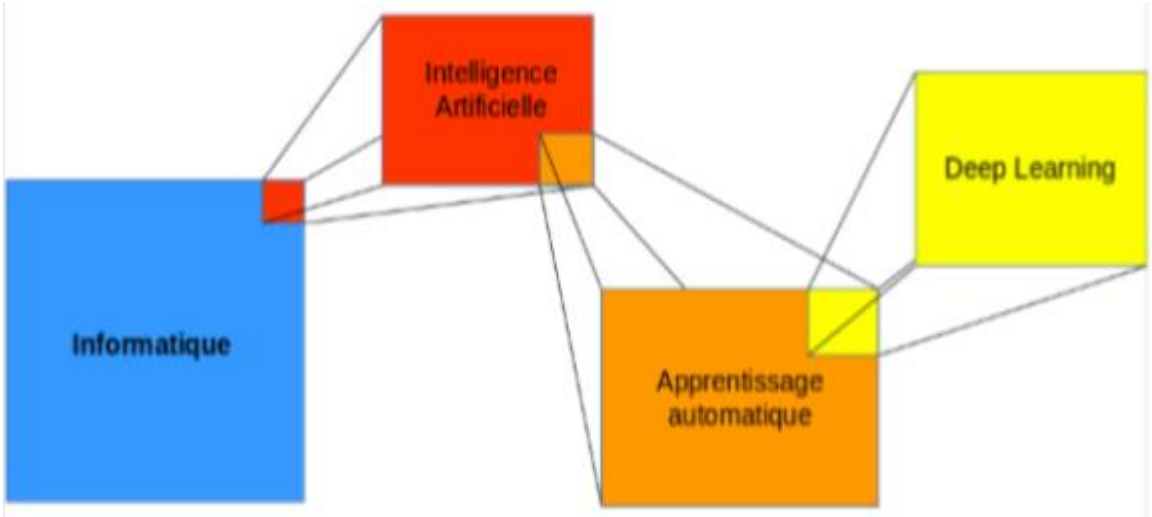

Figure 3.1: Schéma résumant la place du deep learning dans le monde de l'informatique [23]

3.4.1 Où se situe le deep learning dans le monde de l'informatique?

Lorsque l'on parle d'informatique, on regroupe beaucoup de domaines d'étude différents. Pour ne pas se mélanger les pinceaux, nous allons situer la place du deep learning (apprentissage profond en français), dans le domaine de l'informatique. Le deep learning est un sous domaine de l'apprentissage automatique.

L'apprentissage automatique c'est l'art de programmer un ordinateur afin qu'il soit capable d'apprendre de façon autonome et à partir d'exemple.

Le deep learning a permis d'obtenir des résultats impressionnants dans des domaines aussi nombreux que variés:

- Reconnaissance d'image, de texte, de voix, de visage...
- La segmentation dans le domaine médicale, la compréhension d'une scène, d'un texte...
- Génération d'image, de texte, de voix, d'œuvre d'art, de visage humain...
- Voiture autonome, robot autonome...
- Surveillance routière, piétonne...

3.4.2 Principes de l'apprentissage profond

L'apprentissage profond est un paradigme d'apprentissage automatique inspiré de l'anatomie du cerveau humain. Cet apprentissage est associé à une structure algorithmique que l'on appelle un réseau de neurones. Le neurone biologique est une cellule complexe, mais seul son fonctionnement basique a servi d'inspiration au neurone informatique. Le parallèle entre neurone biologique et neurone informatique est illustré figure 3.2.[23]

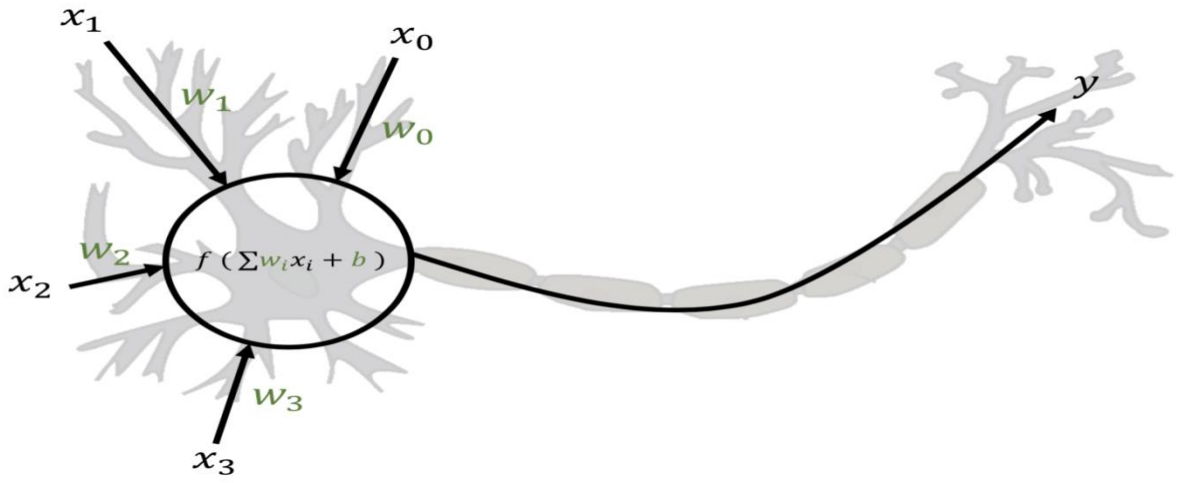

Figure 3.2 : Schéma d'un neurone informatique superpose a un schéma de neurone biologique.[23]

3.4.2.1 Neurone Biologique

Le système nerveux est composé de milliards de cellules : c'est un réseau de neurones biologiques. En effet, les neurones ne sont pas indépendants les uns des autres, ils établissent entre eux des liaisons et forment des réseaux plus ou moins complexes [23].

Le neurone biologique est composé de trois parties principales :

- **Le corps cellulaire** composé du centre de contrôle traitant les informations reçues par les dendrites.
- **Les dendrites** sont les principaux fils conducteurs par lesquels transitent l'information venue de l'extérieur.
- **L'axone** est fil conducteur qui conduit le signal de sortie du corps cellulaire vers d'autres neurones.

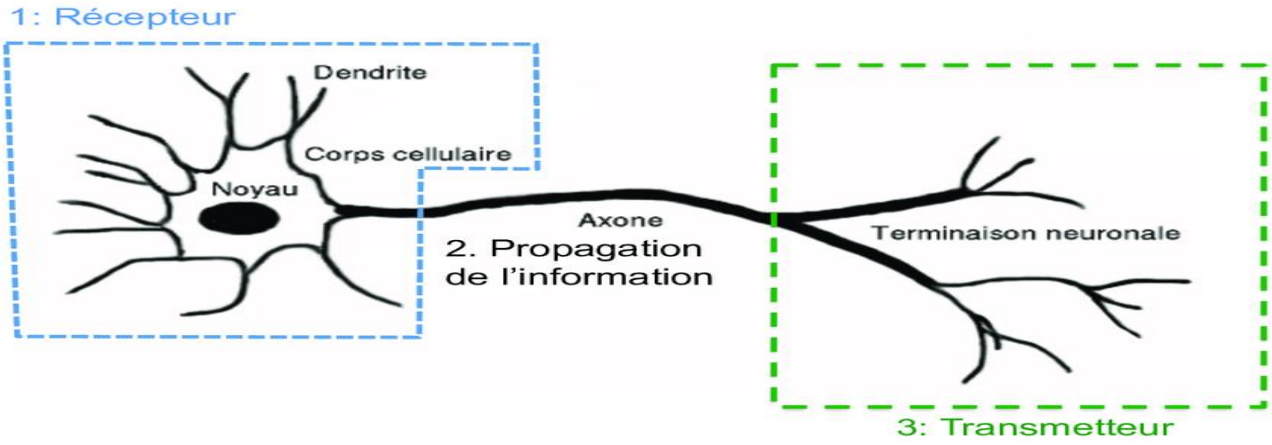

Figure 3.3 : Schéma d'un neurone biologique.[23]

Quant aux synapses, elles font effet de liaison et de pondération entre neurones et permettent donc aux neurones de communiquer entre eux.

4.4.2.2 Le Perceptron

Le perceptron est un algorithme d'apprentissage supervisé de classifier binaires (c'est-à-dire éparant deux classes). Il a été inventé en 1947 par Frank Rosenblatt1 au laboratoire d'aéronautique de l'université Cornell. Il s'agit d'un neurone formel muni d'une règle d'apprentissage qui permet de déterminer automatiquement les poids synaptiques de manière à séparer un problème d'apprentissage supervisé.

Le perceptron peut être vu comme le type de réseau de neurones le plus simple. C'est un classifier linéaire. Ce type de réseau neuronal ne contient aucun cycle (il s'agit d'un réseau de neurones à

propagation avant). Dans sa version simplifiée, le perceptron est monocouche et n'a qu'une seule sortie (booléenne) à laquelle toutes les entrées (booléennes) sont connectées. Plus généralement, les entrées peuvent être des nombres réels(17).

Le mot perceptron est aujourd'hui associé à sa représentation graphique:

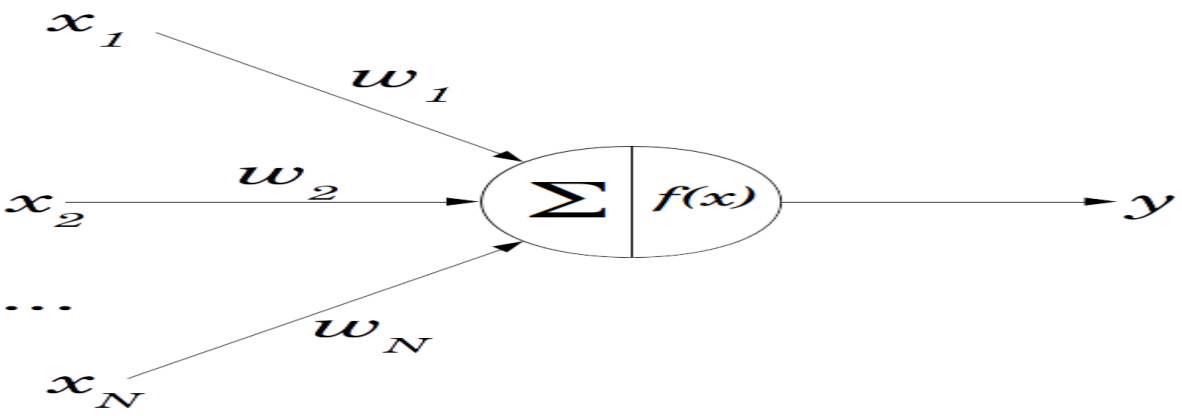

Figure 3.4 : le perceptron est monocouche [17].

Le perceptron est une représentation graphique d'une fonction mathématique composée de deux parties. À gauche, le perceptron calcule la somme pondérée des entrées xi, la partie droite est une fonction appelée fonction d'activation choisie en fonction du type de réseau. De manière générale, il est préférable que cette fonction soit dérivable. La fonction de transfert globale du réseau est donnée par:

$$Y(x_i,w_i) = f\,(w_1x_1 + w_2x_2 + \ldots + w_nx_n)$$

$$Y(x_i,w_i) = \sum_{i=1}^{N} w_ix_i$$

Le perceptron, aussi appelé neurone, est la brique de base des réseaux de neurones. En assemblant ces blocs, il devient possible de créer des réseaux plus complexes.

4.4.2.3 Perceptron multicouche

Le perceptron multicouche (multi layer perceptron MLP) est un type de réseau neuronal artificiel organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie uniquement ; il s'agit donc d'un réseau à propagation directe (feedforward).

Chaque couche est constituée d'un nombre variable de neurones, les neurones de la dernière couche (dite « de sortie ») étant les sorties du système global (Figure 10) [23].

