الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière Électronique
Spécialité Électroniques des Systèmes Embarqués

présenté par

ATIF RAHIL

&

MISSERAOUI NACERA

Evaluation d'un algorithme de Deep learning pour l'amélioration du taux de détection d'objets multiples dans des scènes denses

Proposé par : Mme D. Naceur

Année Universitaire 2020-2021

Remerciements

En préambule à ce mémoire ; Louange à DIEU le tout puissant, le miséricordieux, de nous avoir donné le courage, la force, la santé et la persévérance pour pouvoir effectuer ce travail dans de meilleures conditions.

Ce mémoire n'aurait pas été possible sans l'intervention, consciente, d'un grand nombre de personnes.

Nos remerciements s'adressent en premier lieu à notre promotrice, M^{me} D.NACEUR., pour son amabilité, sa bienveillance et pour son encouragement constant. Elle a dirigé et accompagné de très près, et avec beaucoup de patience, notre travail.

Nos remerciements vont ensuite aux membres du jury, qui ont accepté, sans réserve aucune de lire et d'évaluer ce mémoire à sa juste valeur, et de nous faire part de leur remarques sûrement pertinentes qui, avec un peu de recul, contribueront, sans nul doute, au perfectionnement du présent travail.

Nous tenons enfin à exprimer notre reconnaissance à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce modeste travail ainsi qu'à toute personne qui fera l'effort de lire ce document.

Pour la même occasion, on adresse nos remerciements à tous nos enseignants pour leurs efforts qui ont guidé nos pas et enrichi nos travaux tout au long de nos études universitaires.

Dédicaces

Je dédie le fruit de mes cinq années d'études

Au meilleur des pères Khaled

Autant de phrases ne seraient exprimer ma gratitude et ma reconnaissance. Tu as su m'inculquer le sens de la responsabilité, de l'optimisme et de la confiance en soi face aux difficultés de la vie. En lui transmettant mes plus vifs remerciements pour la peine qu'il s'est donné et l'aide qu'il m'a apporté pour la réalisation et le déroulement de mes études.

A la meilleure des mères Nadia

Affable, honorable, aimable: Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi. Ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études. Tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études.

Qu'ils trouvent en moi la source de leur fierté

A mes grands parents

Qui m'ont accompagné par leurs prières. Puisse ALLAH leur prêter longue vie et beaucoup de santé et de bonheur.

A ma chère sœur Sarah

Ma fidèle accompagnante dans les moments les plus délicats de cette vie mystérieuse. Les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour vous. Tes sacrifices, ton soutien moral, ta gentillesse sans égal.

A ma petite sœur Lina

La douce, au cœur si grand. Aucune dédicace ne saurait exprimer tout l'amour que j'ai pour vous, votre joie et votre gaieté me comblent de bonheur.

Puisse ALLAH vous gardez et vous aidez à réaliser à votre tour vos vœux les plus chers.

A mes chers tantes et oncles, à mes chers cousines et cousins

Veuillez trouver dans ce travail l'expression de mon respect le plus profond et mon affection la plus sincère.

A mon binôme Nacéra

Qui m'a supporté durant la réalisation de ce travail. Et en souvenir de notre sincère et profonde amitié et des moments agréables que nous avons passés ensemble.

A mes chers amis

Achouak, Hadjer, Kaouther, Naouel et Yassmine,
Aymen, Fouad, Mahmoud et Yassine,
Pour leurs indéfectible soutien et en témoignage de l'amitié qui nous uni et des souvenirs de tous les moments que nous avons passé ensemble
et à qui je souhaite un avenir radieux plein de réussite et de bonheur.

A mes chers collègues

Fella, Nermine, Ahmed, Nadir et Oussama.

Sans vos aides, vos conseils et vos encouragements ce travail n'aurait vu le jour.

Rahil

Dédicaces

En premier lieu, je remercie ALLAH de m'avoir permis d'acquérir une infime partie de sa science sans limite.

Je dédie ce modeste travail, tout d'abord à

Mes chers parents

Pour leurs sacrifices et leur encouragement, qui m'ont suivi pendant toute ma carrière d'étudiant.

A mes frères

Mohamed et Moussa

A ma sœur

Fatima Zohra

A mes neveux

Ritadje, Boualem, Roudaina, Sara et Haithem.

A tous mes enseignants

Spécialement mon encadreur Mme Naceur.

A mon binôme Rahil

Ma meilleure amie que je la prends non seulement pour amie mais pour la sœur d'âme, avec laquelle j'ai vécu les meilleurs souvenirs.

A mes chères amis et collègues

Yasmine, Rania, Fella, Nermine et Rahil

Ahmed, Nadir et Sidali

Nacéra

Résumé

La détection des visages est devenue l'une des méthodes les plus populaires dans la détection d'objets. L'application de la détection des visages dans des scènes réelles est toujours confrontée à certains défis en raison de l'influence de facteurs complexes. Les techniques d'apprentissage en profondeur sont apparues comme une stratégie puissante qui a fourni une solution alternative pour imiter la vision humaine. Ces techniques sont basées sur des réseaux neuronaux tels que les réseaux neuronaux convolutifs (CNN).

Notre approche consiste à détecter des visages aussi lointains que possible dans des scènes denses et en temps réel, où de nombreuses interférences limitent l'efficacité des informations pour la détection.

Pour ce faire, on a adopté le modèle YOLO_Face pour la détection, une étude comparative a été faite avec le modèle YOLO5Face.

Mots clés : détection d’objets, détection de visages, vision par ordinateur, apprentissage profond, YOLO, temps réel, scènes denses, OpenCV.

Abstract

Face detection has become one of the most popular methods in object detection. The application of face detection in real scenes still faces some challenges due to the influence of complex factors. Deep learning techniques have emerged as a powerful strategy that has provided an alternative solution to mimic human vision. These techniques are based on neural networks such as convolutional neural networks (CNN).

Our approach is to detect faces as far away as possible in dense, real-time scenes, where many interferences limit the effectiveness of information for detection.

For this purpose, the YOLO_Face model was adopted for detection, a comparative study was made with the YOLO5Face model.

Keywords: object detection, face detection, computer vision, deep learning, YOLO, real time, dense scenes, OpenCV.

الملخص

أصبح الكشف عن الوجه أحد أكثر الطرق شعبية في الكشف عن الأجسام. لا يزال تطبيق كشف الوجه في المشاهد الحقيقية يواجه بعض التحديات بسبب تأثير العوامل المعقدة. ظهرت تقنيات التعلم العميق كاستراتيجية قوية توفر حلا بديلا لتقليد الرؤية البشرية. وتستند هذه التقنيات إلى الشبكات العصبية مثل الشبكات العصبية المعقدة (CNN).

يتمثل نهجنا في كشف الوجوه البعيدة قدر الإمكان في المشاهد الكثيفة والآنية ، حيث تحد العديد من التدخلات من فعالية المعلومات للكشف.

للقيام بذلك، اعتمدنا النموذج YOLO_Face للكشف، أجريت دراسة مقارنة مع النموذج YOLO5Face.

الكلمات الرئيسية: كشف الأجسام، كشف الوجه، رؤية الكمبيوتر، التعلم العميق،YOLO ، المشاهد الآنية الكثيفة، OpenCV.

Liste des abréviations

3D: trois dimensions

ALPAC: Automatic Language Processing Advisory Committee

BN: Batch Normalization

CNN: Convolutional Neural Network

Convnet: Convolutional Network

CPU: Central Processing Unit

DARPA: Defense Advanced Research Projects Agency

DART: Des Moines Area Regional Transit Authority

DL: Deep Learning

fc: fully connected

FGCS: Fifth Generation Computer Systems

fps: frames per second

GPU: Graphical Processing Unit

GPS: General Problem Solver

IA: Intelligence Artificielle

IBM: International Business Machines

iOS: iPhone Operating System

LT: Logic Theorist

MIT: Massachusetts Institute of Technology

ML: Machine Learning

MLP: Multilayer perceptron

NMS: Non-Max suppression

OS: Operating System

ReLU: Rectified Linear Unit

RNN: Recurrent neural network

RVB: Rouge, vert, bleu

SNARC: Stochastic Neural Analog Reinforcement Calculator

SPP-Net: Spatial Pyramid Pooling Network

SSD: Single shot detector

SVM: Support vector machine

YOLO: You Only Look Once

Table des matières

Liste des figures

Introduction générale

Pendant de nombreuses décennies, l'homme rêvait de créer des machines dotées des caractéristiques de l'intelligence humaine, capables de penser et d'agir comme les humains. L'une des idées les plus fascinantes était de donner aux ordinateurs la capacité de "voir" et d'interpréter le monde qui les entoure telles que la détection et le suivi d'objets.

Grâce aux progrès de l'intelligence artificielle (IA), la fiction d'hier est devenue la réalité d'aujourd'hui. La détection automatisée des visages a attiré l'attention dans le domaine de la vision par ordinateur et de la reconnaissance des formes, dont des développements importants ont été réalisés, mais la vision par ordinateur reste un champ d'investigation très actif avec de nombreux problèmes difficiles et non entièrement résolus.

Les premiers systèmes de détection des visages ne pouvaient traiter que des cas simples, mais ils sont désormais plus performants dans diverses situations grâce aux algorithmes d'apprentissage profond qui est la technologie la plus prometteuse d'apprentissage automatique. Il utilise une architecture de neurones artificiels connectés entre eux, inspirée de celle du cerveau. Ce réseau est capable de traiter une grande quantité d'informations et d'apprendre progressivement à partir d'images, textes ou données.

La plupart de ces progrès sont le résultat d'un matériel plus puissant, des jeux de données plus volumineux et de modèles plus grands et une conséquence de nouvelles idées, d'algorithmes et d'architectures réseau améliorées.

En revanche, la précision, le temps de formation et le temps de traitement des vidéos en temps réel pour la détection des visages sont encore des questions de recherche.

L'objectif principal de notre modèle est la détection de visages en temps réel dans des scènes denses, dont nombreuses interférences et bruits (tels que l'occlusion, l'éclairage et la faible résolution) limitent l'efficacité des informations pour la détection des visages.

Notre travail est réparti comme suit :

Dans le premier chapitre, nous présentons un état de l'art de l'intelligence artificielle et de ses sous-domaines (apprentissage automatique et profond), ainsi que la vision par ordinateur.

Dans le deuxième chapitre, nous introduisons les généralités sur les réseaux de neurones convolutifs, sur la détection d'objets, la reconnaissance et la détection de visages et nous terminons par l'étude du modèle choisi.

Dans le troisième chapitre, nous présentons la conception et l'implémentation de notre modèle ainsi que les résultats obtenus et nous étudions les performances de notre modèle en termes de taux de détection et de temps de calcul.

Etat de l’art

Dans sa forme la plus simple, l'intelligence artificielle est un domaine qui combine l'informatique et des ensembles de données robustes pour permettre la résolution de problèmes. Elle englobe également les sous-domaines de l'apprentissage automatique et de l'apprentissage profond, qui sont fréquemment mentionnés en association avec elle [1].

Ces disciplines sont composées d'algorithmes d'IA qui cherchent à créer des systèmes experts qui font des prédictions ou des classifications sur la base de données d'entrée [1].

1.1 Intelligence artificielle

L'intelligence artificielle (IA), est l'un des domaines les plus anciens de l'informatique (plus de 200 ans), et très vastes. Elle est un ensemble de techniques permettant à des machines d'accomplir des tâches et de résoudre des problèmes normalement réservés aux humains *(selon Y. LeCun, présentation au Collège de France)*.

Ces tâches sont parfois très simples pour les humains, moins pour les machines comme reconnaître et localiser les objets dans une image, planifier les mouvements d'un robot pour attraper un objet ou conduire une voiture. Elles requièrent parfois de la planification complexe, comme par exemple jouer aux échecs ou au jeu de go. Les tâches les plus compliquées nécessitent beaucoup de connaissances et de sens commun.

L'objectif étendu de l'intelligence artificielle a donné lieu à de nombreuses questions et débats. À tel point qu'aucune définition unique du domaine n'est universellement acceptée [2].

1.1.1 Historique

Voici un bref aperçu de certains des événements les plus importants de l'IA [2].

1940s

- (1943) Warren McCullough et Walter Pitts publient "A Logical Calculus of Ideas Immanent in Nervous Activity", cet article propose le premier modèle mathématique pour la construction d'un réseau neuronal.
- (1949) Dans son livre The Organization of Behavior : A Neuropsychological Theory, Donald Hebb propose la théorie selon laquelle les chemins neuronaux sont créés à partir d'expériences et que les connexions entre les neurones se renforcent au fur et à mesure qu'elles sont utilisées. L'apprentissage hébbien reste un modèle important en IA.

1950s

- (1950) Alan Turing publie "Computing Machinery and Intelligence", proposant ce que l'on appelle aujourd'hui le test de Turing, une méthode permettant de déterminer si une machine est intelligente.
- (1950) Marvin Minsky et Dean Edmonds, étudiants de premier cycle à Harvard, construisent SNARC, le premier ordinateur à réseau neuronal.
- (1950) Claude Shannon publie l'article "Programming a Computer for Playing Chess".
- (1950) Isaac Asimov publie les "Trois lois de la robotique".
- (1952) Arthur Samuel développe un programme d'auto-apprentissage pour jouer aux dames.
- (1954) L'expérience de traduction automatique Georgetown-IBM traduit automatiquement en anglais 60 phrases russes soigneusement sélectionnées.
- (1956) L'expression "intelligence artificielle" est inventée lors du "Dartmouth Summer Research Project on Artificial Intelligence". Dirigée par John McCarthy, cette conférence, qui a défini la portée et les objectifs de l'IA, est largement considérée comme la naissance de l'intelligence artificielle telle que nous la connaissons aujourd'hui.
- (1956) Allen Newell et Herbert Simon présentent Logic Theorist (LT), le premier programme de raisonnement.

- (1958) John McCarthy développe le langage de programmation d'IA Lisp et publie l'article "Programs with Common Sense". L'article propose l'hypothétique Advice Taker, un système d'IA complet capable d'apprendre par l'expérience aussi efficacement que les humains.
- (1959) Allen Newell, Herbert Simon et J.C. Shaw développent le General Problem Solver (GPS), un programme conçu pour imiter la résolution de problèmes humains.
- (1959) Herbert Gelernter développe le programme Geometry Theorem Prover.
- (1959) Arthur Samuel invente le terme d'apprentissage automatique alors qu'il travaille chez IBM.
- (1959) John McCarthy et Marvin Minsky fondent le projet d'intelligence artificielle du MIT.

1960s

- (1963) John McCarthy crée le AI Lab à Stanford.
- (1966) Le rapport du Comité consultatif sur le traitement automatique des langues (ALPAC) du gouvernement américain décrit en détail le manque de progrès dans la recherche sur la traduction automatique, une initiative majeure de la guerre froide qui promettait la traduction automatique et instantanée du russe. Le rapport ALPAC conduit à l'annulation de tous les projets de traduction automatique financés par le gouvernement.
- (1969) Les premiers systèmes experts réussis sont développés à Stanford avec DENDRAL, un programme XX, et MYCIN, conçu pour diagnostiquer les infections sanguines.

1970s

- (1972) Le langage de programmation logique PROLOG est créé.
- (1973) Le "Rapport Lighthill", détaillant les déceptions de la recherche sur l'IA, est publié par le gouvernement britannique et conduit à de sévères réductions du financement des projets d'intelligence artificielle.
- (1974-1980) La frustration liée aux progrès du développement de l'IA entraîne une réduction importante des subventions universitaires par la DARPA. Si l'on

ajoute à cela le rapport ALPAC et le "Rapport Lighthill" de l'année précédente, le financement de l'intelligence artificielle se tarit et la recherche stagne. Cette période est connue comme le "premier hiver de l'IA".

1980s

- (1980) Digital Equipment Corporations développe R1 (également connu sous le nom de XCON), le premier système expert commercial réussi. Conçu pour configurer les commandes de nouveaux systèmes informatiques, R1 donne le coup d'envoi d'un boom des investissements dans les systèmes experts qui durera une grande partie de la décennie, mettant ainsi fin au premier "hiver de l'IA".
- (1982) Le ministère japonais du commerce international et de l'industrie lance l'ambitieux projet de systèmes informatiques de cinquième génération. L'objectif du FGCS est de développer des performances similaires à celles des superordinateurs et une plate-forme pour le développement de l'IA.
- (1983) En réponse au projet japonais FGCS, le gouvernement américain lance l'Initiative de calcul stratégique (Strategic Computing Initiative) afin de financer la recherche sur le calcul avancé et l'intelligence artificielle par la DARPA.
- (1985) Les entreprises dépensent plus d'un milliard de dollars par an pour les systèmes experts et une industrie entière, connue sous le nom de marché des machines Lisp, se développe pour les soutenir. Des entreprises telles que Symbolics et Lisp Machines Inc. construisent des ordinateurs spécialisés qui fonctionnent avec le langage de programmation Lisp.
- (1987-1993) Au fur et à mesure que la technologie informatique s'améliore, des alternatives moins coûteuses apparaissent et le marché des machines Lisp s'effondre en 1987, marquant le début du "second hiver de l'IA". Au cours de cette période, les systèmes experts s'avèrent trop coûteux à entretenir et à mettre à jour, et finissent par tomber en disgrâce.

1990s

- (1991) Les forces américaines déploient DART, un outil automatisé de planification logistique et d'ordonnancement, pendant la guerre du Golfe.

- (1992) Le Japon met fin au projet FGCS en 1992, invoquant l'incapacité à atteindre les objectifs ambitieux définis dix ans plus tôt.
- (1993) La DARPA met fin à l'Initiative de calcul stratégique en 1993 après avoir dépensé près d'un milliard de dollars et n'avoir pas répondu aux attentes.
- (1997) Deep Blue d'IBM bat le champion du monde d'échecs Gary Kasparov.

2000s

- (2005) STANLEY, une voiture à conduite autonome, remporte le Grand Challenge de la DARPA.
- (2005) L'armée américaine commence à investir dans des robots autonomes comme le "Big Dog" de Boston Dynamics et le "PackBot" d'iRobot.
- (2008) Google réalise des percées dans le domaine de la reconnaissance vocale et introduit cette fonctionnalité dans son application iPhone.

2010-2014

- (2011) Watson d'IBM bat ses concurrents à Jeopardy !
- (2011) Apple lance Siri, un assistant virtuel alimenté par l'IA, dans son système d'exploitation iOS.
- (2012) Andrew Ng, fondateur du projet Google Brain Deep Learning, alimente un réseau neuronal utilisant des algorithmes d'apprentissage profond avec 10 millions de vidéos YouTube comme ensemble d'entraînement. Le réseau neuronal a appris à reconnaître un chat sans qu'on lui dise ce qu'est un chat, inaugurant l'ère de la percée pour les réseaux neuronaux et le financement de l'apprentissage profond.
- (2014) Google fabrique la première voiture à conduite autonome qui réussit un test de conduite dans un État.
- (2014) Sortie d'Alexa, la maison virtuelle d'Amazon.

2015-2021

- (2016) AlphaGo, de Google DeepMind, bat le champion du monde de go, Lee Sedol. La complexité de l'ancien jeu chinois était considérée comme un obstacle majeur à franchir pour l'IA.

- (2016) Le premier "robot citoyen", un robot humanoïde nommé Sophia, est créé par Hanson Robotics et est capable de reconnaissance faciale, de communication verbale et d'expression faciale.
- (2018) Google publie le moteur de traitement du langage naturel BERT, réduisant les obstacles à la traduction et à la compréhension par des applications d'apprentissage automatique.
- (2018) Waymo lance son service Waymo One, permettant aux utilisateurs de toute la région métropolitaine de Phoenix de demander une prise en charge par l'un des véhicules à conduite autonome de l'entreprise.
- (2020) Baidu met son algorithme d'IA LinearFold à la disposition des équipes scientifiques et médicales qui travaillent à l'élaboration d'un vaccin pendant les premiers stades de la pandémie de SRAS-CoV-2. L'algorithme est capable de prédire la séquence d'ARN du virus en seulement 27 secondes, soit 120 fois plus vite que les autres méthodes.

1.1.2 L'intelligence artificielle et l'homme

L'intelligence artificielle n'est pas là pour nous remplacer. Elle augmente nos capacités et nous rend meilleurs dans ce que nous faisons. Les algorithmes d'IA apprennent différemment des humains, ils regardent les choses différemment. Ils peuvent voir des relations et des modèles qui nous échappent. Ce partenariat entre l'homme et l'IA offre de nombreuses possibilités [3]. Il peut :

- Apporter l'analytique aux industries et aux domaines où elle est actuellement sous-utilisée.
- Améliorer les performances des technologies analytiques existantes, comme la vision par ordinateur et l'analyse des séries chronologiques.
- Augmenter les capacités existantes et nous rendre meilleurs dans ce que nous faisons.
- Nous donner une meilleure vision, une meilleure compréhension, une meilleure mémoire et bien plus encore.

1.1.3 Les principaux sous-domaines de l'intelligence artificielle

L'intelligence artificielle travaille avec de grandes quantités de données qui sont d'abord combinées avec un traitement rapide et itératif et des algorithmes intelligents [4].

Les principaux sous-domaines de l'IA sont illustrés sur la figure 1.2.1:

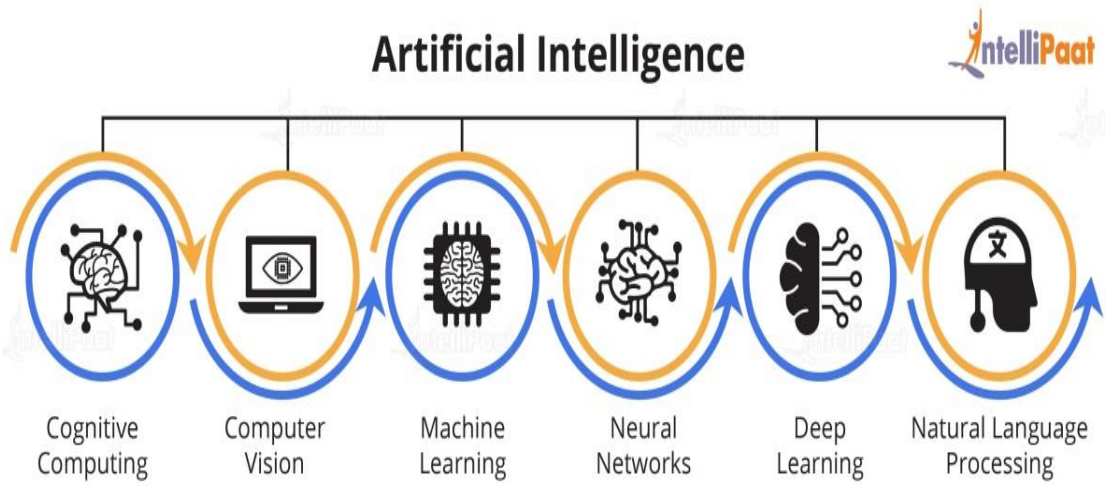

Figure 1.1.1 Les principaux sous-domaines de l'IA [5].

L'IA concerne bien plus le processus et la capacité de réflexion et d'analyse de données surpuissantes que tout format ou fonction particulière. Bien que l'IA évoque des images de robots très performants, semblables à des humains, qui envahissent le monde, elle n'est pas destinée à remplacer les humains. Elle est destinée à améliorer considérablement les capacités et les contributions humaines. Cela en fait un atout commercial très précieux [6].

1.2 Apprentissage automatique

L'IA est devenue un sujet en vogue dans les médias et magazines scientifiques en raison des nombreuses réalisations, dont beaucoup sont le fruit des progrès accomplis dans le domaine de l'apprentissage automatique ou machine learning (ML) [7].

La caractéristique idéale de l'IA est sa capacité à rationaliser et à prendre les mesures qui ont les meilleures chances d'atteindre un objectif spécifique. Un sous-ensemble de l'IA est l'apprentissage automatique, qui fait référence au concept selon

lequel les programmes informatiques peuvent automatiquement apprendre par expérience et acquérir des compétences sans intervention humaine [7].

1.2.1 Qu'est-ce que l'apprentissage automatique ?

L'apprentissage automatique est une application de l'IA qui donne aux systèmes la capacité d'explorer, d'améliorer et de perfectionner automatiquement les différentes expériences sans être programmés [4].

Son aspect itératif est important, car lorsque les modèles sont exposés à de nouvelles données, ils sont capables de s'adapter de manière indépendante [8]. Ils apprennent automatiquement à partir de données antérieures pour acquérir des connaissances par l'expérience, améliorer progressivement son comportement d'apprentissage et produire des décisions et des résultats fiables et reproductibles [9].

1.2.2 Évolution de l'apprentissage automatique

Bien que l'apprentissage automatique ne soit pas nouveau, sa définition précise est encore confuse pour de nombreuses personnes.

Au cours des dernières décennies, l'apprentissage automatique est devenu beaucoup plus efficace et largement disponible. Son concept remonte du milieu du 20ème siècle, dont le mathématicien britannique Alan Turing a imaginé une machine capable d'apprendre [10].

Et avec le temps, différentes techniques de ML ont été développées pour créer des algorithmes capables d'apprendre et de s'améliorer de manière autonome, mais le développement était limité par le manque d'ensembles de données disponibles, et par son incapacité à analyser des quantités massives de données en quelques secondes. Aujourd'hui, des données sont accessibles en temps réel à tout moment [10].

Il s'agit notamment de l'augmentation du volume et de la diversité des données disponibles, de la baisse du coût et de la puissance du traitement informatique et du coût du stockage des données. Tous ces éléments signifient qu'il est possible de produire rapidement et automatiquement des modèles capables d'analyser des données plus importantes et plus complexes et de fournir des résultats plus rapides et plus précis, même à très grande échelle [8].

1.2.3 Fonctionnement de l'apprentissage automatique

L'apprentissage automatique se compose de différents types de modèles d'apprentissage automatique, utilisant diverses techniques algorithmiques. Selon la nature des données et le résultat souhaité, l'un des quatre modèles d'apprentissage suivants peut être utilisé: supervisé, non supervisé, semi-supervisé ou par renforcement. Dans chacun de ces modèles, une ou plusieurs techniques algorithmiques peuvent être appliquées en fonction des ensembles de données utilisés et des résultats prévus. Les algorithmes peuvent être utilisés un par un ou combinés pour obtenir la meilleure précision possible lorsque des données complexes et imprévisibles sont en jeu [11].

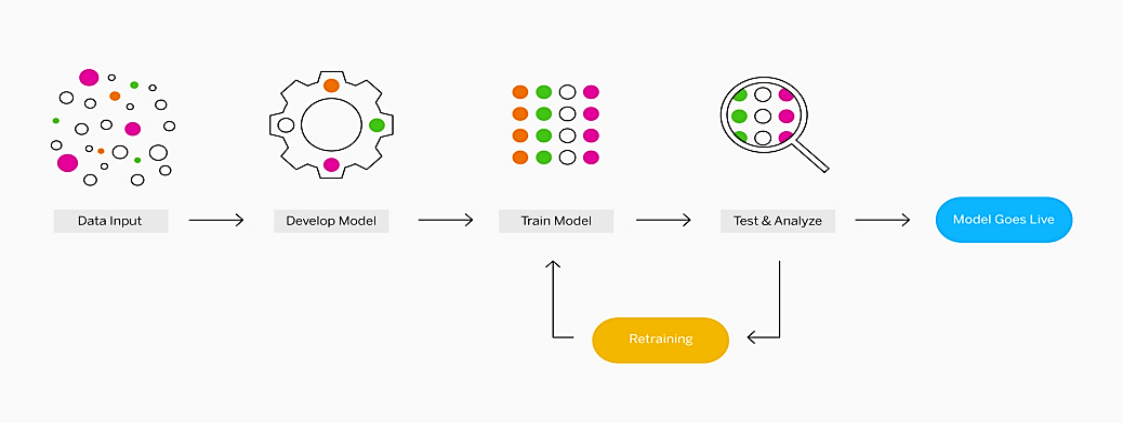

Figure 1.2.1 Comment fonctionne le processus d'apprentissage automatique [11].

1.2.4 Les méthodes d'apprentissage automatique les plus populaires

Deux des méthodes d'apprentissage automatique les plus répandues sont l'apprentissage supervisé et l'apprentissage non supervisé.

a Apprentissage supervisé

L'apprentissage supervisé est considéré comme la classe la plus élémentaire d'algorithmes d'apprentissage automatique [12].