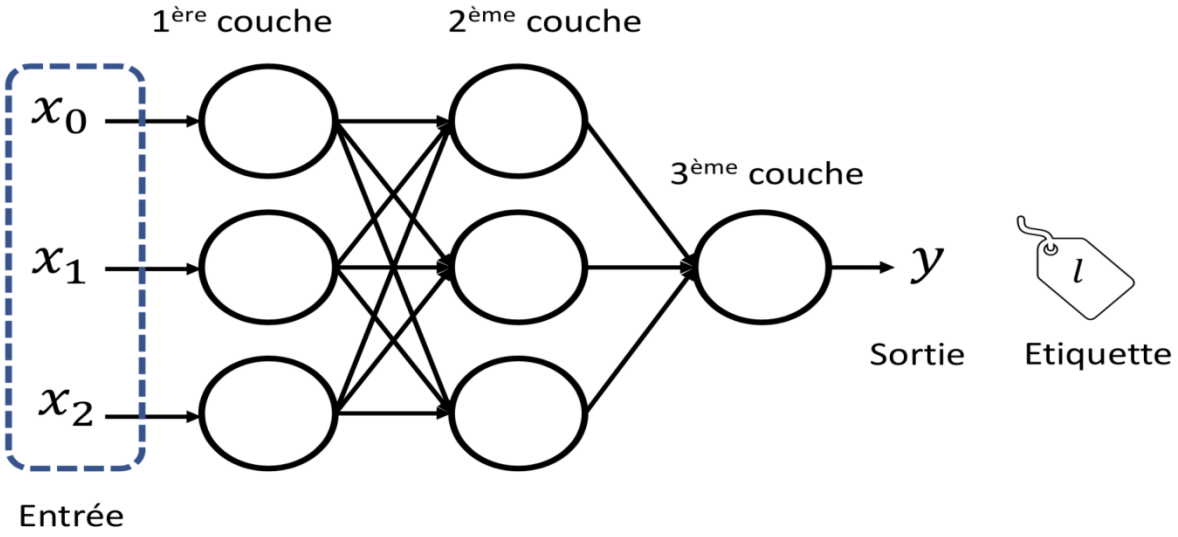

Figure 3.5 : Un perceptron multicouche ou MLP compose de trois couches.[23]

4.4.2.4 Perceptron à couche unique vs multicouches

On distingue deux types de Perceptron : à couche unique et multicouches.

Un Perceptron à couche unique peut apprendre uniquement des fonctions linéaires séparables.

Un Perceptron à couches multiples, aussi appelé réseau neuronal « feedforward », permet de surmonter cette limite et offrent une puissance de calcul supérieure. Il est aussi possible de combiner plusieurs Perceptron pour créer un puissant mécanisme.

3.4.2.5 Quel est le lien entre les neurones biologiques et neurones artificiels?

Un neurone biologique reçoit des signaux électriques d'autres neurones en amont via des points d'entrées (x0, x1, x2 et x3). Ces signaux sont accumulés à l'intérieur du corps du neurone, et si la somme de ces signaux dépasse un certain seuil, le neurone s'active et envoie à son tour un signal à des neurones en aval. Un neurone informatique approxime ces principes. Il accepte en entrée un nombre fixe de nombres réels (représentant les signaux des neurones en amont) et produit en sortie (y) une valeur réelle (représentant le signal envoyé aux neurones en aval). Cette sortie est calculée à partir des entrées via l'équation :

$$Y = f\left(\sum_{i=1}^{N} w_i x_i + b\right)$$

Les entrées sont multipliées par des valeurs w_i que l'on appelle les poids, qui représentent la force de la connexion entre ce neurone et les neurones en amont. La fonction f est une fonction dite d'activation. Il s'agit d'une fonction non-linéaire croissante.

Les premières implémentations de réseaux définissaient f comme une fonction seuil : f renvoie 1 si son argument est strictement positif et 0 sinon, mais d'autres fonctions furent proposées au fil des années comme la fonction unité linéaire rectifiée (Rectified Linear Unit en anglais, ou ReLU). Cette fonction représente l'activation du neurone si celui-ci a accumulé suffisamment de potentiel électrique. La valeur b, que l'on appelle le biais, représente l'appétence ou la résistance du neurone à s'activer. Les poids et le biais (souvent désignés collectivement sous le nom de « poids ») sont les paramètres du neurone : ce sont ces valeurs qui sont modifiées au cours d'un entraînement.

Un neurone informatique peut constituer à lui seul le support d'un algorithme d'apprentissage pour certaines tâches bien définies. En constituant un jeu d'entraînement composé de paires d'entrées réelles et des sorties binaires attendues, et en présentant séquentiellement ces exemples au neurone, il existe un algorithme d'entraînement qui définit comment modifier les poids de celui-ci afin de converger vers la classification attendue.

Cependant, la capacité d'un neurone unique est trop faible pour qu'il soit appliqué ailleurs que sur des cas jouets. Les algorithmes d'apprentissage profond sont basés sur des ensembles de neurones que l'on appelle des réseaux. On appelle « architecture » la structure selon laquelle les neurones sont reliés entre eux. Les premières architectures s'appelaient les perceptron multi-couche (Multi-Layer Perceptron en anglais, ou MLP). Dans un MLP, les neurones sont connectés à la fois parallèlement et séquentiellement selon une organisation en couches .

La première couche est constituée d'un certain nombre de neurones qui prennent en entrée les données. Les sorties des neurones de cette couche servent d'entrée à une deuxième couche de neurones, et ainsi de suite, jusqu'à une dernière couche dont on identifie la sortie à proposition faite par le réseau pour l'étiquetage de l'entrée. Cette structure en couches est inspirée de l'architecture neuronale du cerveau humain [23].

3.4.3 Quelques algorithmes de Deep Learning

Il existe différents algorithmes de Deep Learning. Nous pouvons ainsi citer:

3.4.3.1 Les réseaux de neurones dits récurrents (RNN)

Les réseaux de neurones récurrents sont au cœur de bon nombre d'améliorations substantiels dans des domaines aussi divers que la reconnaissance vocale, la composition automatique de musique, l'analyse de sentiments, l'analyse de séquence ADN, la traduction automatique.

3.4.3.2 Les réseaux de neurones convolutionnel (CNN ou Convolutional Neural Networks). Le problème est divisé en sous parties, et pour chaque partie, un «cluster» de neurones sera créer afin d'étudier cette portion spécifique. Par exemple, pour une image en couleur, il est possible de diviser l'image sur la largeur, la hauteur et la profondeur (les couleurs).

3.4.3.3 Les réseaux antagonistes génératifs (GAN)

les GAN permettent de créer de nouvelles instances de données qui ressemblent aux données sur lesquelles ils ont été formés. Les réseaux antagonistes génératifs sont composés d'un générateur et d'un discriminateur. Ils servent à différentes fins tels que la création ou l'amélioration d'images, la génération de texte, etc.

3.4.3.4 Les réseaux de mémoire à long terme et à court terme (LSTM)

Les LSTM sont des types de réseaux neuronaux récurrents capables d'apprendre et de mémoriser des dépendances à long terme. Ce sont les sorties de ces réseaux que les RNN mémorisent et dont ils se servent comme nouvelles entrées. En plus des utilisations communes aux RNN, les réseaux de mémoire à long terme et à court terme sont utilisés dans la reconnaissance vocale, la composition musicale ou le développement de nouveaux médicaments.

3.5 Les réseaux de neurones convolutionnel(CNN)

Les réseaux de neurones convolutionnel sont à ce jour les modèles les plus performants pour classer des images. Désignés par l'acronyme CNN, de l'anglais « Convolutional Neural Network », ils comportent deux parties bien distinctes. En entrée, une image est fournie sous la forme d'une matrice de pixels. Elle a 2 dimensions pour une image en niveaux de gris. La couleur est représentée par une troisième dimension, de profondeur 3 pour représenter les couleurs fondamentales [Rouge, Vert, Bleu] **[10]**.

3.5.1 Définition

un réseau de neurones convolutifs ou réseau de neurones à convolution (en anglais CNN ou Convolutional Neural Networks) est un type de réseau de neurones artificiels acycliques (feedforward), dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux. Les neurones de cette région du cerveau sont arrangés de sorte qu'ils correspondent à des régions qui se chevauchent lors du pavage du champ visuel. Leur fonctionnement est inspiré par les processus biologiques, ils consistent en un empilage multicouche de perceptrons, dont le but est de prétraiter de petites quantités d'informations. Les réseaux neuronaux convolutifs ont de larges

applications dans la reconnaissance d'image et vidéo, les systèmes de recommandation et le traitement du langage naturel.[24]

3.5.2 Différence entre CNN et perceptron multicouche

Bien qu'efficaces pour le traitement d'images, les perceptrons multicouches (MLP) ont des difficultés à gérer des images de grande taille, en raison de la croissance exponentielle du nombre de connexions avec la taille de l'image, du fait que chaque neurone est « totalement connecté » à chacun des neurones de la couche précédente et suivante. Les réseaux de neurones convolutifs, dont le principe est inspiré de celui du cortex visuel des vertébrés, limite au contraire le nombre de connexions entre un neurone et les neurones des couches adjacentes, ce qui diminue drastiquement le nombre de paramètres à apprendre. Pour un réseau profond tel que AlexNet par exemple, plus de 90 % des paramètres à apprendre sont dus aux 3 couches « complètement connectées » les plus profondes, et le reste concerne les (4) couches convolutives.

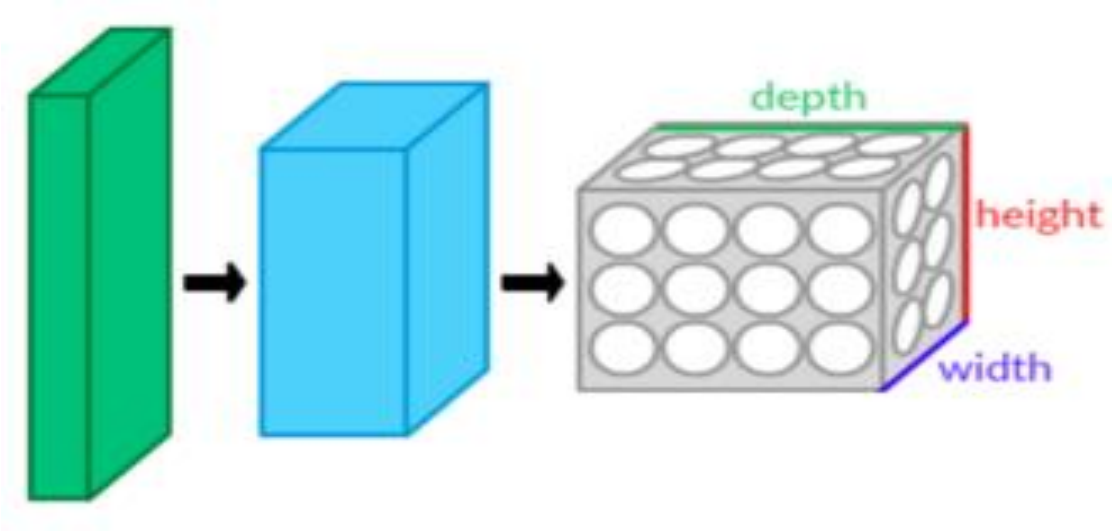

Figure 3.6 : Une couche du CNN en 3 dimensions.[10]

si on prend une image de taille 32 × 32 × 3 (32 de large, 32 de haut, 3 canaux de couleur), un seul neurone entièrement connecté dans la première couche cachée du MLP aurait 3 072 entrées (32*32*3). Une image 200 × 200 conduirait ainsi à traiter 120 000 entrées par neurone ce qui, multiplié par le nombre de neurones, devient énorme.