L'apprentissage supervisé, se définit par l'utilisation d'ensembles de données étiquetées pour former des algorithmes capables de classer des données ou de prédire des résultats avec précision. Au fur et à mesure que les données d'entrée sont

introduites dans le modèle, celui-ci ajuste ses poids jusqu'à ce que le modèle soit ajusté de manière appropriée [13].

L'apprentissage supervisé aide les organisations à résoudre une variété de problèmes du monde réel à l'échelle, comme la classification des spams dans un dossier distinct de votre boîte de réception. Parmi les méthodes utilisées dans l'apprentissage supervisé figurent les réseaux neuronaux, la régression, la machine à vecteurs de support (SVM), etc. [13].

b Apprentissage non-supervisé

La deuxième classe d'algorithmes d'apprentissage automatique est appelée apprentissage non supervisé, dans ce cas, nous n'étiquetons pas les données au préalable, nous laissons plutôt l'algorithme arriver à sa conclusion. Ce type d'apprentissage est important car il est beaucoup plus commun dans le cerveau humain que l'apprentissage supervisé. Les algorithmes d'apprentissage non supervisé sont particulièrement utilisés dans les problèmes de clustering, dans lesquels, étant donné une collection d'objets, nous voulons être en mesure de comprendre et de montrer leurs relations [10].

1.2.5 Applications d'apprentissage automatique

Les applications de l'apprentissage automatique sont présentes partout, dans les domaines de la science, de l'ingénierie et des affaires, et permettent de prendre des décisions plus fondées sur des preuves [9].

1.3 Apprentissage profond

L'apprentissage profond ou deep learning (DL) est un type particulier d'apprentissage automatique qui atteint une grande puissance et une grande flexibilité en apprenant à représenter le monde comme une hiérarchie imbriquée de concepts, chaque concept étant défini par rapport à des concepts plus simples, et les représentations plus abstraites étant calculées en fonction de représentations moins abstraites [14].

Au sein de la révolution que connait l'Intelligence Artificielle aujourd'hui, l'apprentissage profond occupe une place de tout premier plan. Il est au cœur des technologies qui permettent aujourd'hui de réaliser des tâches encore impensables il y a quelques années [7].

1.3.1 Qu'est-ce que l'apprentissage profond ?

L'apprentissage profond est un système avancé inspiré du cerveau humain, qui comporte un vaste réseau de neurones artificiels. Ces neurones sont interconnectés pour traiter et mémoriser des informations, comparer des problèmes ou situations quelconques avec des situations similaires passées, analyser les solutions et résoudre le problème de la meilleure façon possible [15].

1.3.2 Fonctionnement de l'apprentissage profond

Ce type d'apprentissage automatique est appelé "profond", dont il tient son nom de la profondeur des couches des réseaux neuronaux.

Ces réseaux neuronaux profonds se composent de plusieurs de dizaines voire de centaines de couches de nœuds interconnectés, chacune s'appuyant sur la couche précédente pour affiner et optimiser la prédiction ou la catégorisation. Cette progression des calculs dans le réseau est appelée propagation vers l'avant. Les couches d'entrée et de sortie d'un réseau neuronal profond sont appelées couches visibles. La couche d'entrée est celle où le modèle d'apprentissage profond ingère les données à traiter, et la couche de sortie est celle où la prédiction ou la classification finale est effectuée [16].

Toutes les couches entre les deux sont appelées couches cachées (Hidden layers). Plus on augmente le nombre de couches, plus les réseaux de neurones apprennent des choses compliquées, abstraites, correspondant de plus en plus à la manière dont un humain raisonne [10].

Un autre processus appelé rétropropagation utilise des algorithmes, comme la descente de gradient, pour calculer les erreurs dans les prédictions, puis ajuste les poids et les biais de la fonction en remontant les couches dans le but d'entraîner le modèle. Ensemble, la propagation vers l'avant et la rétropropagation permettent à un

réseau neuronal de faire des prédictions et de corriger les erreurs en conséquence. Au fil du temps, l'algorithme devient progressivement plus précis [16].

Comme exemple, un système d'apprentissage profond qui traite des images de la nature et recherche des marguerites Gloriosa reconnaîtra à la première couche, une plante. Au fil des couches neuronales, il identifiera ensuite une fleur, puis une marguerite, et enfin une marguerite Gloriosa. Parmi les exemples d'applications d'apprentissage profond, citons la reconnaissance vocale, la classification d'images et l'analyse pharmaceutique [11].

Ce qui précède décrit le type le plus simple de réseau neuronal profond dans les termes les plus simples. Cependant, les algorithmes d'apprentissage profond sont incroyablement complexes, et il existe différents types de réseaux neuronaux pour répondre à des problèmes ou des ensembles de données spécifiques [16].

- Perceptrons multicouches (MLP)
- Réseaux de neurones convolutifs (CNN)
- Réseaux de neurones récurrents (RNN)

1.3.3 Applications d'apprentissage profond

L'un des principaux avantages d'apprentissage profond est l'analyse et l'apprentissage des quantités massives de données, ce qui lui a permis de pouvoir dans tous les domaines [10].

Les applications d'apprentissage profond dans le monde réel font partie de notre quotidien, mais dans la plupart des cas, elles sont si bien intégrées dans les produits et services que les utilisateurs n'ont pas conscience du traitement complexe des données qui se déroule en arrière-plan [16].

L'apprentissage profond a permis d'obtenir des résultats impressionnants dans des domaines aussi nombreux que variés

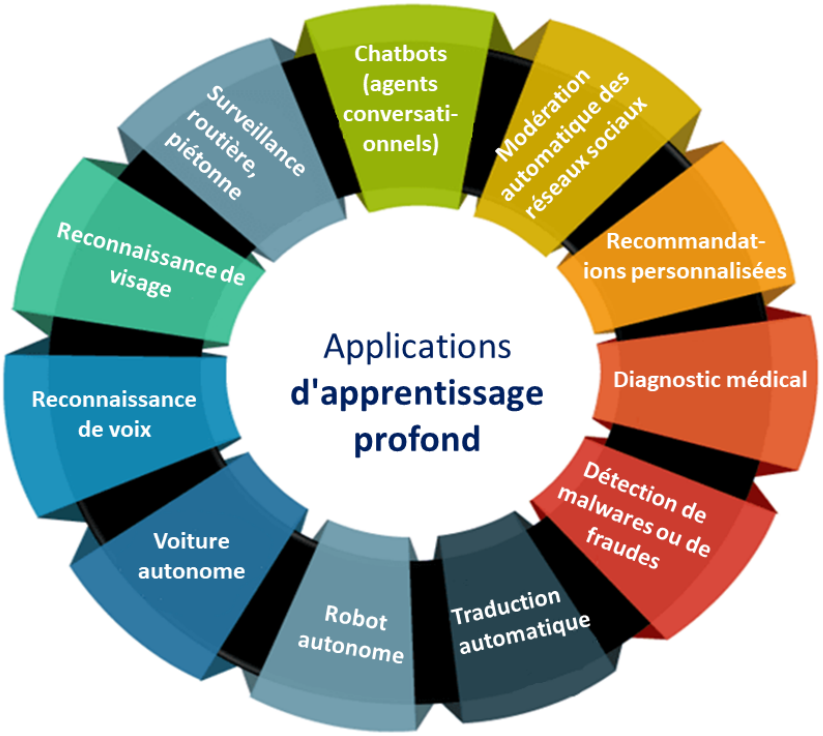

Figure 1.3.1 Applications d'apprentissage profond.

1.4 Vision par ordinateur

Dans le domaine de l'intelligence artificielle, la vision par ordinateur est l'une des tâches les plus intéressantes et les plus stimulantes. La vision par ordinateur agit comme un pont entre les logiciels informatiques et les visualisations qui nous entourent [17].

Elle se concentre sur la création de systèmes numériques capables de traiter, d'analyser et de donner un sens aux données visuelles (images ou vidéos) de la même manière que les humains. Le concept de vision par ordinateur repose sur l'apprentissage des ordinateurs à traiter une image au niveau du pixel et à la comprendre. Techniquement, les machines tentent de récupérer les informations visuelles, de les traiter et d'interpréter les résultats grâce à des algorithmes logiciels spéciaux [18].

La vision par ordinateur fonctionne à peu près comme la vision humaine, sauf que les humains ont une longueur d'avance. La vision humaine a l'avantage de pouvoir s'entraîner, à distinguer les objets, à déterminer leur distance, à savoir s'ils sont en mouvement et si quelque chose ne va pas dans une image [19].

En revanche, il faut beaucoup de temps et de données d'entraînement pour qu'une machine puisse identifier ces objets. Mais avec les récentes avancées en matière de matériel et de l'intelligence artificielle, ce domaine de la vision par ordinateur est devenu beaucoup plus facile et plus intuitif [20].

Si l'IA permet aux ordinateurs de penser, la vision par ordinateur leur permet de voir, d'observer et de comprendre [19].

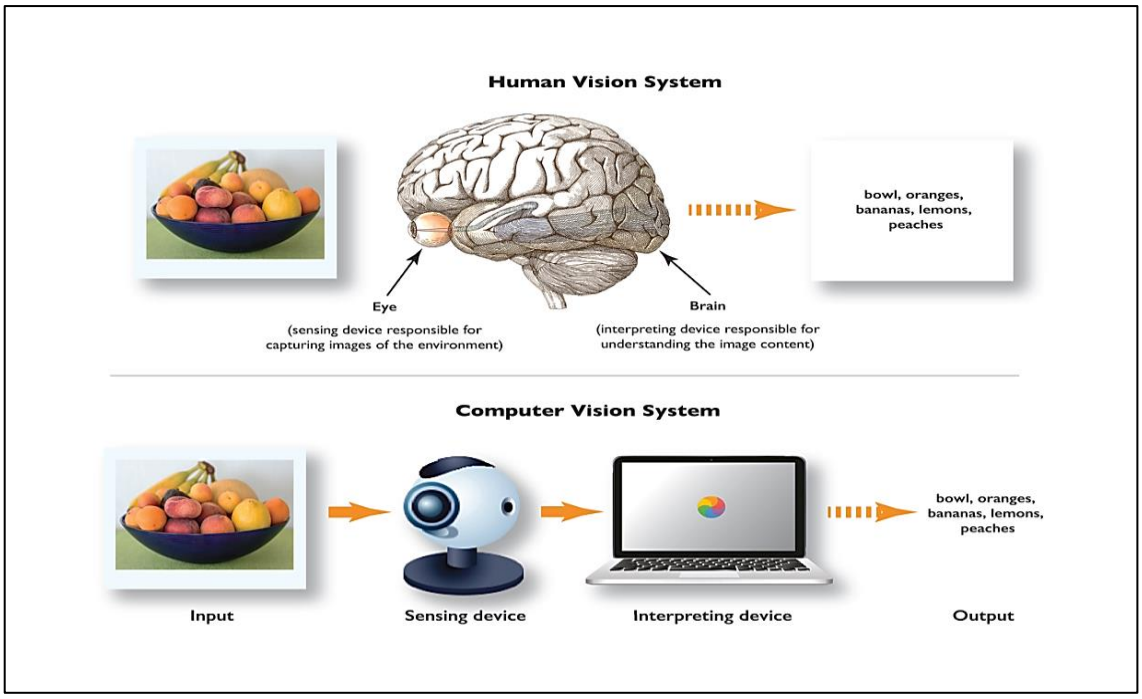

Figure 1.4.1 Fonctionnement de la vision par ordinateur par rapport à la façon dont les humains traitent les données visuelles [21].

1.4.1 Fonctionnement de la vision par ordinateur

La technologie de vision par ordinateur tend à imiter la façon dont le cerveau humain fonctionne. Mais comment notre cerveau parvient-il à reconnaître les objets visuels ? Selon l'une des hypothèses les plus répandues, notre cerveau s'appuie sur des motifs pour décoder les objets individuels. Ce concept est utilisé pour créer des systèmes de vision par ordinateur [21].

Les algorithmes de vision par ordinateur que nous utilisons aujourd'hui sont basés sur la reconnaissance des formes. Nous entraînons les ordinateurs sur une quantité massive de données visuelles - les ordinateurs traitent les images, étiquettent les objets qu'elles contiennent et trouvent des motifs dans ces objets [21].

La vision par ordinateur peut émuler des tâches humaines de base telles que la détection et la reconnaissance des visages ou la détection des objets [22]. Les éléments de la figure 1.5.2 sont les principaux composants de la vision par ordinateur qui sont largement utilisés :

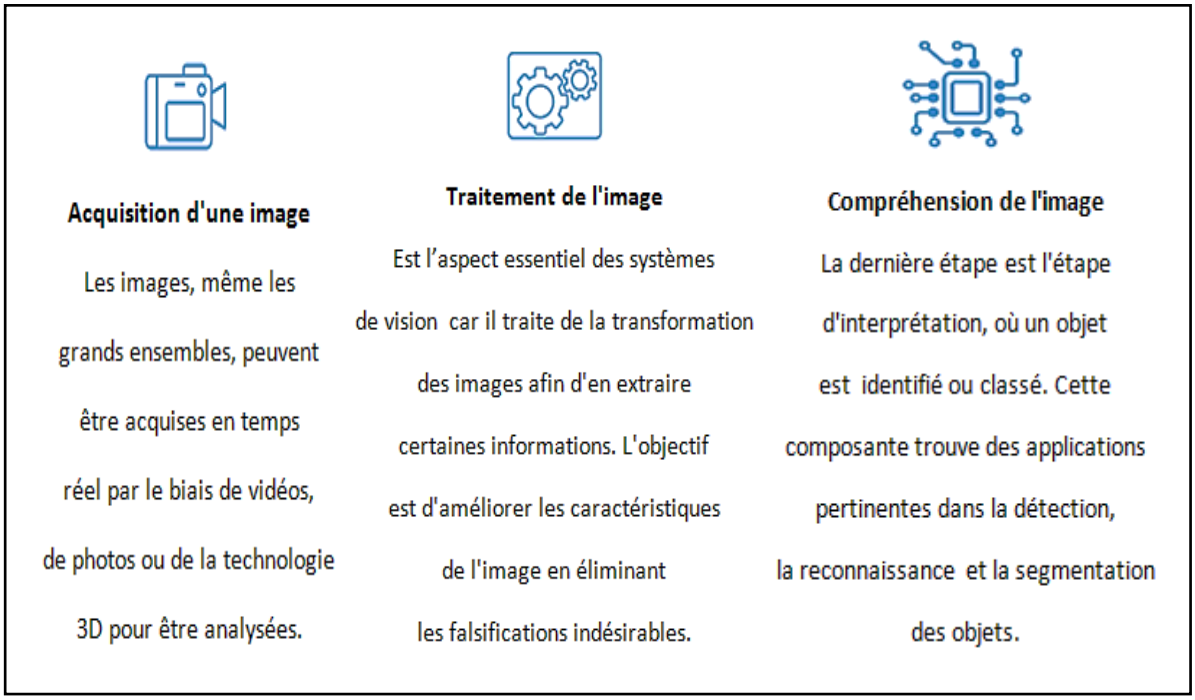

Figure 1.4.2 Les principaux composants de la vision par ordinateur.

1.4.2 Applications

Certaines personnes pensent que la vision par ordinateur est quelque chose qui vient d'un futur lointain de la conception. C'est faux. La vision par ordinateur est déjà intégrée dans de nombreux domaines de notre vie grâce aux progrès rapides de l'IA [21]. Vous trouverez ci-dessous quelques exemples notables de la manière dont nous utilisons cette technologie aujourd'hui :

- La surveillance
- Reconnaissance faciale
- Détection
- Voitures à conduite autonome
- Drones

1.5 Pourquoi l'apprentissage profond ?

Pour comprendre le processus récent de la technologie de vision par ordinateur, nous devons nous plonger dans les algorithmes sur lesquels cette technique s'appuie.

Elle se repose sur l'apprentissage profond, un sous-ensemble spécifique de l'apprentissage automatique, qui sont disponibles et aident à construire des modèles.

L'apprentissage profond est important pour une seule raison, la précision obtenue est significative et utile pour des tâches importantes. L'apprentissage automatique est utilisé pour la classification d'images et de textes depuis des décennies, mais il a du mal à franchir le seuil, il y a une précision de base que les algorithmes doivent avoir pour fonctionner dans des environnements professionnels. L'apprentissage profond a permis enfin de franchir cette limite dans des domaines où personne n'était en mesure de le faire auparavant [23].

L'apprentissage profond est très populaire aujourd'hui parce qu'il permet aux machines d'obtenir des résultats à la hauteur des performances humaines [9].

Par exemple, dans le domaine de la reconnaissance profonde des visages, les modèles d'IA atteignent une précision de détection (par exemple, Google FaceNet a atteint 99,63 %) supérieure à celle que les humains peuvent obtenir (97,53 %) [9].

Les algorithmes de ML fonctionnent bien pour une grande variété de problèmes. Cependant ils ont échoués à résoudre quelques problèmes majeurs de l'IA telle que la reconnaissance vocale et la reconnaissance d'objets [24].

Une des grandes différences entre l'apprentissage profond et les algorithmes de ML traditionnelles est qu'il s'adapte bien, plus la quantité de données fournie est grande plus les performances d'un algorithme de DL sont meilleurs. Contrairement à plusieurs algorithmes de ML classiques qui possèdent une borne supérieure à la quantité de données qu'ils peuvent recevoir appelée "plateau de performance", les modèles de DL n'ont pas de telles limitations (théoriquement) et ils sont même allés jusqu'à dépasser la performance humaine dans des domaines comme le traitement d'image [24].

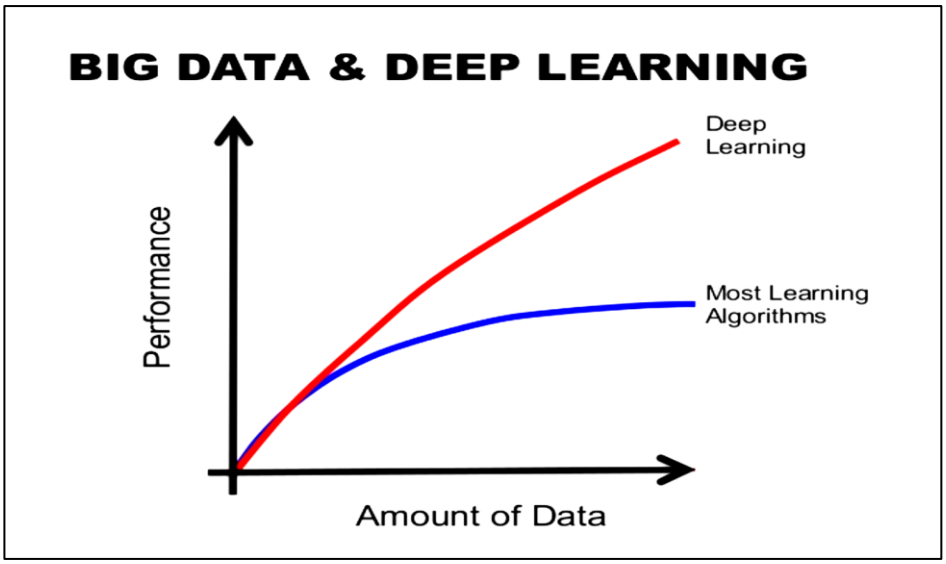

Figure 1.5.1 La différence de performance entre l'apprentissage profond et la plupart des algorithmes de ML en fonction de la quantité de données [24].

Autre différence entre les algorithmes de ML traditionnelles et les algorithmes de DL est l'étape de l'extraction de caractéristiques. Dans les algorithmes de ML traditionnelles l'extraction de caractéristiques est faite manuellement, c'est une étape difficile et coûteuse en temps et requiert un spécialiste en la matière alors qu'en apprentissage profond cette étape est exécutée automatiquement par l'algorithme [24].

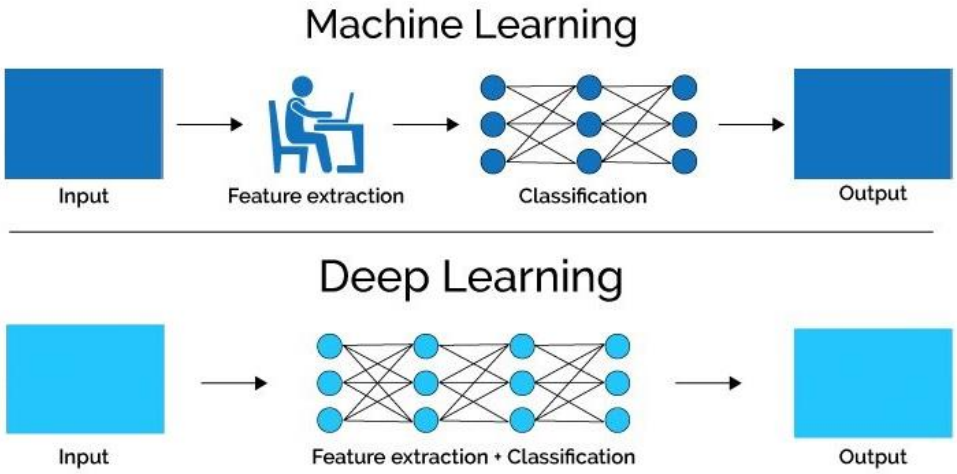

Figure 1.5.2 Comparaison entre l'apprentissage automatique et l'apprentissage profond.

L'apprentissage profond nécessite des machines haut de gamme, contrairement aux algorithmes d'apprentissage automatique traditionnels. Un GPU est capable d'effectuer plusieurs tâches simultanément. L'exécution d'un réseau neuronal, que ce soit lors de l'apprentissage ou de l'application du réseau, peut être très bien réalisée à l'aide d'un GPU [9].

Réseaux de neurones et Détection d’objets

2.1 Les réseaux de neurones convolutifs

Au cours des dernières années, les modèles de réseaux neuronaux profonds ont suscité un intérêt croissant pour la résolution de divers problèmes de vision [25], telles que la détection d'objets, la détection de visage ou la reconnaissance, l'estimation de pose et plus encore.

L'un des cadres d'apprentissage profond les plus réussis est l'architecture des réseaux neuronaux convolutifs (CNN) [25], qui offrent une amélioration spectaculaire des performances par rapport aux algorithmes de traitement d'image traditionnels [26].

En général, le principal avantage du CNN est qu'il a tendance à être un moyen plus puissant et plus précis de résoudre les problèmes de classification [27].

Étant donné une série d'images ou de vidéos du monde réel, avec l'utilisation de CNN, le système d'IA apprend à extraire automatiquement les caractéristiques de ces entrées pour effectuer une tâche spécifique sans aucune supervision humaine [28].

Les réseaux convolutifs sont extrêmement efficaces dans les domaines où des données volumineuses et non structurées sont impliquées, comme la classification d'images [29].

Cependant, en raison de la capacité de CNN à considérer les images comme des données, il s'agit de la solution la plus répandue pour les problèmes de vision par ordinateur et d'apprentissage automatique dépendant de l'image [27].

2.1.1 Réseau de neurones convolutif

Un réseau de neurones à convolution, appelé aussi Convnet pour «Convolutional Network», ou encore CNN pour «Convolutional Neural Network» [30].

Un réseau neuronal convolutif est un puissant réseau neuronal avec une architecture multicouche qui utilise des filtres pour extraire des caractéristiques des images. Il le fait également de manière à conserver les informations relatives à la position des pixels [31]. En d'autres termes, il s'agit de réseaux neuronaux qui sont

particulièrement aptes à construire des caractéristiques complexes à partir de caractéristiques moins complexes [32].

2.1.2 Architecture d'un réseau de neurone convolutif

L'architecture de CNN repose sur plusieurs réseaux de neurones profonds consistant en une succession de couches de convolution dédiés à l'extraction automatique de caractéristiques, tandis que la seconde partie est dédiée à la classification [33], figure 2.2.1.

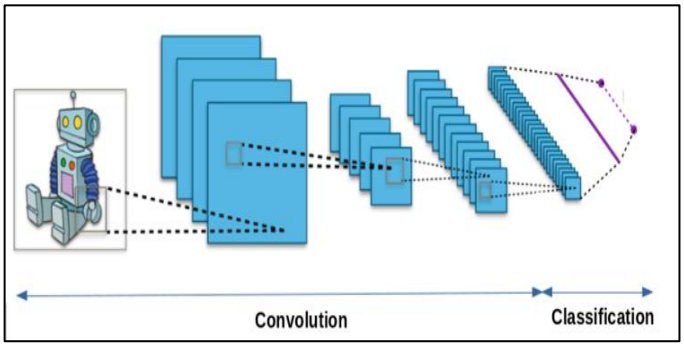

Figure 2.1.1 Architecture d'un réseau de neurones convolutif [30].

Dans les réseaux de neurones convolutifs, chaque couche agit comme un filtre de détection pour la présence de caractéristiques spécifiques ou des motifs présents dans les données d'origine. Les premières couches d'un réseau convolutif détectent des caractéristiques qui peuvent être reconnues et interprétées facilement. Les couches ultérieures détectent de plus en plus des caractéristiques plus abstraites. La dernière couche du réseau convolutif est capable de faire une classification ultra-spécifique en combinant toutes les caractéristiques spécifiques détectées par les couches précédentes [10].

Les CNN se composent généralement de 3 types de couches :

- Couche de convolution
- Couche de mise en commun
- Couche entièrement connectée

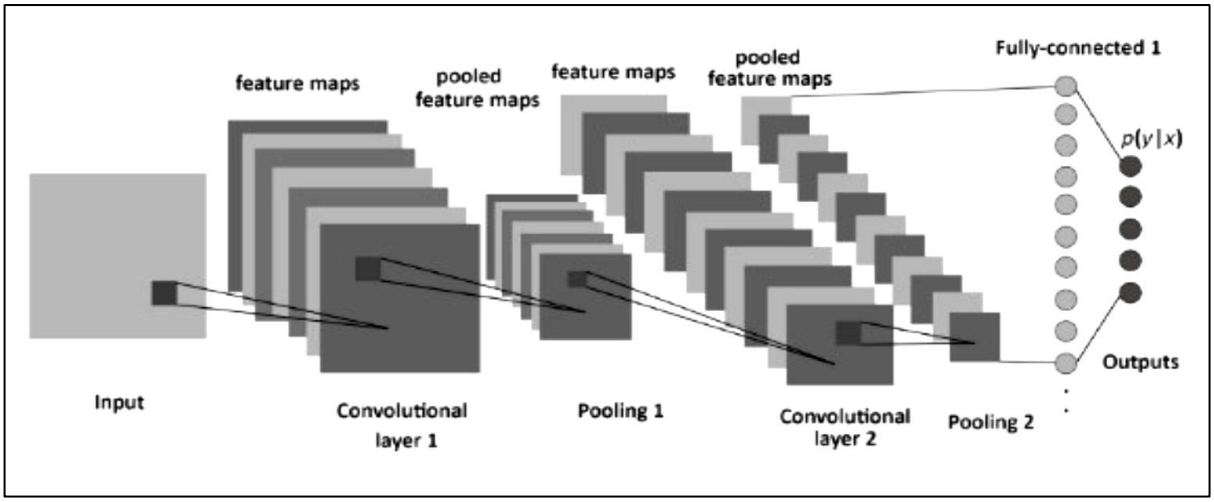

Figure 2.1.2 Couches du réseau CNN [34].

a Couche de convolution

La couche de convolution est parfois appelée couche d'extraction de caractéristiques. Elle est le bloc de construction de base d'un CNN, et c'est là que la majorité des calculs sont effectués [35].

Le but principal de la convolution dans le cas d'un ConvNet est d'extraire des caractéristiques de l'image d'entrée [10].

Tout d'abord, une partie de l'image est connectée à la couche Convolution pour effectuer une opération de convolution et calculer le produit scalaire entre le champ récepteur (c'est une région locale de l'image d'entrée ayant la même taille que celle du filtre) et le filtre comme le montre la figure 2.2.3. Le résultat de l'opération est un entier unique du volume de sortie. Ensuite, nous faisons glisser le filtre sur le champ récepteur de la même image d'entrée par une foulée (stride) et refaisons la même opération. Cette opération est répétée par le même processus encore et encore jusqu'à ce que toute l'image soit parcourue [33].

Supposons que l'entrée soit une image couleur, composée d'une matrice de pixels en 3D. Cela signifie que l'entrée aura trois dimensions - hauteur, largeur et profondeur qui correspondent au RVB d'une image. Le détecteur de caractéristiques, également appelé noyau ou filtre, se déplace dans les champs réceptifs de l'image

pour vérifier la présence de la caractéristique. Ce processus est connu sous le nom de convolution [35].