Les réseaux de neurones convolutifs visent à limiter le nombre d'entrées tout en conservant la forte corrélation « spatialement locale » des images naturelles. Par opposition aux MLP,

3.5.3 Les couches de réseau de neurones convolutifs CNN

Il existe quatre types de couches pour un réseau de neurones convolutifs : la couche de convolution, la couche de pooling, la couche de correction ReLU et la couche fully-connected . Une architecture de réseau de neurones convolutifs est formée par un empilement de couches de traitement :

3.5.3.1 La couche de convolution :

Quand on lui présente une nouvelle image, le CNN ne sait pas exactement si les caractéristiques seront présentes dans l'image ou où elles pourraient être, il cherche donc à les trouver dans toute l'image et dans n'importe quelle position. En calculant dans toute l'image si une caractéristique est présente, nous faisons un filtrage. Les mathématiques que nous utilisons pour réaliser cette opération sont appelés une convolution, de laquelle les réseaux de neurones à convolution tiennent leur nom. La couche de convolution est la composante clé des réseaux de neurones convolutifs, et constitue toujours au moins leur première couche Son but est de repérer la présence d'un ensemble de features dans les images reçues en entrée .

Elle fonctionne comme un extracteur de caractéristiques des images. Une image est passée à travers une succession de filtres, ou noyaux de convolution, On obtient pour chaque paire (image, filtre) une carte d'activation, ou feature map, qui nous indique où se situent les features dans l'image : plus la valeur est élevée, plus l'endroit correspondant dans l'image ressemble à la feature. Finalement, les valeurs des dernières feature maps sont concaténées dans un vecteur. Ce vecteur définit la sortie du premier bloc, et l'entrée du second.

La convolution, d'un point de vue simpliste, est le fait d'appliquer un filtre mathématique à une image. D'un point de vue plus technique, il s'agit de faire glisser une matrice par-dessus une image, et pour chaque pixel, utiliser la somme de la multiplication de ce pixel par la valeur de la matrice. Cette technique nous permet de trouver des parties de l'image qui pourraient nous être intéressantes. Prenons la Figure ci-dessous à gauche comme exemple d'image et la Figure à droite comme exemple de filter Dans le cas d'image, les valeurs sont binaires. Dans un cas réel, les valeurs devraient varier entre 0 et 244. Dans le cas de filtre, les valeurs sont représentées par des 1 et 0. Dans un cas réel, ces valeurs sont continues et peuvent être positives ou négatives.[24]

Appliquer le filtre sur l'image : dans la matrice image M, nous pouvons voir que chaque valeur des pixels de l'image tuile est multipliée par chaque valeur correspondante du filtre (3x1, 0x0, 1x-1 ….). Puis additionner tous ces valeurs pour obtenir une seule valeur '-4' qui fera partie d'une nouvelle image convoluée.

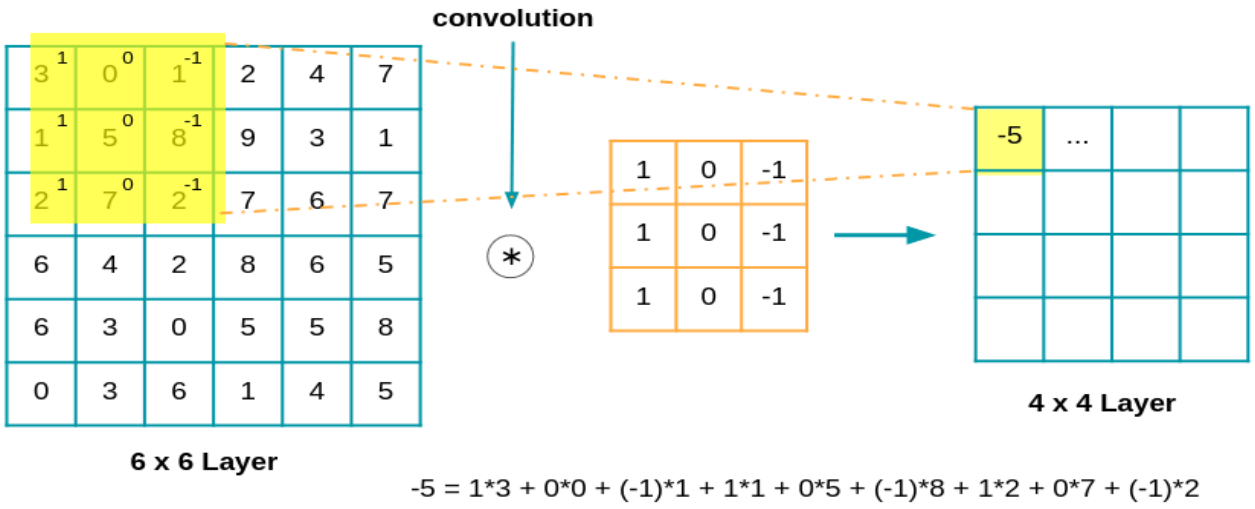

Figure 3.7 : Exemple La couche de convolution [24]

Noté qu'une convolution 3x3 de profondeur 1 effectuée sur une carte de caractéristiques d'entrée 6x6, également de profondeur 1. Comme il y a neuf emplacements 3x3 possibles pour extraire les tuiles de la carte de caractéristiques 6x6, cette convolution génère une carte de caractéristiques de sortie 4x4.

Un réseau de neurones à convolution contient de multiples filtres et ces filtres sont appliqués sur l'image d'origine. Après la première étape nous avons donc autant de nouvelles images que de filtres. La phase de convolution peut aussi être vue comme des couches de neurones cachées où chaque neurone n'est connecté qu'à quelques neurones de la couche suivante.[24]

Les filtres sont aussi adaptés à chaque itération d'apprentissage car les valeurs des filtres mathématiques utilisés sont des poids comme dans les réseaux de neurones multicouches.

3.5.3.2 La couche de pooling :

Un autre outil très puissant utilisé par les CNNs s'appelle le Pooling. Qui est une méthode permettant de prendre une large image et d'en réduire la taille tout en préservant les informations les plus importantes qu'elle contient.Ce type de couche est souvent placé entre deux couches de convolution : elle reçoit en entrée plusieurs feature maps, et applique à chacune d'entre elles l'opération de pooling.

La méthode utilisée consiste à imaginer une fenêtre de 2 ou 3 pixels qui glisse au-dessus d'une image, comme pour la convolution. Mais, cette fois-ci, nous faisons des pas de 2 pour une fenêtre de taille 2, et des pas de 3 pour 3 pixels. La taille de la fenêtre est appelée « kernel size » et les pas s'appellent «strides » Pour chaque étape, nous prenons la valeur celon les méthodes courantes de Pooling, cette valeur constitue un nouveau pixel dans une nouvelle image. [24]

En pratique, il existe deux méthodes courantes de Pooling :

- **Le Max Pooling** : calcule la valeur maximale pour chaque fenêtre de la carte des caractéristiques, c'est-à-dire que l'on garde la valeur maximale du pixel de cette région de l'image.
- **L'Average Pooling** : calcule la valeur moyenne de chaque fenêtre de la carte des caractéristiques, c'est-à-dire que l'on conserve la moyenne de tous les pixels contenus dans la région de l'image.

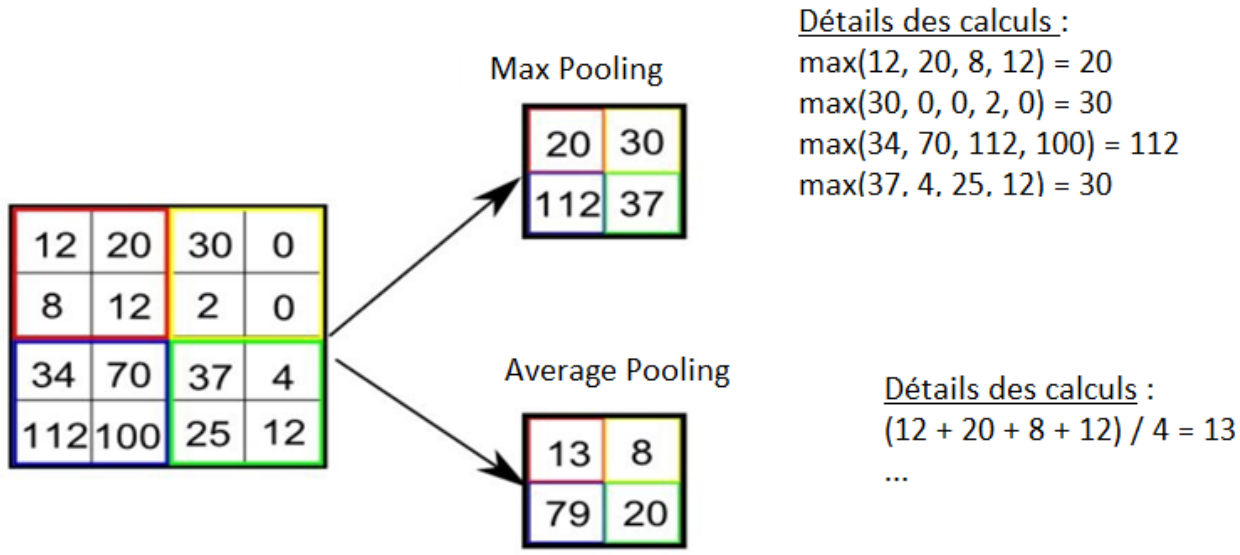

Figure 3.8 : Illustration des techniques de Max pooling & Average pooling [24].

Le pooling permet de gros gains en puissance de calcul. Cependant, en raison de la réduction agressive de la taille de la représentation (et donc de la perte d'information associée), la tendance actuelle est d'utiliser de petits filtres (type 2 × 2). Il est aussi possible d'éviter la couche de pooling mais cela implique un risque de sur-apprentissage plus important.

3.5.3.3 La couche de correction (Segmoide,ReLU,....)

Souvent, il est possible d'améliorer l'efficacité du traitement en intercalant entre les couches de traitement une couche qui va opérer une fonction mathématique (fonction d'activation) sur les signaux de sortie. On a notamment :

fonction sigmoïde la fonction sigmoïde est la fonction d'activation la plus ancienne et la plus populaire [84], elle est définie comme :

$$S(x)=\frac{1}{1+e^{-z}}.$$

e est la constante exponentielle, à peu près égale à 2,71828. Un neurone qui utilise un sigmoïde comme fonction d'activation est appelé un neurone sigmoïde. Nous fixons d'abord la variable z à la somme pondérée des entrées, puis la transmettons à la fonction sigmoïde.

$$Z=b+\sum_{i}^{\square} wixi$$

La sigmoïde a des propriétés non-linéaires telles que :

- bornée inférieurement par 0
- saturée quand les entrées deviennent grandes
- bornée supérieurement par 1

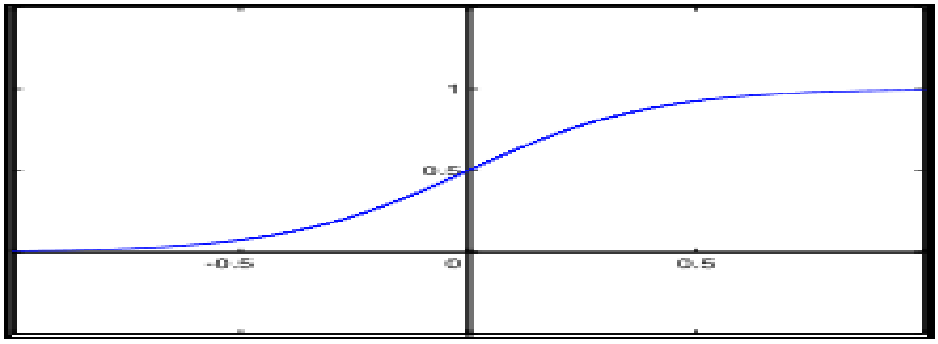

Figure 3.9 : Fonction sigmoïde [24].

Fonction ReLU est une fonction d'activation très couramment utilisée. Acronyme de Rectified Linear Unit (unité linéaire rectifiée), elle permet tout simplement de remplacer les résultats négatifs par zéro. La fonction ReLU est interprétée par la formule: f(x) = max (0, x)

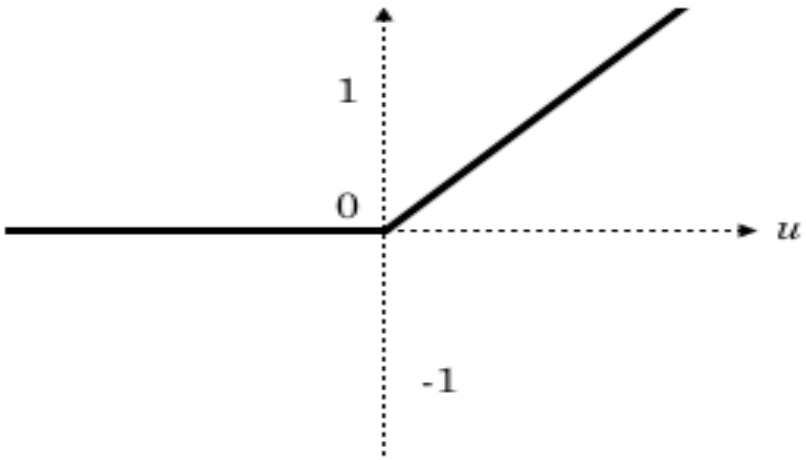

Figure 3.10 : Fonction relu [24].

Fonction softmax est une généralisation de la fonction logistique qui prend en entrée un vecteur Z=(z1,z2,.....zk) de *K* nombres réels et qui en sort un vecteur σ(z) de *K* nombres réels strictement positifs et de somme 1.

La fonction est définie par :

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

Pour tout j € {1,…,k}

C'est-à-dire que la composante ***j*** du vecteur **σ(z)** est égale à l'exponentielle de la composante ***j*** du vecteur **z** divisée par la somme des exponentielles de toutes les composantes de **z**.

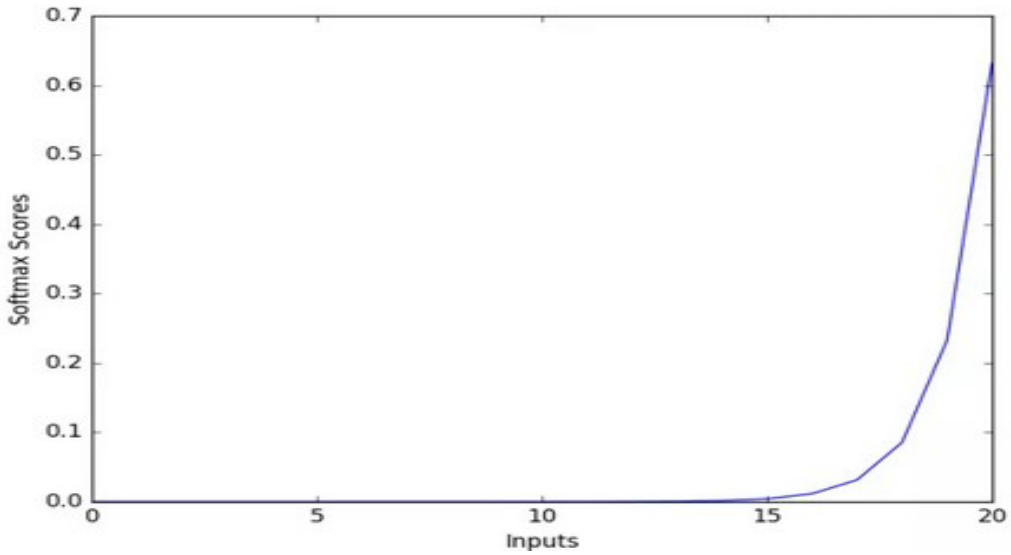

Figure 3.11 : Fonction SoftMax [24].

Souvent, la correction ReLU est préférable, car il en résulte la formation de réseau neuronal plusieurs fois plus rapide, sans faire une différence significative à la généralisation de précision.

3.5.3.4 La couche de entièrement connectée (FC)

La couche de entièrement connectée constitue toujours la dernière couche d'un réseau de neurones, Après plusieurs couches de convolution et de max-pooling, le raisonnement de haut niveau dans le réseau neuronal se fait via des couches entièrement connectées. Les neurones dans une couche entièrement connectée ont des connexions vers toutes les sorties de la couche précédente (comme on le voit régulièrement dans les réseaux réguliers de neurones) [10]. Leurs fonctions d'activations peuvent donc être calculées avec une multiplication matricielle suivie d'un décalage de polarisation.

Ce type de couche reçoit un vecteur en entrée et produit un nouveau vecteur en sortie. Pour cela, elle applique une combinaison linéaire puis éventuellement une fonction d'activation aux valeurs reçues en entrée.

la couche FC permet de classifier l'image en entrée du réseau : elle renvoie un vecteur de taille N, où N est le nombre de classes dans notre problème de classification d'images. Chaque élément du vecteur indique la probabilité pour l'image en entrée d'appartenir à une classe.

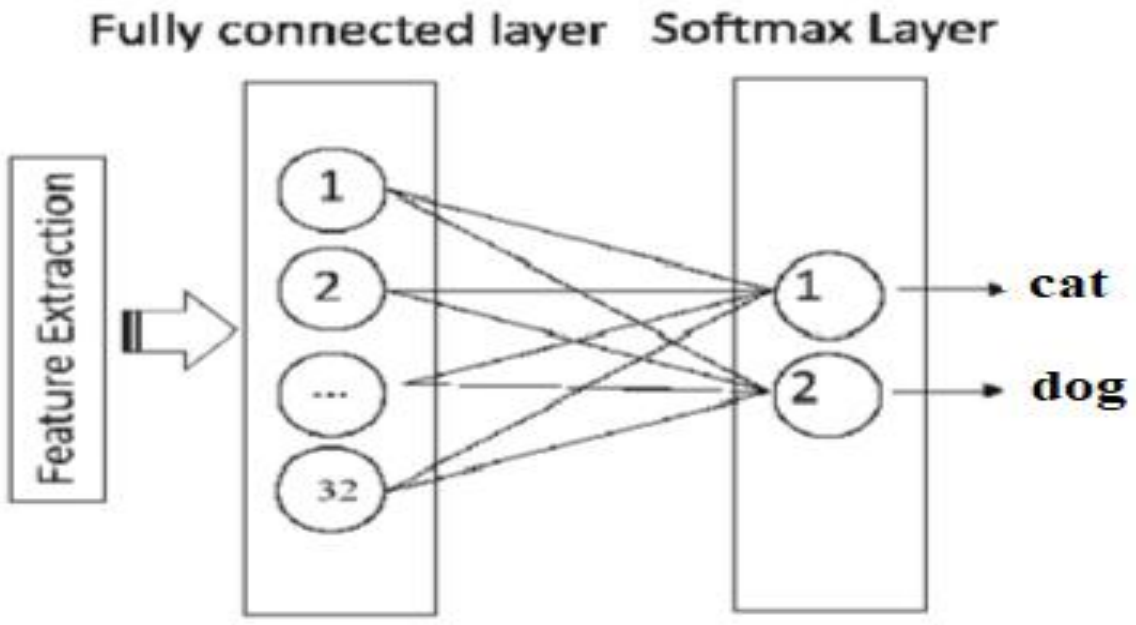

Figure 3.12 : Illustration la couche de entièrement connectée[10].

Par exemple, si le problème consiste à distinguer les chats des chiens, le vecteur final sera de taille 2 : le premier élément (respectivement, le deuxième) donne la probabilité d'appartenir à la classe "chat" (respectivement "chien"). Ainsi, le vecteur [0.9 0.1] signifie que l'image a 90% de chances de représenter un chat.

Chaque valeur du tableau en entrée "vote" en faveur d'une classe. Les votes n'ont pas tous la même importance : la couche leur accorde des poids qui dépendent de l'élément du tableau et de la classe.

Pour calculer les probabilités, la couche fully-connected multiplie donc chaque élément en entrée par un poids, fait la somme, puis applique une fonction d'activation (logistique si N=2, softmax si N>2) :

Ce traitement revient à multiplier le vecteur en entrée par la matrice contenant les poids. Le fait que chaque valeur en entrée soit connectée avec toutes les valeurs en sortie explique le terme entièrement connectée.

3.5.4 Architecture d'un CNN

Un CNN est simplement un empilement de plusieurs couches de convolution, pooling, correction ReLU et fully-connected. Chaque image reçue en entrée va donc être filtrée, réduite et corrigée plusieurs fois, pour finalement former un vecteur. Dans le problème de classification, ce vecteur contient les probabilités d'appartenance aux classes.

Tous les réseaux de neurones convolutifs doivent commencer par une couche de convolution et finir par une couche fully-connected. Les couches intermédiaires peuvent s'empiler de différentes

manières, à condition que la sortie d'une couche ait la même structure que l'entrée de la suivante. Par exemple, une couche fully-connected, qui renvoie toujours un vecteur, ne peut pas être placée avant une couche de pooling, puisque cette dernière doit recevoir une matrice 3D.

En général, un réseau de neurones empile plusieurs couches de convolution et de correction ReLU, ajoute ensuite une couche de pooling (facultative), et répète ce motif plusieurs fois ; puis, il empile des couches fully-connected. Plus il y a de couches, plus le réseau de neurones est "profond" : on est en plein dans le Deep Learning!

La première couche de convolution apprend des features simples, qui représentent des éléments de structure rudimentaires de l'image (contours, coins...) Plus les couches de convolution sont "hautes", c'est-à-dire loin de l'entrée du réseau, plus les Features apprises sont complexes : celles-ci se composent des Feature plus simples des couches précédentes. Un carré est un exemple de Feature complexe, formée de contours et de coins.

Les couches de convolution les plus hautes apprennent donc des Features sophistiquées : par exemple couche convolution 2, dans le cas de la reconnaissance de chat ci-dessous, elles peuvent correspondre aux oreilles, nez ou œil.. **[10].**

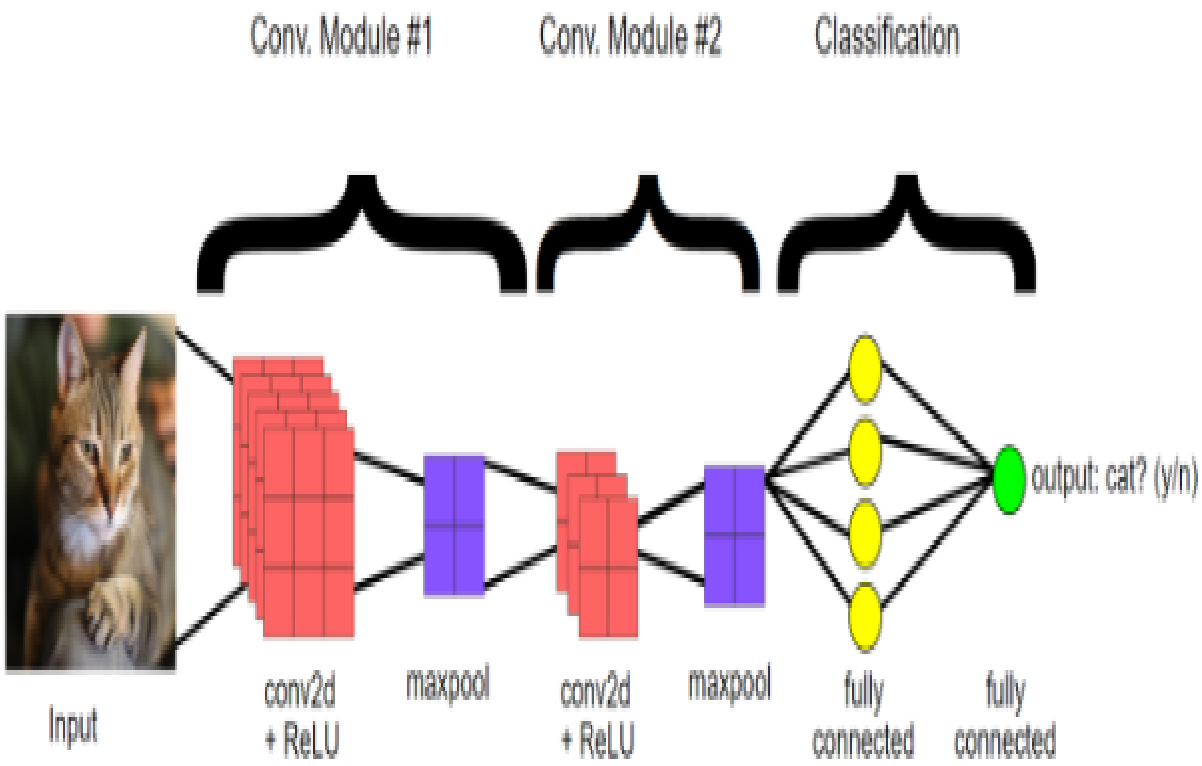

Figure 3.13 : Illustration Architecture d'un CNN [10].