La taille du filtre est généralement une matrice 3x3 et la sortie finale de la série de produits scalaires de l'entrée et du filtre est appelée carte de caractéristiques, carte d'activation ou caractéristique convoluée [35].

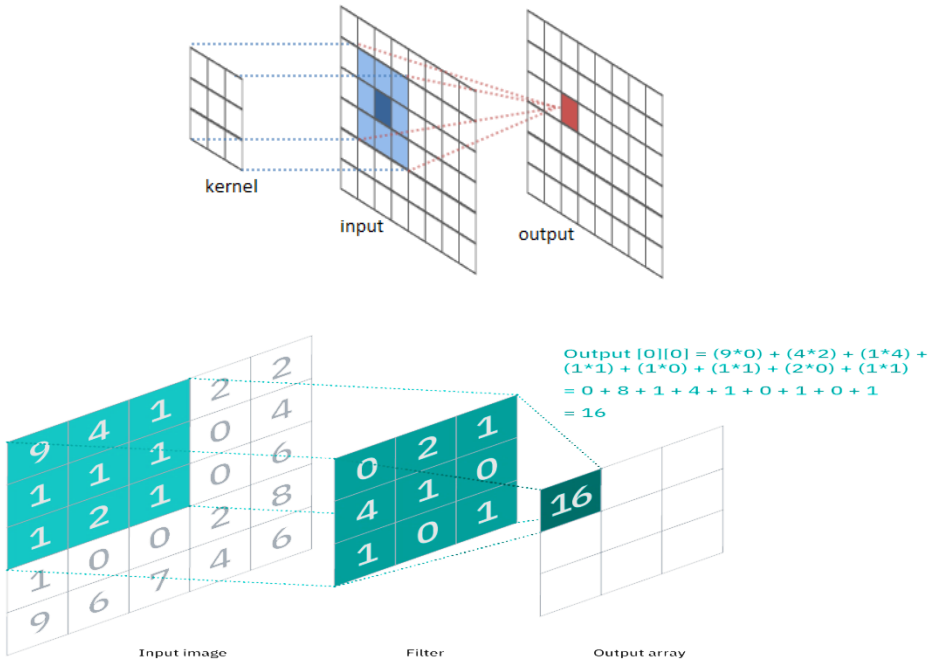

Figure 2.1.3 Entrée convoluée avec un noyau (filtre) [35].

Comme nous pouvons le voir dans l'image de la figure 2.2.3, chaque valeur de sortie de la carte de caractéristiques ne doit pas nécessairement être connectée à chaque valeur de pixel de l'image d'entrée. Elle doit seulement être connectée au champ réceptif, où le filtre est appliqué. Étant donné que la carte de sortie ne doit pas nécessairement correspondre directement à chaque valeur d'entrée, les couches convolutionnelles (et de mise en commun) sont communément appelées couches "partiellement connectées". Cependant, cette caractéristique peut également être décrite comme une connectivité locale [35].

Après chaque opération de convolution, le CNN applique une transformation ReLU (Rectified Linear Unit) à la carte des caractéristiques pour augmenter la non-linéarité en sortie [35].

b Couche de mise en commun (Pooling Layer)

Pour réduire la complexité de calcul et obtenir une représentation d'image hiérarchique, chaque séquence de couches de convolution est suivie d'une couche de mise en commun [36].

Lorsque nous utilisons le remplissage dans la couche de convolution, la taille de l'image reste la même. Les couches de mise en commun sont donc utilisées pour réduire la taille de l'image. Elles fonctionnent par échantillonnage dans chaque couche en utilisant des filtres [31].

Dans les architectures classiques de réseaux de neurones convolutifs, les couches de convolution sont suivies par des couches de sous échantillonnage (couche d'agrégation). Cette dernière réduit la taille des cartes de caractéristique pour but de diminuer la taille de paramètres, et renvoie les valeurs maximales des régions rectangulaires de son entrée, tout en conservant les informations importantes [33].

Elle se présente sous plusieurs formes, mais la plus couramment utilisée est le max pooling, figure 2.2.4.

Bien qu'une grande quantité d'informations soit perdue dans la couche de mise en commun, elle présente également un certain nombre d'avantages pour le CNN. Elles permettent de réduire la complexité, d'améliorer l'efficacité et de limiter le risque de sur-apprentissage [35].

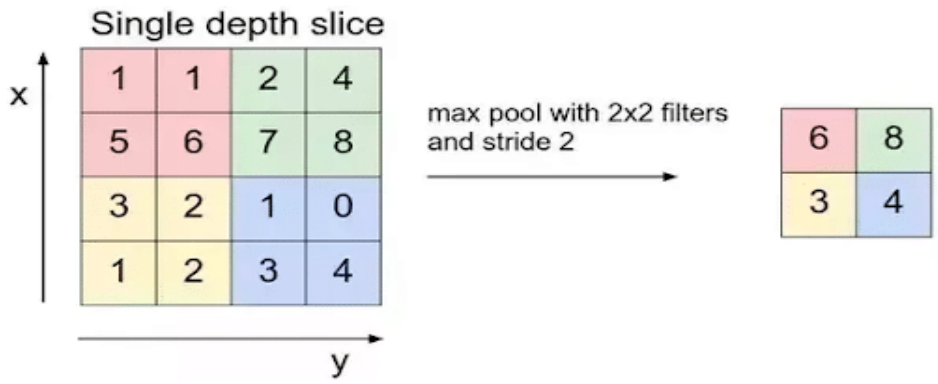

Figure 2.1.4 Processus du Max pooling (prend la plus grande valeur de chaque fenêtre) [37].

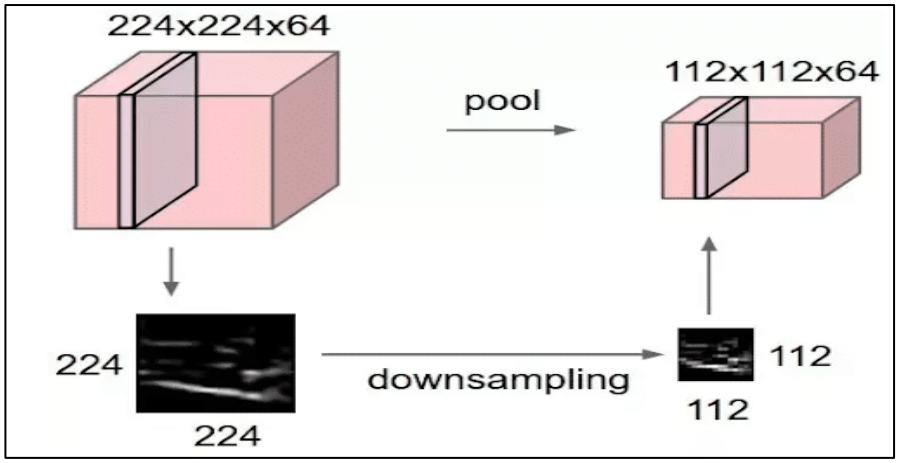

Figure 2.1.5 Exemple de principe du Pooling [37].

❖ **À quoi sert la partie convolutive des modèles CNN ?**

En effet, comme nous l'avons vu ci-dessus, la partie de convolution va avoir pour effet de réduire la dimension de la carte de caractéristiques que l'on obtient après la convolution (en comparaison avec la taille de l'image en entrée). Si l'on répète ce processus plusieurs fois, en prenant comme nouvelle entrée (sur laquelle nous allons effectuer la convolution) la sortie de la convolution précédente, nous allons diminuer de plus en plus la taille de la carte de caractéristiques, et donc nous diminuons également le nombre de calculs [30], figure 2.2.6.

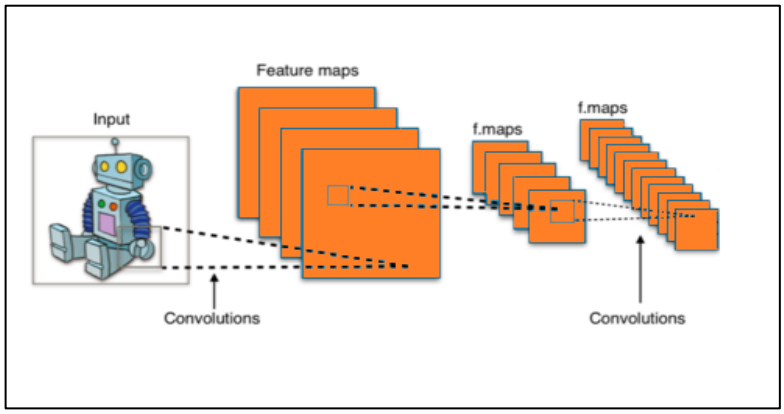

Figure 2.1.6 Réduction de la taille de la carte de caractéristiques [30].

c La couche entièrement connectée (Fully-Connected)

À la sortie de la partie convolutive nous avons un long vecteur qui comprend les caractéristiques les plus pertinentes de l'image. Nous connectons chaque valeur de ce

vecteur à un neurone du réseau de la partie classification, que l'on appelle réseau de neurone entièrement connecté [30].

À la fin des couches de convolution et de mise en commun, les réseaux utilisent généralement des couches entièrement connectées dans lesquelles chaque pixel est considéré comme un neurone distinct, comme dans un réseau neuronal classique. Ces couches contiennent autant de neurones que le nombre de classes à prédire [31].

La couche entièrement connectée (***fc***) est utilisée à la fin pour affecter la caractéristique à la probabilité de classe après l'extraction et la consolidation des caractéristiques de la couche convolutive et la mise en commun ultérieure, respectivement. Ces couches utilisent des fonctions d'activation linéaires ou des fonctions d'activation softmax [38].

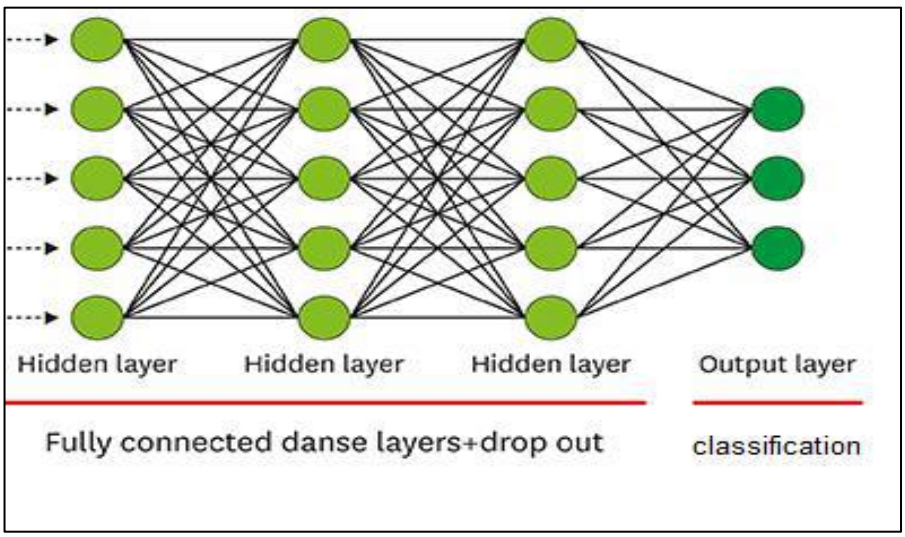

Figure 2.1.7 Couche entièrement connectée [10].

Les performances et l'efficacité d'un CNN sont déterminées par son architecture. Cela inclut la structure des couches, la façon dont les éléments sont conçus et les éléments présents dans chaque couche [26]. De nombreux CNN ont été créés, voici quelques-uns des modèles les plus efficaces :

- AlexNet
- Google Net
- VGGNet

- ResNet
- Xception

2.2 Détection d'objets

La détection d'objets est le concept de base pour la reconnaissance des objets, la reconnaissance faciale, et bien d'autres.

2.2.1 Qu'est-ce que la détection d'objets ?

La détection d'objets est une technologie informatique liée à la vision par ordinateur et au traitement d'images qui traite de la détection d'instances d'objets sémantiques d'une certaine classe (tels que des humains, des bâtiments ou des voitures) dans des images et des vidéos numériques. La détection d'objets a des applications dans de nombreux domaines de la vision par ordinateur, y compris la récupération d'images et la vidéosurveillance [39].

Le problème de la détection d'objets est plus complexe que la classification, qui peut également reconnaître des objets mais n'indique pas où se trouve l'objet dans l'image. De plus, la classification ne fonctionne pas sur les images contenant plus d'un objet. Par contre la détection implique à la fois la classification de chaque objet dans l'image et sa localisation [40].

La localisation d'image est utilisée pour déterminer où se trouvent les objets dans une image. Une fois identifiés, les objets sont marqués d'un cadre de délimitation. La détection d'objets s'étend sur cela et classe les objets qui sont identifiés. Ce processus est basé sur des CNN tels qu'AlexNet, Fast RCNN et Faster RCNN [26].

La localisation et la détection d'objets peuvent être utilisées pour identifier plusieurs objets dans des scènes complexes. Cela peut ensuite être appliqué à des fonctionnalités telles que le comptage de personnes, la détection de visages, la détection de texte, la détection de pose ou la reconnaissance de plaques minéralogiques [26].

L'emplacement est indiqué en dessinant un cadre de délimitation autour de l'objet. Le cadre de délimitation peut ou non localiser avec précision la position de l'objet [41].

Il est naturel que la détection d'objets anticipe trop de cadres de délimitation. De plus, chaque case a un score de confiance qui indique la probabilité que le modèle pense réellement que l'image contient un objet. Au final, toutes les cases dont le score tombe en dessous d'un bord spécifique sont supprimées [42].

La capacité de localiser l'objet à l'intérieur d'une image définit les performances de l'algorithme utilisé pour la détection. La détection de visage est l'un des exemples de détection d'objets [41].