3.5.5 Choix des paramètres des couches

Les réseaux de neurones convolutifs utilisent plus d'hyper paramètres qu'un perceptron multicouche standard. Même si les règles habituelles pour les taux d'apprentissage et des constantes de régularisation s'appliquent toujours, il faut prendre en considération les notions de nombre de filtres, leur forme et la forme du max pooling.

- **Nombre des filtres :** En pratique, un CNN apprend seul les valeurs de ces filtres pendant le processus de l'entrainement. Les paramètres tels que le nombre de filtres, sont spécifiés par le scientifique avant le lancement du processus de l'entrainement. Plus nous avons des filtres, plus les caractéristiques d'image sont extraites et plus le réseau devient performant au niveau de l'extraction des caractéristiques et de la classification des images.
- **Forme du filtre :** Les formes de filtre varient grandement dans la littérature. Ils sont généralement choisis en fonction de l'ensemble de données.
- **Forme du Max Pooling:** Les valeurs typiques sont 2 × 2. De très grands volumes d'entrée peuvent justifier un pooling 4 × 4 dans les premières couches. Cependant, le choix de formes plus grandes va considérablement réduire la dimension du signal, et peut entraîner la perte de trop d'information.

3.5.6 Avantages d'un CNN dans le domaine de la reconnaissance d'images

Comparé aux réseaux neuronaux conventionnels, le CNN offre de nombreux avantages :

- Il convient aux applications d'apprentissage automatique et d'Intelligence artificielle avec de grandes quantités de données d'entrée telles que la reconnaissance d'images.
- Le réseau fonctionne de manière robuste et est insensible à la distorsion ou à d'autres changements optiques.
- Il peut traiter des images enregistrées dans différentes conditions d'éclairage et dans multi focus. Les caractéristiques typiques d'une image sont ainsi facilement identifiées.
- Il nécessite beaucoup moins d'espace de stockage que les réseaux de neurones entièrement maillés. Le CNN est divisé en plusieurs couches locales partiellement maillées. Les couches de convolution réduisent considérablement les besoins de stockage.
- Le temps de formation d'un CNN est également considérablement réduit. Grâce à l'utilisation de processeurs graphiques modernes, les CNN peuvent être formés de manière très efficace.
- Il est la technologie de pointe pour L'apprentissage profond et la classification dans la reconnaissance d'images (image recognition).

3.6 Conclusion

Dans ce chapitre, nous avons parlé du Deep Learning et le Machin Learning en général, et sont relations avec Intelligence artificiel. Puis, nous avons présenté les réseaux de neurones avec leur architecture détaillée. et leur utilisation dans traitement d'images et la fusion d'image spécifiquement. Et à la fin nous prestons quelques Avantages d'un CNN dans le domaine de la reconnaissance d'images.

Le prochain chapitre, traite les détails de la conception, ainsi que la méthode et les outils utilisés pour la réalisation de notre application.

Chapitre 4

Implémentation et résultat

4.1 Introduction

Ces derniers années les approches basées sur Deep Learning ont montré des résultats prometteurs dans la fusion des images multi-focus, dans le chapitre quatrième que nous avons étudies sur les méthodes de fusion d'images basées sur deep learning , nous avons trouvé plusieurs algorithmes qui donnent de bons résultats comme, CNN, GAN, KNN, ……

Dans ce chapitre, nous avons aussi implémenté la méthode CNN, qui aide à surmonter les problèmes de perte de détails spatiaux et de flou qui sont observé dans les images, Cette approche surpasse les précédentes méthodes. Nous allons faire la conception et présenter aussi la mise en œuvre de notre application en utilisant le langage Python et bibliothèque d'apprentissage automatique Pytorch . Le réseau CNN sera entraîné on google colab ,On commençant tout d'abord par une présentation du langage de programmation choisi. Ensuite nous présentons l'architecture de CNN et résultat de l'exécution de notre application.

4.2 Méthode proposée

Le méthode de fusion d'images multi-focus dans le domaine spatial basée sur schéma de division de blocs, dans lequel chaque image source est divisée en un certain nombre de blocs de taille fixe et domaine spatial utilise des caractéristiques spatiales locales telles que le gradient, l'écart type local. La méthode proposée introduit une architecture qui utilise des réseaux de neurones convolutifs (CNN) entraînés sur trois dataset contenant chacun les images originales, le gradient des images dans les directions horizontale et verticale. les dataset proposés prépare un type simple d'image multi-focus pour obtenir les meilleures performances de fusion.

4.3 Réseaux de neurones convolutifs et processeurs graphiques

Entraîner un réseau de neurones convolutif est très coûteux : plus les couches s'empilent, plus le nombre de convolutions et de paramètres à optimiser est élevé. L'ordinateur doit être en mesure de stocker plusieurs giga-octets de données et de faire efficacement les calculs. C'est pourquoi les

fabricants de matériel informatique multiplient les efforts pour fournir des processeurs graphiques (GPU) performants, capables d'entraîner rapidement un réseau de neurones profond en parallélisant les calculs.Une unité de traitement graphique (GPU) est un matériel similaire à une unité de traitement centrale (CPU), à la différence qu'elle est conçue exclusivement pour les images et que toute fonction qu'elle remplit est appelée traitement d'image GPU. Contrairement au processeur et à la plupart des processeurs d'image, le traitement d'image GPU traite chaque image comme une image tridimensionnelle (3D), même si l'image est bidimensionnelle (2D). Les images complexes ont souvent des textures et un processeur graphique peut charger plusieurs textures à la fois.[25]

4.4 Logiciels et Bibliothèques Utilisés dans l'implémentation

4.4.1 Paython :

Python est un langage de programmation de haut niveau interprété (il n'y a pas d'étape de compilation) et orienté objet avec une sémantique dynamique. Il est très sollicité par une large communauté de développeurs et de programmeurs. Python est un langage simple, facile à apprendre et permet une bonne réduction du cout de la maintenance des codes. Les bibliothèques (packages) python encouragent la modularité et la réutilisabilité des codes. Python et ses bibliothèques sont disponibles (en source ou en binaires) sans charges pour la majorité des plateformes et peuvent être redistribués gratuitement.

4.4.2 Google Colab :

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur.
Est un service hébergé de notebooks Jupyter qui ne nécessite aucune configuration et permet d'accéder gratuitement à des ressources informatiques, dont des GPU [26].

Colab permet :

-D'améliorer vos compétences de codage en langage de programmation Python.
-De développer des applications en Deep Learning en utilisant des bibliothèques Python populaires telles que Keras, TensorFlow, PyTorch et OpenCV.
-D'utiliser un environnement de développement (Jupyter Notebook) qui ne nécessite aucune configuration.
-Développement et entraînement de réseaux de neurones
-Expérimentation avec les TP

Principales fonctionnalités :

- Plusieurs types de GPU
- Quels que soient vos besoins en performances et votre budget, les GPU NVIDIA K80, P100, P4, T4, V100 et A100 offrent une variété d'options de calcul adaptées à votre charge de travail.
- Des performances flexibles
- Profitez d'un parfait équilibre entre le processeur, la mémoire, le disque hautes performances et jusqu'à 12GPU par instance en fonction de votre charge de travail individuelle. En outre, avec la facturation à la seconde, vous ne payez que ce que vous utilisez.
- Tous les avantages de Google Cloud
- Exécutez vos charges de travail GPU sur Google Cloud Platform afin de profiter des technologies de stockage, de mise en réseau et d'analyse de données les plus performantes du secteur.[26]

4.4.3 PyTorch :

PyTorch est une bibliothèque d'apprentissage automatique pour Python utilisée principalement pour le traitement du langage naturel. Le logiciel open source a été développé par les équipes d'intelligence artificielle de Facebook Inc. en 2016. PyTorch offre deux fonctionnalités importantes, notamment le calcul tensoriel, ainsi que des réseaux de neurones profonds fonctionnels

PyTorch utilise un module Autograd pour calculer la différenciation automatique. Bref, un enregistreur détaille les opérations effectuées puis les rejoue pour synthétiser les gradients. Cela permet de gagner du temps dans le développement des réseaux de neurones car la différenciation des données est effectuée rapidement lors de la passe avant. Le package optim de PyTorch permet à un utilisateur de définir un optimiseur qui mettra à jour les poids automatiquement. Cependant, lorsque les utilisateurs souhaitent créer leur propre modèle personnalisé, ils peuvent tirer parti du module nn.module de PyTorch. Compte tenu des différents modules, PyTorch vous permet d'implémenter différents types de couches telles que des couches convolutives, des couches récurrentes et des couches linéaires, entre autres [26].

4.4.4 Scikit-learn

Scikit-learn est une bibliothèque libre Python dédiée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria et Télécom ParisTech.

Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec des autres bibliothèques libre Python, notamment NumPy et SciPy.

4.5. Création Dataset:

Afin d'obtenir dataset proposées, On Utilise langage de programmation Matlab et plus de 100 images de haute qualité télécharger en cite Kaggle come la **Figure 4.1**.