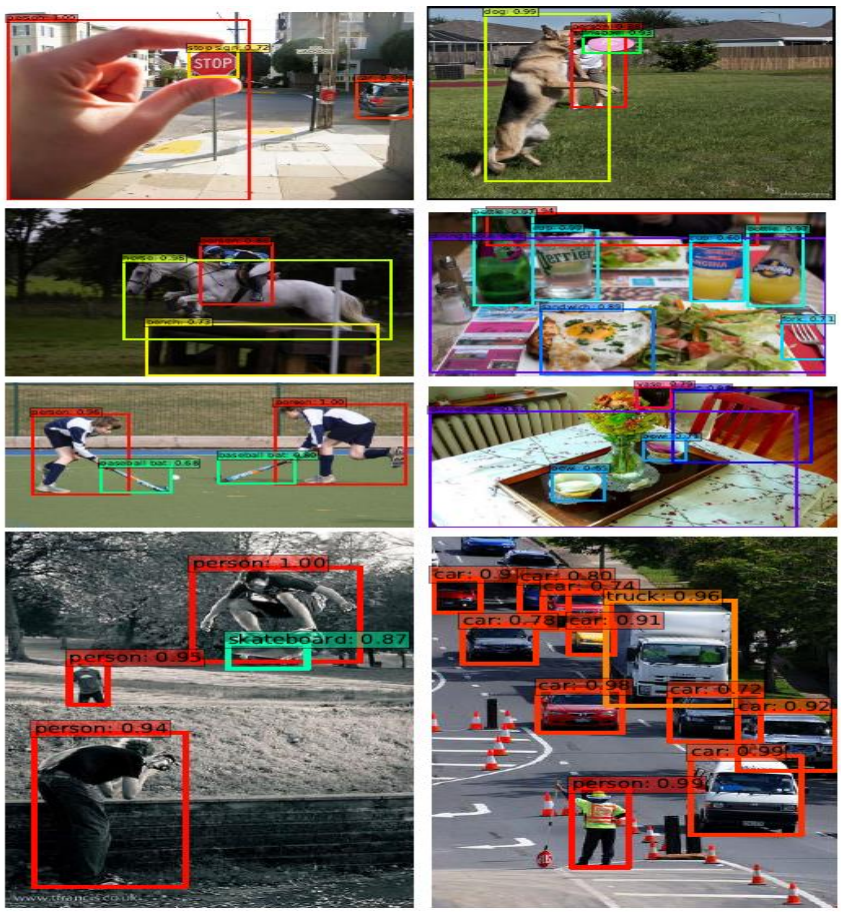

Figure 2.2.1 Détection d'objets [43].

La détection d'objets peut être effectuée à l'aide de techniques traditionnelles de traitement d'images ou de réseaux modernes d'apprentissage profond [44].

1. **Les techniques de traitement d'images** ne nécessitent généralement pas de données historiques pour la formation et sont non supervisées par nature.
   - **Avantages:** Par conséquent, ces tâches ne nécessitent pas d'images annotées, où les humains étiquettent les données manuellement (pour la formation supervisée).
   - **Inconvénients:** Ces techniques sont limitées à de multiples facteurs, tels que les scénarios complexes (sans arrière-plan unicolore), l'occlusion (objets partiellement cachés), l'éclairage et les ombres, et l'effet de fouillis.
2. **Les méthodes d'apprentissage profond** dépendent généralement d'une formation supervisée. Les performances sont limitées par la puissance de calcul des GPU, qui augmente rapidement d'année en année.
   - **Avantages:** La détection d'objets par apprentissage profond est nettement plus robuste aux occlusions, aux scènes complexes et aux éclairages difficiles.
   - **Inconvénients:** Une énorme quantité de données d'entraînement est nécessaire; le processus d'annotation des images est laborieux et coûteux.

2.3.2 Comparaison des détecteurs d'objets à un étage et à deux étages

Actuellement, les cadres de détection d'objets basés sur l'apprentissage approfondi peuvent être principalement divisés en deux types (figure 2.3.2) : notamment les méthodes basées sur la proposition de régions (Détecteurs à deux étages) et les méthodes basées sur la régression (Détecteurs à un étage) [10].

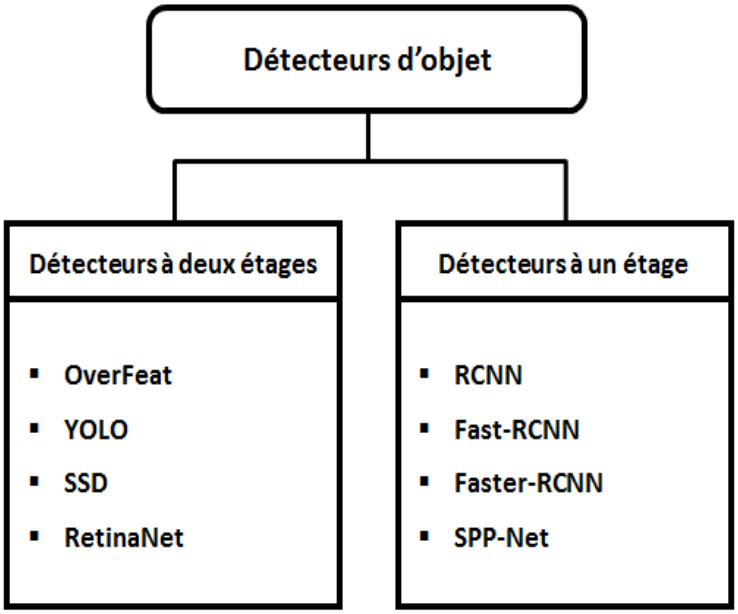

Figure 2.2.2 Principales catégories de détecteurs d'objets.

a Détecteurs à deux étages

Les détecteurs à deux étages ont donné une précision plus élevée avec de meilleures performances et rapportent des résultats idéaux mieux que les détecteurs à un étage dans la détection d'objets [10], mais ils sont généralement plus lents que les détecteurs à un étage car ils ont deux étapes : d'abord l'extraction de propositions de régions, puis la classification de chaque proposition et la prédiction de la boîte englobante [45].

Même en limitant le nombre de régions à traiter, les améliorations de performance ne sont pas suffisantes pour faciliter le fonctionnement en temps réel [45].

b Détecteurs à un étage

Les détecteurs à un étage sont beaucoup plus rapides et plus recherchés pour les applications de détection d'objets en temps réel, mais ont des performances relativement médiocres par rapport aux détecteurs à deux étages [46]. Ils peuvent réduire les calculs en supprimant l'étape de proposition de région et en formulant la détection d'objets comme un problème de régression dense [45].

Au lieu d'avoir deux réseaux pour deux tâches différentes (proposition de région et classification de région), un seul ConvNet est utilisé pour les deux tâches simultanément. Un seul réseau neuronal prédit les boîtes englobantes et les probabilités de classe directement à partir d'images complètes en une seule évaluation [45].

2.3 Détection de visages

La détection de visages est une version spécialisée de la détection d'objets. C'est la première étape nécessaire pour tous les algorithmes d'analyse faciale, y compris l'alignement des visages, la reconnaissance des visages, la vérification des visages et l'analyse des visages. Elle consiste à détecter la présence et la localisation précise d'un ou plusieurs visages humains dans une image numérique ou à partir d'une série d'images provenant d'un dispositif de capture vidéo [47].

Si plusieurs visages sont présents, chaque visage est entouré d'une boîte englobante et nous connaissons donc l'emplacement des visages.

La détection du visage nécessite un haut niveau de codage à l'aide d'un algorithme capable de détecter des images en mouvement à partir d'un flux vidéo en cours d'exécution et de capturer les différentes poses qui constituent son ensemble de données [48].

2.4.1 Les principaux défis

Les défis de la détection des visages sont les raisons qui réduisent la précision et le taux de détection de la reconnaissance faciale. Ces défis sont les suivants : arrière-plan complexe, trop de visages dans les images, expressions bizarres, éclairages, résolution moindre, occlusion du visage, couleur de la peau, distance et orientation, etc. [49].

- **Expression inhabituelle :** Les visages humains dans une image peuvent présenter des expressions faciales inattendues ou étranges.
- **Éclairage :** Certaines parties de l'image peuvent présenter une illumination ou des ombres très élevées ou faibles.

- **Types de peau :** La détection de visages de couleurs différentes est un défi pour la détection et nécessite une plus grande diversité d'images d'entraînement.
- **La distance :** Si la distance à la caméra est trop élevée, la taille de l'objet (taille du visage) peut être trop petite.
- **L'orientation :** L'orientation du visage et l'angle par rapport à la caméra ont un impact sur le taux de détection des visages.
- **Arrière-plan complexe :** Un nombre élevé d'objets dans une scène réduit la précision et le taux de détection.
- **Nombreux visages dans une image :** Une image contenant un grand nombre de visages humains est très difficile à détecter avec précision.
- **Occlusion des visages :** Les visages peuvent être partiellement cachés par des objets tels que des lunettes, des foulards, des mains, des cheveux, des chapeaux et d'autres objets, ce qui a un impact sur le taux de détection.
- **Faible résolution :** Les images à faible résolution ou le bruit de l'image ont un impact négatif sur le taux de détection.

Un autre important défi est la nécessité de détecter des visages en temps réel. Cela nécessite souvent de grandes performances, ou on peut sacrifier la précision pour gagner en vitesse [10].

Pour cela un détecteur idéal devrait avoir une:

- Haute précision de localisation et de reconnaissance: le détecteur doit être capable de localiser et de reconnaître les objets dans les images avec précision.
- Haute efficacité en temps et en mémoire: la tâche de détection doit s'exécuter à une fréquence d'images suffisante (fps) avec une utilisation de mémoire et de stockage acceptable [10].

2.4.2 Avantages de la détection des visages

En tant qu'élément clé des applications d'imagerie faciale, telles que la reconnaissance et l'analyse des visages, la détection des visages présente divers avantages pour les utilisateurs [50], notamment :

- **Amélioration de la sécurité** : La détection des visages améliore les efforts de surveillance et permet de traquer les criminels et les terroristes. La sécurité personnelle est également renforcée puisque les pirates n'ont rien à voler ou à modifier, comme les mots de passe.
- **Facile à intégrer :** La technologie de détection et de reconnaissance des visages est facile à intégrer, et la plupart des solutions sont compatibles avec la majorité des logiciels de sécurité.
- **Identification automatisée :** Dans le passé, l'identification était effectuée manuellement par une personne, ce qui était inefficace et souvent inexact. La détection des visages permet d'automatiser le processus d'identification, ce qui permet de gagner du temps et d'accroître la précision.

2.4.3 Applications de la détection des visages

- **Surveillance des foules :** La détection des visages est utilisée pour détecter les foules dans les zones publiques ou privées fréquentées.
- **Interaction homme-machine :** De nombreux systèmes basés sur l'interaction homme-machine utilisent la reconnaissance faciale pour détecter la présence d'humains.
- **Photographie :** Certains appareils photo numériques récents utilisent la détection des visages pour l'autofocus. Les applications mobiles utilisent la reconnaissance faciale pour détecter les régions d'intérêt dans les diaporamas.
- **Extraction de caractéristiques facials :** Les caractéristiques faciales telles que le nez, les yeux, la bouche, la couleur de la peau, etc. peuvent être extraites des images.
- **Classification par sexe.** Les applications sont conçues pour détecter les informations relatives au sexe à l'aide de méthodes de détection des visages.
- **Reconnaissance des visages :** Un système de reconnaissance des visages est conçu pour identifier et vérifier une personne à partir d'une image numérique ou d'une trame vidéo.
- **Marketing :** La détection des visages devient de plus en plus importante pour le

marketing, l'analyse du comportement des clients ou la publicité ciblée.

- **Présence :** La reconnaissance faciale est utilisée pour détecter la présence des personnes. Elle est souvent combinée à la détection biométrique pour la gestion des accès [49].

2.4.4 Choix d'algorithmes (Méta-modèle)

a Faster R-CNN

Parmi les meilleurs détecteurs à un étage le Faster R-CNN, qui est un modèle de réseau sophistiqué et rigoureux qui atteint la plus haute précision de détection d'objets. Néanmoins, le processus de génération des propositions de régions rend la vitesse du réseau à deux étapes bien inférieure aux exigences de la détection en temps réel [51].

Afin d'augmenter la vitesse des détecteurs d'objets basés sur l'apprentissage profond, Single Shot Detector (SSD) et You Only Look Once (YOLO) utilisent une stratégie de détection en une étape [51].

b YOLO

YOLO est un algorithme de détection d'objets en temps réel qui identifie des objets spécifiques dans des vidéos, des flux en direct ou des images.

En tant que détecteur à une seule étape, YOLO effectue la classification et la régression de la boîte englobante en une seule étape, ce qui le rend beaucoup plus rapide que la plupart des réseaux neuronaux convolutifs, comme Faster R-CNN qui nécessite des milliers d'évaluations pour une seule image. Cette rapidité a permis à YOLO d'être exécutable en temps réel [44].

L'aspect limitant et désavantageux de l'algorithme YOLO est

- Il ne traite pas toujours bien les petits objets.
- Il ne gère pas, en particulier, les objets groupés de manière rapprochée.
- Localisations incorrectes.

c YOLO_Face

L'application de la détection des visages est généralement liée aux scènes en temps réel telles que la surveillance et la vidéo, au lieu d'être limitée à la détection d'images. Cela signifie que l'architecture de détection de visage avec de bonnes performances doit répondre à la fois aux exigences de précision et de vitesse [52].

Le détecteur de visages YOLO_Face qui est basé sur YOLO v3 a été proposé afin de préserver la vitesse en temps réel sans sacrifier trop de précision de détection. Il est considéré comme l'un des plus rapides algorithmes avec une précision compétitive par rapport aux autres détecteurs Faster RCNN, SSD, RetinaNet, etc.

Cette étude se concentre sur la manière de détecter rapidement et efficacement les visages dans des scènes denses en temps réel. Par conséquent, le YOLO_Face est choisi comme réseau de base.

2.4 Reconnaissance de visages

La technologie de la reconnaissance faciale a beaucoup évolué au cours des vingt dernières années. Aujourd'hui, les machines sont capables de vérifier automatiquement des informations d'identité pour des transactions sécurisées, pour des tâches de surveillance et de sécurité, et pour le contrôle d'accès [53].

La reconnaissance des visages est la tâche qui consiste à identifier un objet déjà détecté comme étant un visage connu ou inconnu, en utilisant un jeu de données de visages afin de valider ce visage d'entrée [54].

Souvent le problème de la reconnaissance des visages est confondu avec le problème de la détection des visages, en fait, la détection des visages n'est qu'une partie de la reconnaissance des visages [54].

La procédure de reconnaissance des visages humains se compose essentiellement de deux phases [47] :

- **La détection de visage** qui consiste à chercher et à détecter un ou plusieurs visages en parcourant une image numérique ou une vidéo.