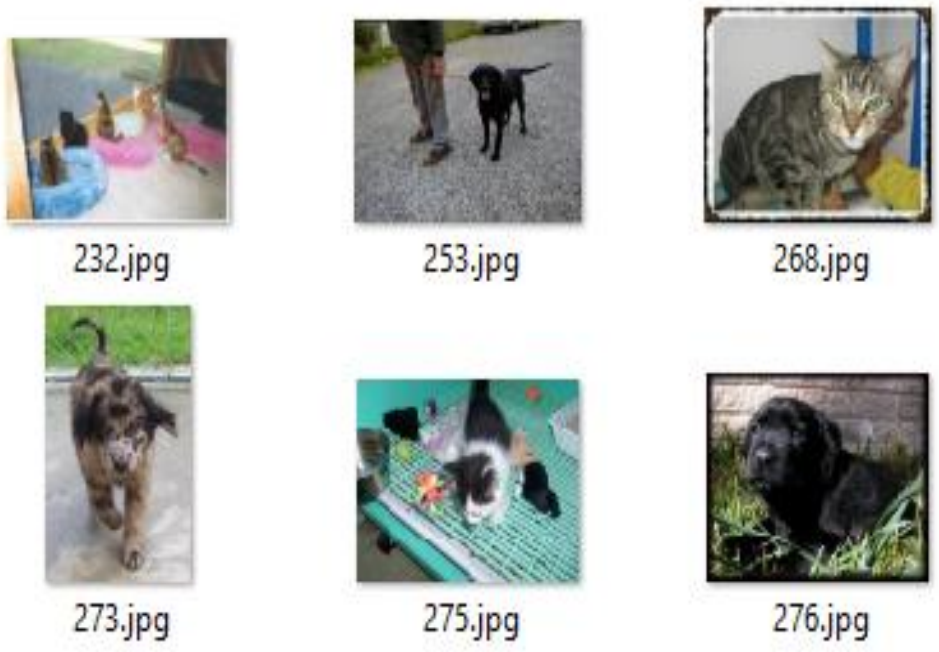

Figure 4.1: quelques exemples d'images utilisées dans la création dataset .

d'abord, converties les images en niveaux de gris, Afin de créer une condition non focalisée, chaque image est passée à travers quatre filtres gaussiens différents avec l'écart type de 9, 11, 13 et 15. Par conséquent, cinq types d'images dont l'image originale et les quatre versions de l'image floue. Ensuite, les gradients dans les directions horizontale (Gx) et verticale (Gy) sont appliqués pour chacun de ces cinq types d'images. Nous allons créer trois groupes contenant chacun les images originales, le gradient dans les directions horizontale (Gx) et verticale (Gy) des images. Ensuite, chaque type d'images de ces trois groupes est divisé en 32 × 32 blocs ou patchs. Cependant, nous savons avec une connaissance préalable que chaque patch provient de la version non floue de l'image ou de l'une des quatre versions floues des images. Puisque nous allons créer nos trois ensembles de données proposés avec les connaissances préalables, Le schéma de Figure 4.2 montre en détail procédure de création de patch.

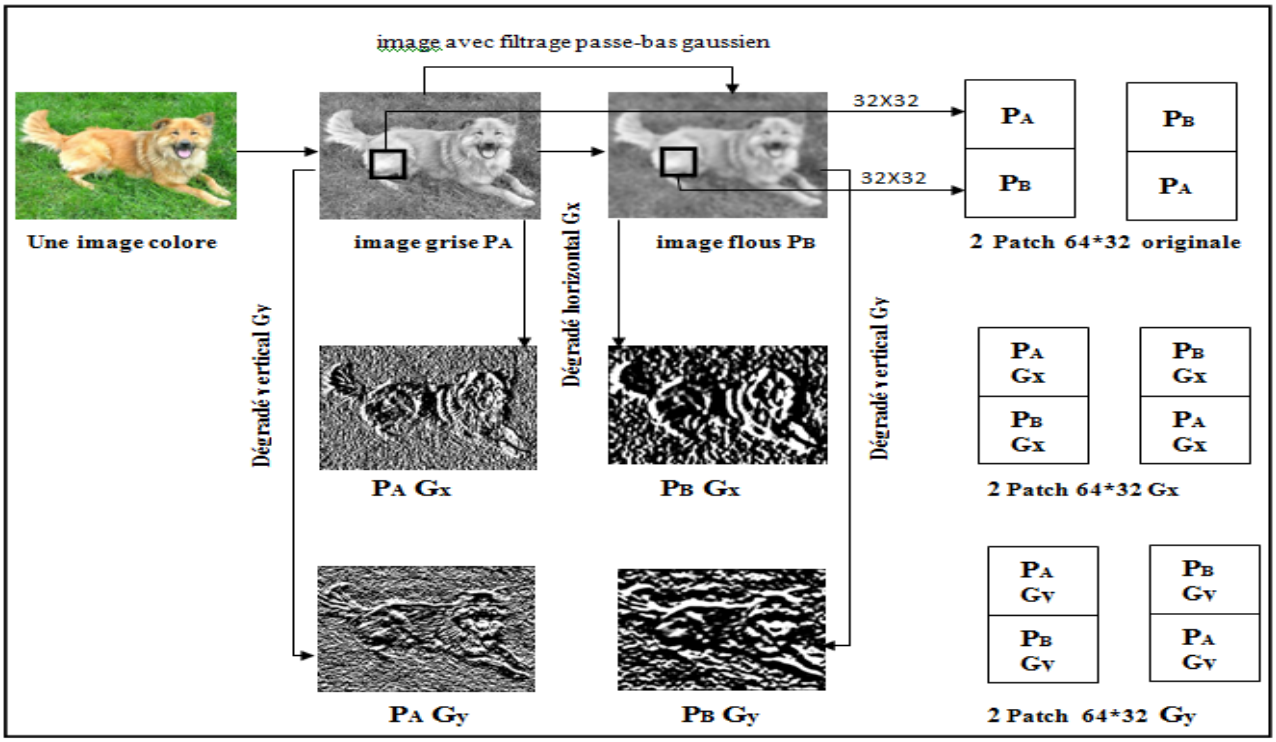

Figure 4.2 : **Le schéma de création Dataset qui est utilisé dans l'entraînement.**

Avec cette procédure de création de patch à partir des images d'origine pour le jeu de données d'origine, les patchs du gradient dans les directions horizontale (Gx) et verticale (Gy) des images sont créés dataset Gx et Gy. 12600 patchs pour entraîne et 3400 patchs pour test en Figure 4.3

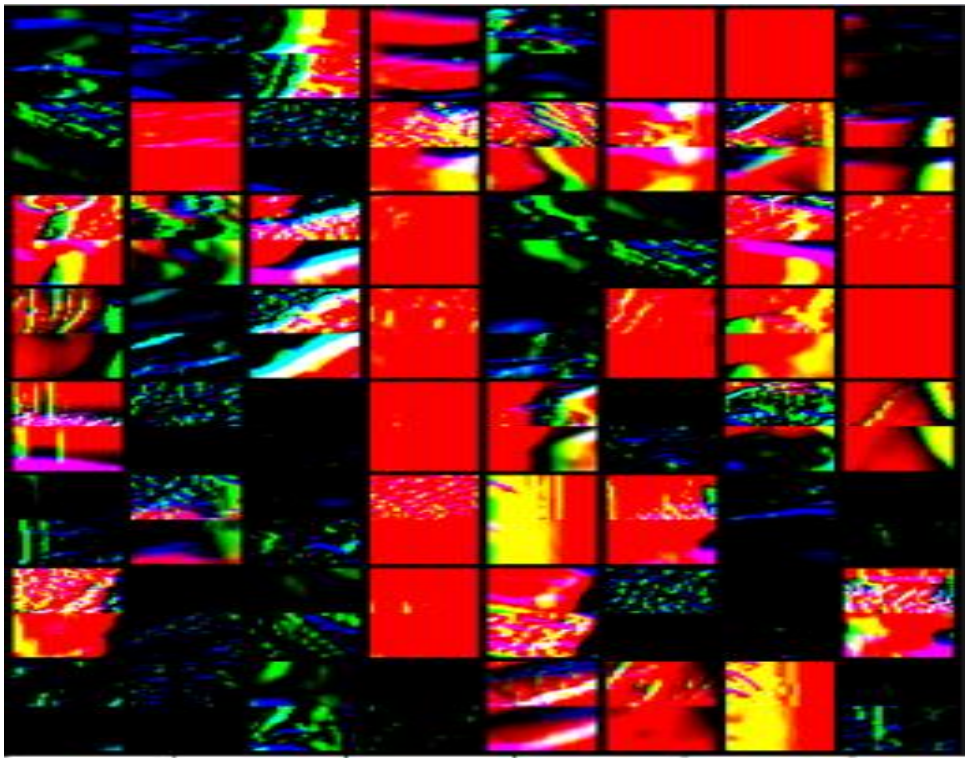

Figure 4.3: les images aléatoires de dataset.

4.6. Architecture de notre modèle CNN

Il existe de nombreuses architectures pour le modèle CNN, mais cet article tente d'implémenter les CNN simples afin de simplifier le problème. Cet article considère quelques nombres de couches convolutives en raison de la simplicité de la fusion d'images multi-focus par rapport aux problèmes avancés tels que la détection d'objets.

L'architecture proposée contient les couches convolutives avec une taille de noyau (kernel) de 3 × 3, stride 1 × 1, une taille de remplissage (padding) de 1 × 1 et une fonction d'activation non linéaire de ReLU(0.1,inplace = True). Max pooling de 2 × 2 est utilisé pour cette architecture. L'ensemble des images d'origine et les images Gx (gradient d'image dans les directions horizontales) et les images Gy (gradient d'image dans les directions verticales) sont trois entrées alimentées le réseau CNN, La couche FC est cartographiée sur les deux derniers neurones pour détecter les étiquettes focalisées et non focalisées.

Notre réseau que nous présentons dans la Figure 4.4 sera structuré avec les 6 couches suivantes :

Conv1 ---> Conv2 ---> Conv3 ---> Conv4 ---> Conv5 ---> Linear.

les trois entrée d'image sont de taille 32*32, les images passe d'abord à la première couche de convolution. Cette couche est composée de 64 filtres de taille 3*3, Chacune de nos couches de convolution est suivie d'une couche BatchNorm2d qui est applique une normalisation sur les entrées pour obtenir une moyenne nulle et une variance unitaire et augmenter la précision du réseau, chaque couche BatchNorm2d est suivie d'une fonction d'activation ReLU cette fonction force les neurones à retourner des valeurs positives, après cette convolution 32 features maps de taille 32*32 seront créés.

Les 64 feature maps qui sont obtenus auparavant ils sont donnés en entrée de la deuxième couche de convolution qui est composée aussi de 128 filtres et un couche BatchNorm2d , une fonction d'activation RELU, ensuite on applique Maxpooling pour réduire la taille de l'image ainsi la quantité de paramètres et de calcul. À la sortie de cette couche, nous aurons 128 feature maps de taille 16*16. On répète la même chose avec les couches de convolutions trois,

Après le couche trois, l'images d'origine passe à la quatrième couche de convolution qui est composée 256 filtrés un couche BatchNorm2d, une fonction d'activation RELU, ensuite on applique Maxpooling , prend les images Gx et Gy et les concatène et passe à la cinquième couche de convolution qui est composée 256 filtrés un couche BatchNorm2d , ensuite on applique Maxpooling.

les sorties qui sont obtenus de la couche quatrième et cinquième nous allons les concatène et passe La couche Linear (FC) est mappé à les deux neurones pour la prédiction finale qui indiquent les étiquettes focalisées et non focalisées. Elle est la couche finale de notre réseau.

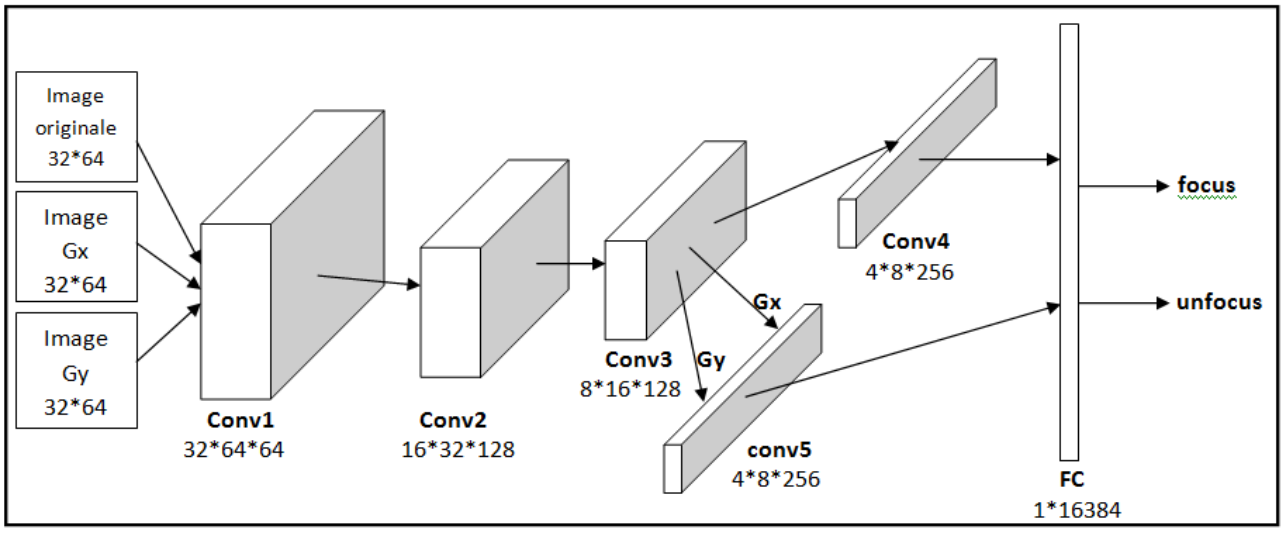

Figure 4.4: Le schéma de l'architecture CNN proposée

4.7. Entraînement et teste du réseau

Pour créer un réseau neuronal avec PyTorch, vous allez utiliser le package torch.nn. Ce package contient des modules, des classes extensibles et tous les composants requis pour générer des réseaux neuronaux. nous avons construit toutes les procédures et calculs comme illustre in Figure 4.5.

```
CNN(
  (conv1): Sequential(
    (0): Conv2d(1, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): LeakyReLU(negative_slope=0.1, inplace=True)
  )
  (conv2): Sequential(
    (0): Conv2d(64, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (1): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): LeakyReLU(negative_slope=0.1, inplace=True)
    (3): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (conv3): Sequential(
    (0): Conv2d(128, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (1): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): LeakyReLU(negative_slope=0.1, inplace=True)
    (3): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (conv4): Sequential(
    (0): Conv2d(128, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (1): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): LeakyReLU(negative_slope=0.1, inplace=True)
    (3): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (conv5): Sequential(
    (0): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (1): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): LeakyReLU(negative_slope=0.1, inplace=True)
    (3): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (fc1): Linear(in_features=16384, out_features=2, bias=True)
)
```

Figure 4.5: **Le configuration de modèle CNN**

4.7.1.Entraînement

nous entraînerons le modèle sur les 12600 images, Il est maintenant temps d'entrainement le modèle ci-dessus. Nous devons accéder à chaque élément manuellement et être disposés en boucle pour en faire un entrainement continu. La procédure est la suivante

Nous commençons par itérer à travers 100 époques dans nos données d'entraînement, Ensuite, nous effectuons une passe arrière dans laquelle nous mettons à jour nos poids afin d'améliorer notre modèle, Les gradients sont alors mis à zéro avant chaque mise à jour à l'aide de la fonction (optimizer.zero grad()),Les nouveaux gradients sont ensuite calculés à l'aide de la (fonction loss.backward()), Enfin, nous utilisons la fonction optimiser.step() pour mettre à jour les poids, résultats pour chaque époque illustre dans la Figure 4.6 .

```
Epoch [1/100], Loss: 0.764395
Epoch [2/100], Loss: 0.255601
Epoch [3/100], Loss: 0.154903
Epoch [4/100], Loss: 0.170637
Epoch [5/100], Loss: 0.089485
Epoch [6/100], Loss: 0.128929
Epoch [7/100], Loss: 0.111777
Epoch [8/100], Loss: 0.064695
Epoch [9/100], Loss: 0.101184
Epoch [10/100], Loss: 0.092986
Epoch [11/100], Loss: 0.091257
Epoch [12/100], Loss: 0.052046
Epoch [13/100], Loss: 0.068314
Epoch [14/100], Loss: 0.043261
Epoch [15/100], Loss: 0.091109
Epoch [16/100], Loss: 0.038759
Epoch [17/100], Loss: 0.067400
Epoch [18/100], Loss: 0.048487
Epoch [19/100], Loss: 0.034903
Epoch [20/100], Loss: 0.035326
Epoch [21/100], Loss: 0.023131
```

Figure 4.6 : illustre les sorties de l'exécution du code entrainement.

L'entraînement de notre CNN a pris ≈ 3252 secondes sur GPU, Nous pouvons voir que la perte de validation diminue à mesure que les époques augmentent. Visualisons les pertes d'entraînement en traçant le graphique du Figure 4.7.

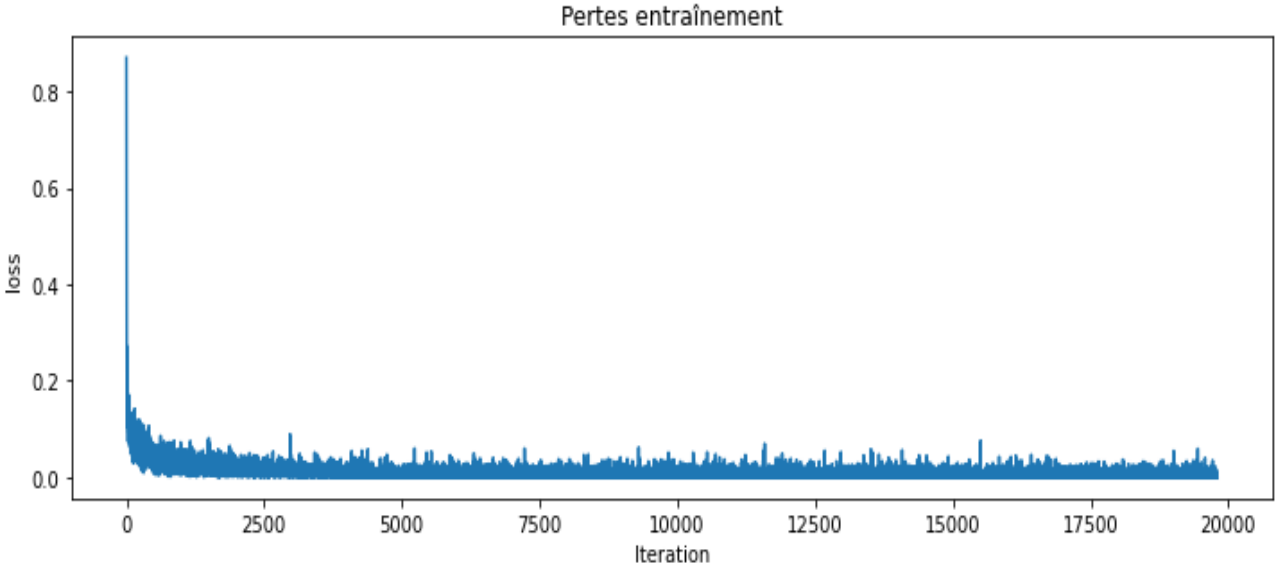

Figure 4.7 :le graphique illustre la perte d'entraînement de chaque itération

Nous avons atteint une précision de traine de 97,4 % avec cette architecture CNN, c'est une très bonne précision.

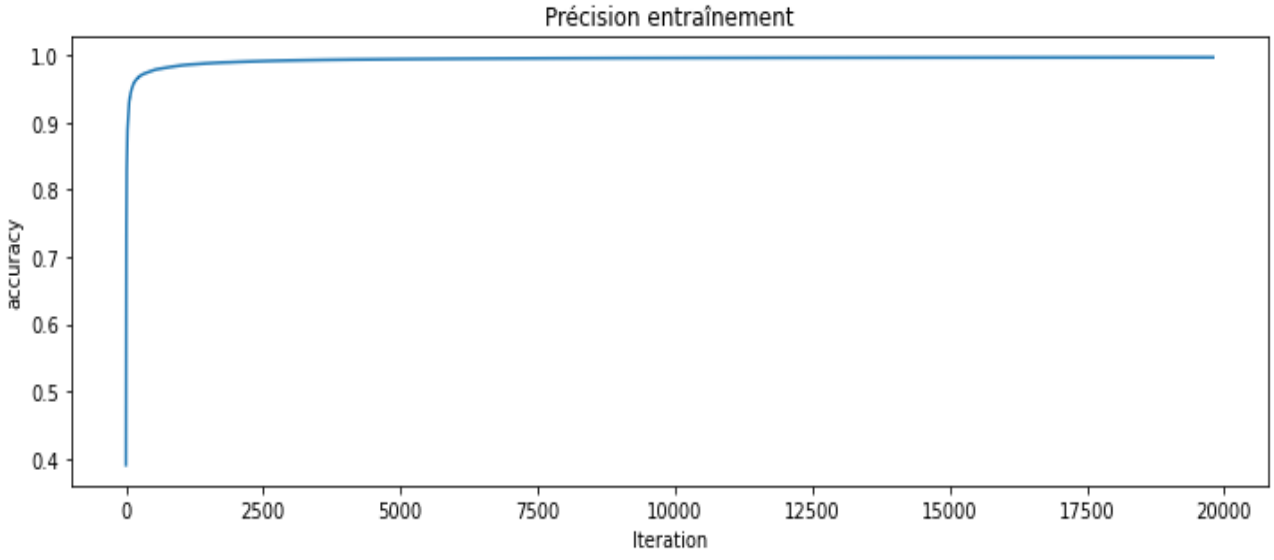

Figure 4.8 :le graphique illustre la précision d'entraînement de chaque itération

Le graphique des précisions et des pertes montre comment notre modèle améliore sa précision après chaque époque. Mais nous devons vérifier si le réseau a appris quelque chose. Nous vérifierons cela au cours de la phase de test

4.7.2.Teste

Pour tester notre modèle, définissons nœuds de précision. Il évaluera notre modèle après chaque itération de test, ce qui vous aidera à suivre les performances de votre modèle. Après chaque itération, le modèle est testé sur 3400 images de test, qui ne seront pas vues dans la phase d'entrainement. Lorsque nous évaluons sur notre ensemble de test, nous atteignons une précision de ≈ 99 %, ce qui est assez bon la Figure 4.9 illustre les sorties de l'exécution de test.

```
accuracy: 99.99
corrects: 3423.52
Toatal: 3424.00
Total Time of Training 961.605605602264404296875s
```

Figure 4.9 :les sorties de l'exécution des images test

4.8. Application :

La fenêtre principale de notre application illustre dans la Figure 4.10 est contaient trois bouton, Cliquer sur le bouton "Upload file" dans la fenêtre (interface) pour ouvrir une boite de dialogue pour choisir un fichier d'image 1 à partir d'un emplacement local, De la même façon on choisit un fichier d'image 2