- **La reconnaissance de visage** qui consiste à comparer le visage détecté à celui se trouvant dans le jeu de données de visages reconnus afin de savoir à qui appartient-il.

De façon plus indirecte, la détection de visage est la première étape vers des applications plus évoluées, qui nécessitent la localisation du visage [47].

Implémentation et Résultats

3.1 Introduction

La première partie de ce chapitre est une présentation de l'architecture du modèle YOLO v3 ainsi que YOLO_Face. La seconde partie est l'implémentation du modèle choisi YOLO_Face dont on a présenté le matériel et le logiciel nécessaire au fonctionnement de ce dernier. Ensuite on a passé à l'entrainement avec les résultats obtenus, ainsi on a effectué une étude comparative avec le YOLO5Face afin de tester la fiabilité de notre modèle.

3.2 YOLO v3

YOLO v3 est un modèle de détection d'objets en temps réel en une seule étape qui s'appuie sur YOLO v2 avec plusieurs améliorations. Les améliorations comprennent l'utilisation d'un nouveau réseau fédérateur, Darknet-53 [55].

3.2.1 Architecture YOLO v3

YOLO v3 utilise une variante de Darknet, qui a à l'origine un réseau de 53 couches formé sur Imagenet. Pour la tâche de détection, 53 couches supplémentaires sont empilées sur ce réseau, ce qui nous donne une architecture sous-jacente entièrement convolutive de 106 couches pour YOLO v3 [56].

Dans YOLO v3, la détection est effectuée en appliquant des noyaux de détection 1 x 1 sur des cartes de caractéristiques de trois tailles différentes à trois endroits différents du réseau [56].

La forme du noyau de détection est S x S x (B x (5 + C)). Ici, **B** est le nombre de boîtes de délimitation qu'une cellule de la carte de caractéristiques peut prédire, '5' est pour les 4 attributs de boîte de délimitation et une confiance d'objet et **C** est le nombre de classes [56].

YOLO v3 utilise l'entropie croisée binaire pour calculer la perte de classification pour chaque étiquette, tandis que la confiance de l'objet et les prédictions de classe sont prédites par régression logistique [56].

a Architecture CNN du Darknet-53

Dans YOLO v3, une architecture plus profonde d'extracteur de caractéristiques appelée Darknet-53 est utilisée. Il est principalement composé de filtres 3 x 3 et 1 x 1 avec des connexions de saut.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 3.2.1 Architecture CNN du Darknet-53 [56].

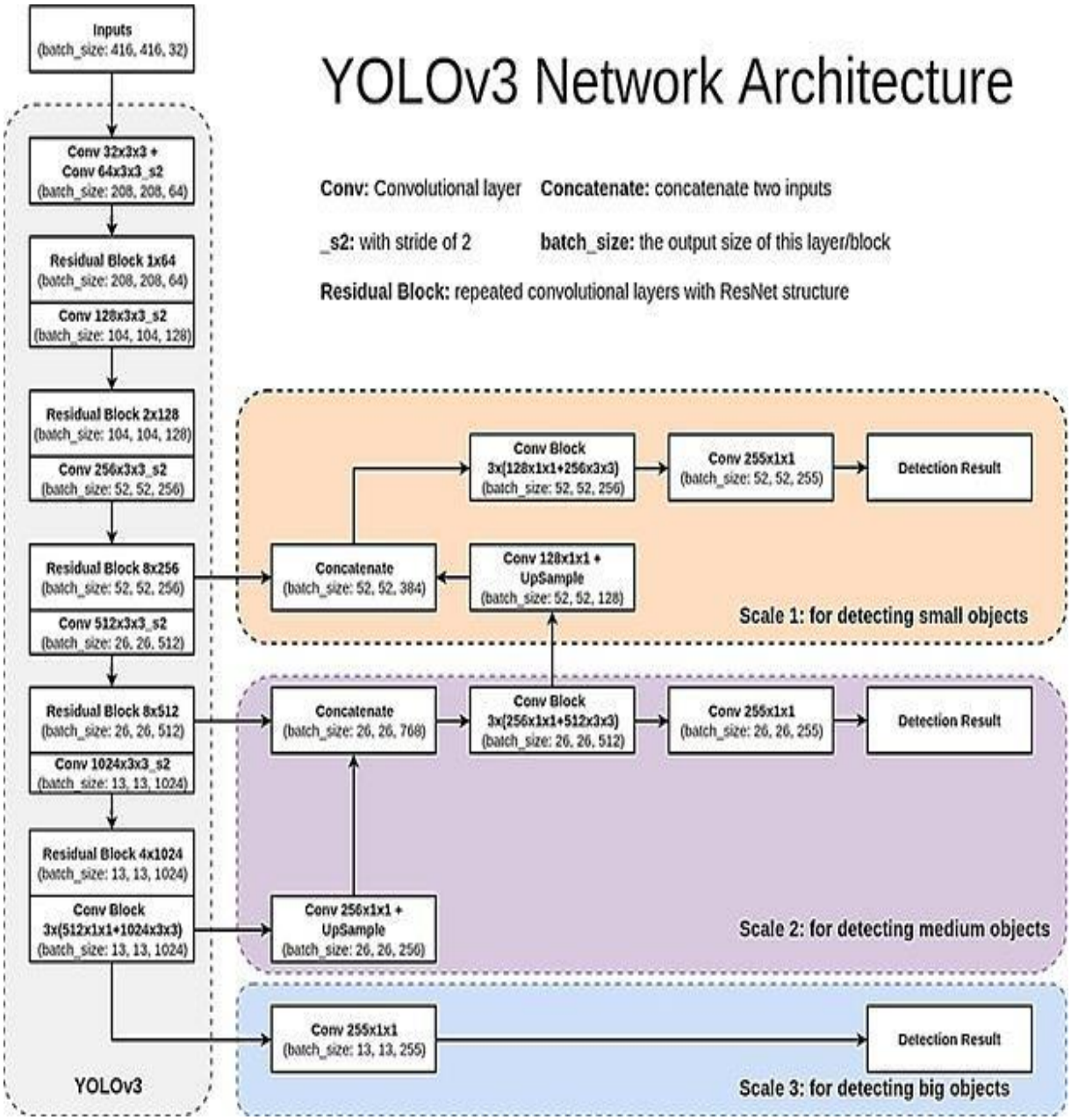

Figure 3.2.2 Diagramme d'architecture de YOLO v3 [57].

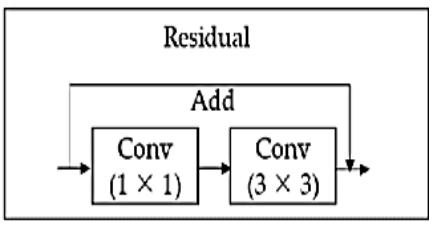

Figure 3.2.3 Le bloc Residual.

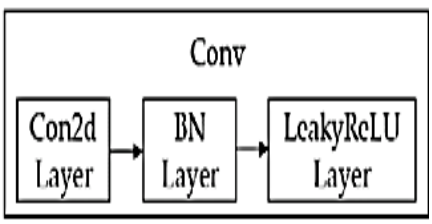

Figure 3.2.4 Le bloc Conv.

❖ **Couches de convolution dans YOLO v3**

L'algorithme contient 53 couches de convolution qui ont été, chacune suivie d'une couche de normalisation par lot (BN) et d'une activation Leaky ReLU.

La couche de convolution est utilisée pour convoluer des filtres multiples sur les images et produire des cartes de caractéristiques multiples.

Aucune forme de regroupement n'est utilisée et une couche convolutionnelle avec un pas de 2 est utilisée pour sous-échantillonner les cartes de caractéristiques.

Cela permet d'éviter la perte de caractéristiques de bas niveau souvent attribuée à la mise en commun [56].

3.3 YOLO_Face

Le problème posé ici était de détecter des visages dans une image donnée. À cette fin, l'algorithme YOLO v3 a été utilisé.

Pour YOLO v3, le nombre de filtres dans l'architecture Darknet est de 255. Ce chiffre provient de 1 x 1 x (B x (5 + C)) = 3× (5+80), mais comme on n'a qu'une seule classe, le nombre de filtres doit être modifié en 3× (5+1), soit 18 !

3.4 Implémentation

3.4.1 Matériel utilisé

Notre projet a été développé sur un PC portable avec les caractéristiques suivantes :

- Processeur : Intel(R) Core (TM) i5-4210U CPU, 1.70 GHz
- Capacité Mémoire (RAM) : 4.00 Go
- Capacité disque dur : 500 Go
- Système d'exploitation : Windows 10 Professionnel
- WebCam

3.4.2 Outils de développement

a Python 

Si l'on s'attarde sur la philosophie qui a présidé à la création du langage Python, on peut dire que ce langage a été conçu pour sa lisibilité et sa nature moins complexe.

Une mise en œuvre simple dans le langage Python aide l'ingénieur à valider une idée.

Cela dépend donc entièrement du type de tâche pour laquelle on souhaite utiliser l'apprentissage profond, par exemples pour un projet de vision par ordinateur, les données d'entrée sont l'image ou la vidéo [58].

Imaginons que tout ce qui existe autour est sous forme de données et ces données sont brutes, inadéquates, incomplètes, non structurées et volumineuses. Python peut servir de guide à l'apprentissage profond pour résoudre tous ces problèmes [58].

Ce langage de programmation présente de nombreuses caractéristiques intéressantes telles que [58]:

- Il est multiplateforme, c'est-à-dire qu'il fonctionne sur de nombreux systèmes d'exploitation: Windows, Mac OS X, Linux, Android, iOS, depuis les mini-ordinateurs Raspberry Pi jusqu'aux supercalculateurs.
- Il est gratuit, vous pouvez l'installer sur autant d'ordinateurs que vous voulez (même sur votre téléphone!).
- C'est un langage de haut niveau, il demande relativement peu de connaissance sur le fonctionnement d'un ordinateur pour être utilisé.
- C'est un langage interprété. Un script Python n'a pas besoin d'être compilé pour être exécuté, contrairement à des langages comme le C ou le C++. Il convient bien à des scripts d'une dizaine de lignes qu'à des projets complexes de plusieurs dizaines de milliers de lignes et il est relativement simple à prendre en main.

b L'environnement de développement « Visual studio code »

Visual Studio Code combine la simplicité d'un éditeur de code source ultra-rapide avec de puissants outils de développement, tels que la complétion de code et le débogage IntelliSense [59].

Il s'agit d'un éditeur parfait pour une utilisation quotidienne qui vous permet de passer moins de temps à manipuler votre environnement et plus de temps à mettre vos idées en pratique [59].

c Bibliothèques utilisées

Tensorflow

TensorFlow est la meilleure bibliothèque de toutes, car elle est conçue pour être accessible à tous. La bibliothèque TensorFlow intègre différentes API pour construire à l'échelle des architectures d'apprentissage profond telles que CNN ou RNN.

TensorFlow est basé sur le calcul de graphe, il permet au développeur de visualiser la construction du réseau neuronal avec Tensorboad. Cet outil est utile pour déboguer le programme. Enfin, Tensorflow est conçu pour être déployé à grande échelle. Il fonctionne sur le CPU et le GPU.

OpenCV

OpenCV est une énorme bibliothèque open-source pour la vision par ordinateur, l'apprentissage automatique et le traitement d'images. OpenCV supporte une grande variété de langages de programmation comme Python, C++, Java, etc. Il peut traiter des images et des vidéos pour identifier des objets, des visages ou même l'écriture d'un être humain [61].

- **OpenCV-Python**

OpenCV-Python est une bibliothèque de liaisons Python conçue pour résoudre les problèmes de vision par ordinateur [62].

Numpy

NumPy, qui signifie Numerical Python, il s'agit d'un module d'extension pour Python, majoritairement écrit en C. Ainsi, les fonctions et fonctionnalités mathématiques et numériques précompilées de Numpy garantissent une grande vitesse d'exécution [63].

En outre, NumPy enrichit le langage de programmation Python avec de puissantes structures de données, mettant en œuvre des tableaux multidimensionnels et des matrices. L'implémentation vise même les matrices et tableaux de grande taille, mieux connus sous le nom de "big data". Le module fournit une grande bibliothèque de fonctions mathématiques de haut niveau pour opérer sur ces matrices et tableaux [63].

Keras

Keras est une bibliothèque Python open source puissante et facile à utiliser pour développer et évaluer des modèles d'apprentissage profond [64].

Elle est capable de fonctionner au-dessus de TensorFlow ou Theano. Elle a été développée dans le but de permettre une expérimentation rapide. Passer de l'idée au résultat en un minimum de temps est essentiel pour mener de bonnes recherches [65].

Matplotlib

Matplotlib.pyplot est une bibliothèque de traçage utilisée pour les graphiques 2D dans le langage de programmation python. Elle peut être utilisée dans les scripts python, les shells, les serveurs d'applications web et autres boîtes à outils d'interface utilisateur graphique [66].

L'un des plus grands avantages de la visualisation est qu'elle nous permet d'avoir un accès visuel à d'énormes quantités de données sous forme de visuels faciles à interpréter. Matplotlib se compose de plusieurs graphiques comme la ligne, la barre, la dispersion, l'histogramme, etc. [67].

Pillow

Pillow est construit sur la base de PIL (Python Image Library). PIL est l'un des modules importants pour le traitement d'images en Python [68].

C'est une bibliothèque relativement complète qui offre la possibilité de manipuler des images avec une grande simplicité. Elle dispose d'un large éventail de fonctions qui touchent à différents domaines allant du filtrage aux graphiques, en passant par la manipulation des pixels [69].

La bibliothèque se positionne donc comme une bibliothèque généraliste dans le domaine du traitement d'images, et ne se démarque pas dans un domaine spécifique.

3.5 Entrainement et résultats

- La 1ère étape consiste a entraîné notre modèle sur un jeu de données dont le choix a été porté sur la base WiderFace [70].

3.5.1 Jeu de données WiderFace

Dans les photos, les vidéos et les scènes en temps réel, l'échelle du visage est généralement assez petite, tandis que l'environnement de fond a tendance à être très compliqué. Par conséquent, la formation d'un détecteur de visage robuste nécessite un ensemble de données de WiderFace avec suffisamment d'informations sur les visages minuscules ainsi que du bruit pour simuler les caractéristiques des données du scénario d'application [52].

Ce jeu de données de référence pour la détection de visages comprend 32 203 images et 393 703 visages étiquetés avec un haut degré de variabilité en termes d'échelle, de pose et d'occlusion [49].