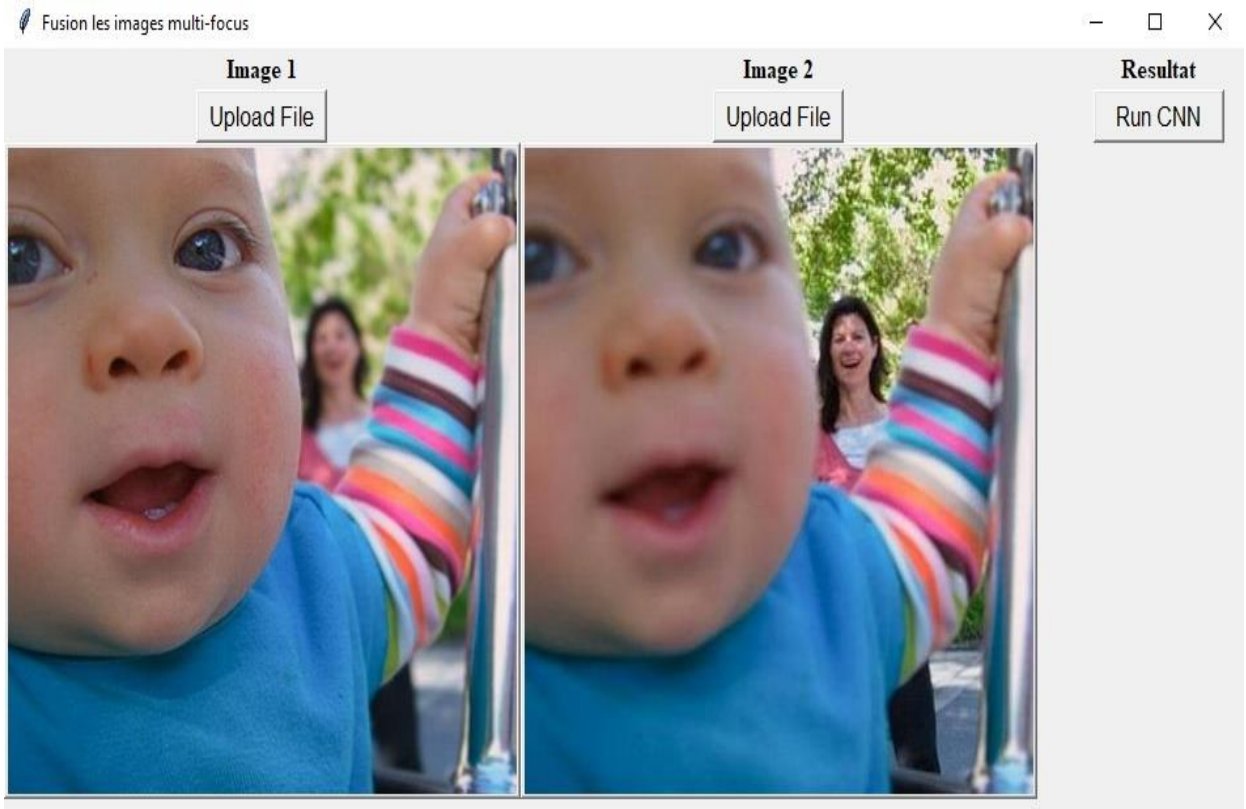

Figure 4.10 :La fenêtre principale de notre application

Utiliser le bouton "run CNN" pour la fusion deux les images qui ont été choisis et produit une image focalisée illustre en la Figure 4.11

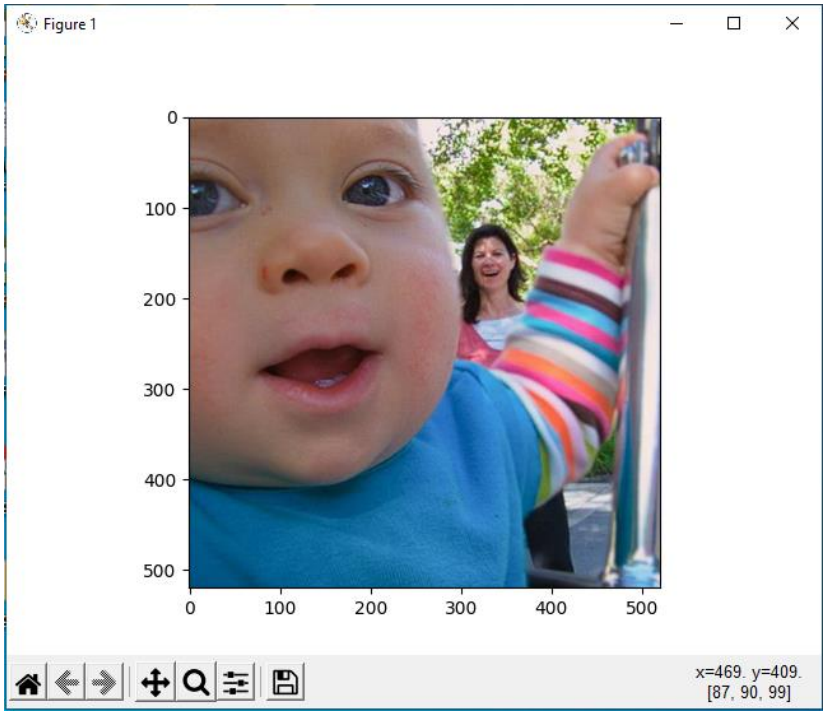

Figure 4.11 : La fenêtre de l'image focalisée

4.9. Le schéma de fusion

Les images multi-focus d'entrée doivent être transformées en images en niveaux de gris pour construire une carte de décision s'il s'agit d'images couleur ; La méthode proposée pourrait être facilement étendue pour fusionner plus de deux images d'entrée. Les images multi-focus d'entrée A et B sont introduites dans le réseau proposé préformé selon la stratégie d'alimentation de patch qui est utilisée pour créer les trois ensembles de données come la Figure 4.2 , Les patchs extraits des images multi-focus d'entrée sont superposés pour alimenter le réseau pré-entraîné afin de simuler la fusion d'images pixel par pixel. Ensuite, le réseau proposé pré-entraîné renvoie les étiquettes de 0 et 1 qui indiquent les étiquettes focalisées et non focalisées, respectivement. Avec cette procédure, chaque pixel est contribué plusieurs fois pour obtenir l'étiquette de focalisé et non focalisé. Chaque patch qui est envoyé au réseau met à jour la carte de score des images multi-focus d'entrée avec la méthode de fusion proposée comme

$$\text{M(r,c)} = \begin{cases} M(r:\ r+b, c:c+b) \mathrel{+}= -1\ if\ \text{étiquette} = 0 \\ M(r:\ r+b, c:c+b) \mathrel{+}= \ 1\ if\ \ \text{étiquette} = 1 \end{cases}$$

où r, c et M indiquent la ligne et la colonne des images d'entrée et la carte de décision, b est la taille de la largeur et de la hauteur des patchs extraits des images multi-focus d'entrée pour la construction des patchs et il doit être redimensionné à 32 pour alimenter le réseau pré-entraîné. la carte de décision segmentée initiale de la méthode proposée est construite comme ci-dessous :

$$D(r,c) = \begin{cases} 1 & if \quad M(r,c) > 0 \\ 0 & if \quad else \end{cases}$$

Enfin, l'image fusionnée finale est calculée comme ci-dessous :

F(r,c) = D(r,c) x A(r,c) + (1- D(r,c) x B(r,c))

où A(r,c) et B(r,c) sont des images multi-focus d'entrée.

4.10 Conclusion

Nous avons présenté dans ce chapitre une approche de fusion des images multi-focus basée sur les réseaux de neurones convolutifs , Nous avons utilise l'environnement de développement Python.

A la fin nous avons présenté notre application en donnant quelques captures d'écran qui expliquent le déroulement et le fonctionnement de notre travail, la méthode proposée permet fusion d'images multi focalise dans un seule image focus précision 99%. Dans les évaluations qualitatives et quantitatives.

Conclusion générale

Pour réaliser notre travail de fusion des images, nous avons utilisé le DeepLearning, une méthode d`apprentissage qui a montré ses performances ces dernières années. Nous avons donc choisi la méthode CNNs car elle permet de présenter différents types de couches utilisées dans la classification : la couche convolutionnelle, la couche de rectification, la couche de pooling et la couche fully-connected Cette choix est justifié par la simplicité et l'efficacité de la méthode.

La fusion d'images est un mécanisme permettant d'améliorer la qualité des images. Les applications importantes de la fusion d'images incluent l'imagerie médicale, l'imagerie microscopique la télédétection, la vision par ordinateur, les robots intelligents et les systèmes de surveillance. Nous avons rencontré quelques obstacles dans la phase d'implémentation, l'utilisation d'un CPU a fait que le temps d'exécution était trop couteux. Afin de régler ce problème on doit utiliser des réseaux de neurones convolutionnels plus profonds déployé sur un GPU.

Dans ce projet nous avons proposé une approche pour la fusion des images multi-focus en une seule image focalisée, qui est s'appuye sur la formation en réseau de neurones convolutionnel CNN en trois dataset différent, Le résultat obtenu lors de la phase de test confirme obtenir des meilleurs résultats en terme de précision et de corrects, bien que les capacités des activités réalisées dans le domaine de fusion des images multi-focus, aucune méthode n'est jugée parfaite à 100%.

Bibliographiques

[01] http://faubertlab.com/perception-visuelle/ (consulté le 05/03/2022)

[02] Alexandre Benoit. Le système visuel humain au secours de la vision par ordinateur. Traitement du signal et de l'image]. Institut National Polytechnique de Grenoble - INPG, 2007. Français. fftel-00193715,

[03] Marie Rousserie. Attention visuelle. Médecine humaine et pathologie. 2015. ffdumas-01243535f

[04] Sophie Lemonnier. L'allocation de l'attention visuelle lors d'une situation multi-tâche et dynamique : l'approche de carrefour en conduite.. Psychologie. Université paris 8, 2011. Français.

[05] Baluch, F., & Itti, L. Mechanisms of top-down attention. Trends in Neuroscience. (2011) .

[06] Chun, M. M., & Jiang, Y. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive Psychology . (1998) .

[07] Treisman, A. M., & Gelade, A feature-integration theory of attention. Cognitive Psychology . G. (1980).

[08] Summer_eld, C., & Egner, Expectation (and attention) in visual cognition. Trends in Cognitive Science, T. (2009).

[09] BABAHENINI . S, L'apport de la perception/l'attention visuelle à l'amélioration de la fusion d'images multi-focus ,Mémoire de Doctorat LMD, Université Mohamed Khider – BISKRA,2021.

[10] Mokri M, Classification des images avec les réseaux de neurones convolutionnels , Mémoire de PFE, Université de Tlemcen, 2017.

[11] https://www.police-scientifique.com/photographie/image-numerique (consulté le 10/04/2022).

[12] Jean marie. La liaison automatique des plusieurs images perçues sur un scanner ISP(Institut Supérieur Pédagogique de Bukavu) - licencié en pédagogie; Option : Informatique de Gestion 2008.

[13] N MERABET,M MAHLIA, recherche d'images par le contenu, université abou bakrbelkaid– tlemcen.2011.

[14] D. Zeroual, "Implémentation d'un environnement parallèle pour la compression d'images a l'aide des fractales", Thèse de magister en informatique, Université de Batna, 2006.

[15] KAAZAOUI .A et KAAZAOUI.K, La fusion d'image multi-focus, Mémoire de PFE, Université Ahmed Draia – Adrar 2017.

[16] Shutao Li, James T Kwok, and Yaonan Wang. Combination of images with diverse focuses using the spatial frequency. Information fusion, 2(3):169–176, 2001.

[17] Wei Huang and Zhongliang Jing. Evaluation of focus measures in multi-focus image fusion. Pattern recognition letters, 28(4):493–500, 2007.

[18] R.C.Gonzalez, R.E.Woods, Digital Image Processing, 2nd ed., Prentice Hall.2001

[19] RajaKumari K1, Anu Priya S, Survey on contourlet based fusion techniques for multimodality image fusion,International Journal of Computer Engineering and Applications, Volume IX, Issue III, March 15 www.ijcea.comISSN 2321

[20] P.T. Burt and E.H. Andelson, "The Laplacian pyramid as a compact Image code," IEEE Transactions on Communications, , 1983

[21] https://datascientest.com/intelligence-artificielle-definition (consulté le 20/04/2022)

[22] https://www.oracle.com/dz/data-science/machine-learning/what-is-machine-learning/ (consulté le 22/04/2022)

[23] Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), Université d'Angers et Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS), Lyon.2021

[24] BARREIRO LINDO, Interprétation d'images basée sur la technologie des réseaux de neurones, Bachelor HES, Haute École de Gestion de Genève (HEG-GE) 2018

[25] SANNY Abdul-Qadir, Système d'aide à l'estimation du rendement agricole par une méthode de deeplearning : cas de la tomate, Mémoire de PFE, Université d'Abomey-Calavi 2021.

[26] Tetbirt A et Khelifi D, Fusion d'images Basée sur l'apprentissage profond, Mémoire de PFE, Université SAAD DAHLAB de BLIDA 2020.