Le jeu de données WiderFace est organisé en 61 classes et, pour chaque classe, 40 %, 10 % et 50 % des données sont sélectionnées de manière aléatoire pour les ensembles d'entrainement, de validation et de test, respectivement. En outre, les images sont divisées en trois niveaux (Easy, Medium, Hard) en fonction des difficultés de détection. WiderFace dépasse de loin les jeux de données existants en termes de

volume de données, de volume de tags, de diversité des visages, ce qui représente un grand défi pour la détection des visages [71].

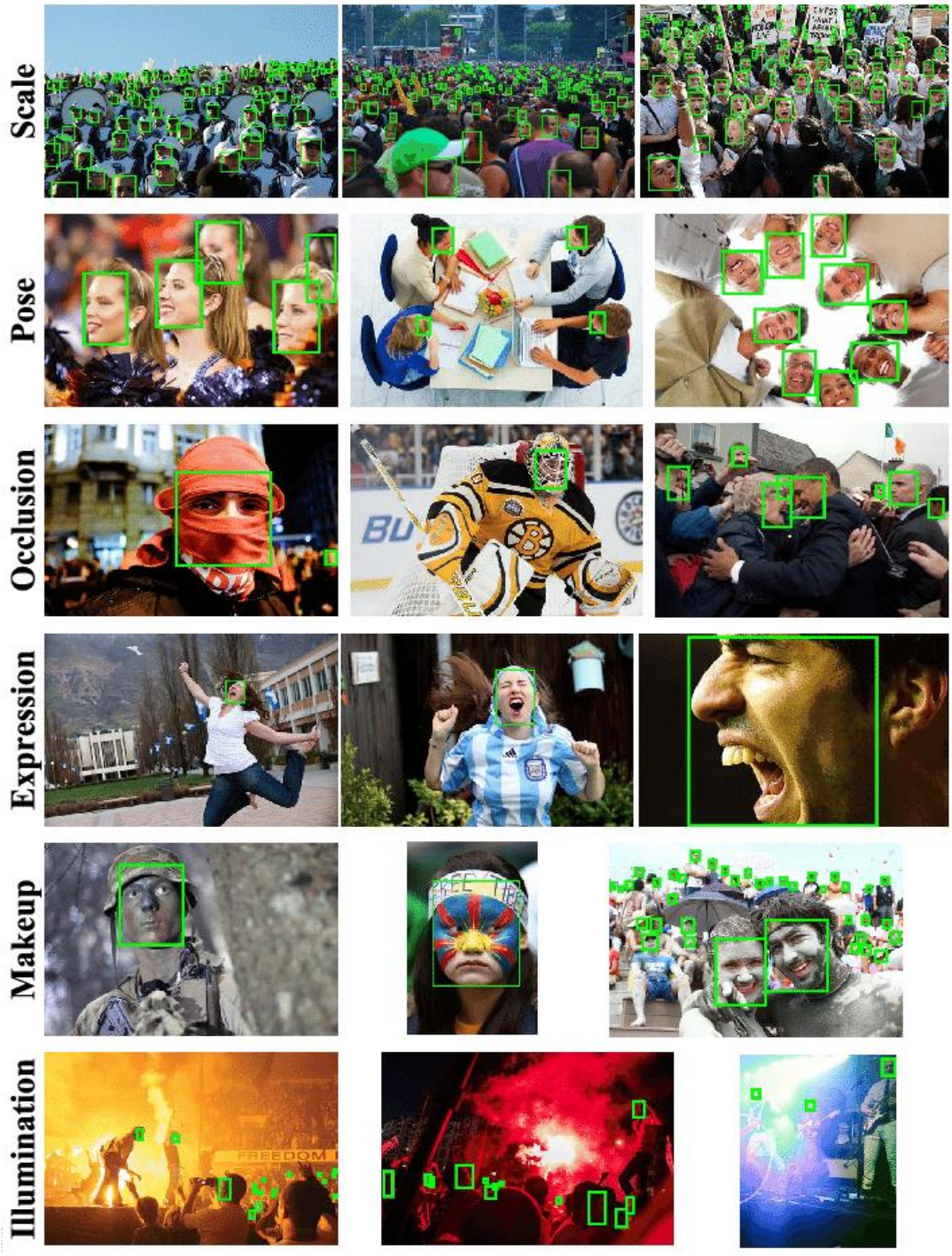

Figure 3.5.1 Quelques images du jeu de données WiderFace avec un haut degré de variabilité [70].

- Avant de passer à l'étape suivante les images doivent être redimensionnées à 416 x 416 pixels avant d'être introduites dans notre modèle ou les dimensions peuvent également être spécifiées lors de l'exécution du fichier python.
- Une fois que le modèle a terminé l'entrainement, l'image d'entrée est passée au modèle avec les tailles des boîtes d'ancrage à utiliser dans les trois couches de sortie YOLO.

Le modèle renvoie les boîtes d'ancrage prédites pour l'image, les boîtes redondantes sont supprimées en utilisant la Suppression Non Maximale (NMS), puis les boîtes prédites finales sont ajustées en fonction de la taille de l'image. Les boîtes finales obtenues, qui sont les visages prédits, sont tracées sur l'image d'entrée et affichées.

3.5.2 Paramètres utilisés

- **Seuil de confiance:** Confiance minimale pour qu'une boîte soit détectée

CONF_THRESHOLD = 0.5

- **Seuil de suppression non-maximum:** Il permet de surmonter le problème de la détection d'un objet plusieurs fois dans une image. Pour ce faire, il prend les cases avec une probabilité maximale et supprime les cases proches avec des probabilités non maximales (inférieures au seuil prédéfini).

NMS_THRESHOLD = 0.4

- **La taille d'image:** 416×416

IMG_WIDTH = 416

IMG_HEIGHT = 416

3.5.3 Résultats sur des images fixes

Les figures ci-dessous nous montrent les différents résultats sur des images fixes obtenus avec YOLO_Face.

.

Figure 3.5.2 Résultats du modèle YOLO_Face.

3.5.4 Détection de visages vidéo

La détection de visages vidéo permet de détecter des visages à partir d'une vidéo. Une image est une image unique qui capture une instance statique unique d'un événement naturel. D'un autre côté, une vidéo contient de nombreuses instances d'images statiques affichées en une seconde, induisant l'effet de visionner un événement naturel. OpenCV fournit une interface très simple pour capturer une vidéo «VideoCapture». Cette fonction peut ouvrir un fichier vidéo ou une séquence de fichiers image ou un périphérique de capture ou un flux vidéo IP pour la capture vidéo.

3.5.5 Détection de visages de webcam en temps réel

L'objectif de la détection de visages de la webcam en temps réel est la détection et le suivi simultanés des instances dans les vidéos en direct.

3.6 Résultats

Les résultats obtenus dans les trois cas précédents montrent la robustesse et la fiabilité de notre modèle YOLO_Face à détecter les visages dans des scènes denses en temps réel malgré que de nombreuses interférences rendent ces scènes plus compliquées.

3.7 Etude comparative

Cette partie représente une comparaison entre le YOLO_Face et le YOLO5Face afin de tester la fiabilité de notre modèle.

Tout d'abord, on décrit le YOLO v5 et le YOLO5Face

YOLO v5 est une version récente de la famille de modèles YOLO. Il a été publié en mai 2020. Il est l'un des modèles de détection d'objets les plus rapides et les plus précis. La plus grande contribution de YOLO v5 est qu'il a été écrit dans le cadre PyTorch et il est beaucoup plus léger et facile à utiliser.

YOLO v5 met en œuvre le goulot d'étranglement CSP pour formuler des caractéristiques d'image. Le CSP résout les problèmes de gradient dupliqué dans d'autres dorsales ConvNet plus importantes, ce qui permet d'obtenir moins de paramètres et moins de FLOPS pour une importance comparable. Ceci est extrêmement important pour la famille YOLO, où la vitesse d'inférence et la petite taille du modèle sont de la plus haute importance [72].

YOLO5Face est un détecteur de visages basé sur le détecteur d'objets YOLOv5. Alors que de nombreux détecteurs de visages utilisent des modèles conçus pour la détection de visages, les réalisateurs du YOLO5Face ont traité la détection de visages comme une tâche générale de détection d'objets [73].

Ils ont ajouté une tête de régression à cinq points de repère et utilisé la fonction de perte Wing. Ils ont concédé des détecteurs avec différentes tailles de modèle, d'un grand modèle pour obtenir les meilleures performances, à un modèle super petit pour la détection en temps réel sur un dispositif embarqué ou mobile. Les résultats de l'expérience sur le jeu de données WiderFace montrent que leurs détecteurs de visages peuvent atteindre des performances de pointe, dépassant les détecteurs de visages désignés plus complexes [73].

Figure 3.7.1 Résultat du modèle YOLO5FACE [73].

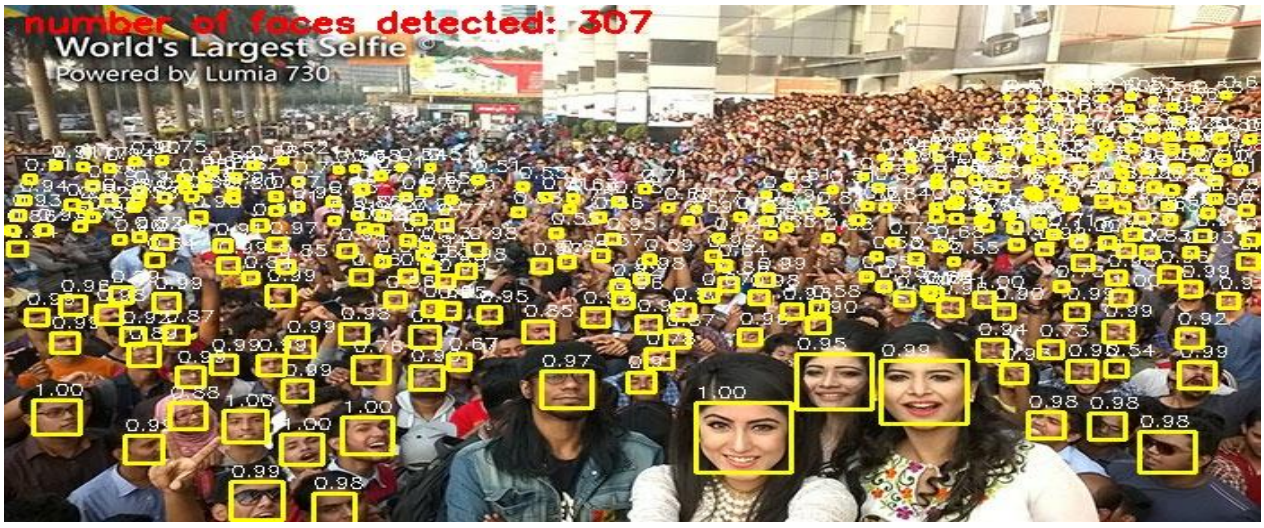

Figure 3.7.2 Résultat du modèle YOLO_Face.

On remarque que les résultats obtenus sont proches pour notre modèle et le modèle YOLO5Face pour les visages du premier plan. Quant à la détection des visages très éloignés, YOLO5Face les détectent avec une considérable précision.

3.8 Conclusion

D'après les résultats obtenus et l'étude comparative, on peut conclure que le modèle YOLO_Face a l'intérêt d'être robuste permettant la détection de visages dans des scènes complexes. Comparé à d'autres modèles des améliorations restent toujours à apporter sur ce modèle.

Conclusion générale

Ce présent travail a consisté à choisir un modèle permettant de détecter l'ensemble des visages en temps réel dans des scènes denses, où de nombreuses interférences et bruits (tels que l'occlusion, l'éclairage et la faible résolution) limitent l'efficacité des informations pour détecter les petits visages dans une scène réelle.

Aujourd'hui, la détection d'objets est au cœur de la plupart des logiciels et programmes d'IA basés sur la vision. Plus précisément la détection des visages qui consiste à détecter la présence et la localisation précise d'un ou plusieurs visages humains dans une image ou une série d'images provenant d'un dispositif de capture vidéo.

Au cours de ce travail, nous avons adopté le modèle YOLOV3 pour la détection, les résultats sont très satisfaisants. Pour une étude comparative, nous avons également tenté de tester la fiabilité de la nouvelle méthodologie du modèle récent à une seule étape YOLO5Face basé sur le détecteur YOLOV5 (une version améliorée de YOLO).

Sur la base des résultats obtenus, le modèle YOLO5Face est plus efficace par rapport au modèle YOLO_Face, bien que les résultats soit très proches.

Au vue de ces résultats, ce travail pourrait servir de guide pour entraîner le modèle YOLO5Face avec un jeu de données personnel.

Références Bibliographiques

[1] “Artificial Intelligence (AI).” 2020. IBM Cloud Education; Accès: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

[2] “Artificial Intelligence.” Built In; Accès: https://builtin.com/artificial-intelligence

[3] “Artificial Intelligence-What it is and why it matters.” Analytics Software & Solutions (SAS); Accès: https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html

[4] “Machine Learning Vs. Artificial Intelligence: Understanding The Key Differences.” 2018. Technostacks; Accès: https://technostacks.com/blog/ai-vs-ml

[5]“What is Artificial Intelligence?” IntelliPaat; Accès: https://intellipaat.com/blog/what-is-artificial-intelligence/

[6] “What is AI? Learn about Artificial Intelligence.” Oracle Cloud infrastructure; Accès: https://www.oracle.com/artificial-intelligence/what-is-ai/

[7] M. A. ALLAL. Utilisation du deep learning dans la radio cognitive, mémoire de Master professionnel en Réseaux et Systèmes Distribués (R.S.D), Université Abou Bakr Belkaid, Tlemcen, 2018.

[8] “Machine Learning: What it is and why it matters.” Analytics Software & Solutions (SAS); Accès: https://www.sas.com/en_us/insights/analytics/machine-learning.html

[9] G. Boesch. “What’s the difference between Machine Learning and Deep Learning?.” 2021.

[10] H. Trad. La détection d’objet avec OpenCV et deep learning, mémoire de Master professionnel en Réseaux et Télécommunication, Université Mohamed Khider, Biskra, 2020.

[11] What Is Machine Learning? ; Accès: https://insights.sap.com/what-is-machine-learning/

[12] S. K. Chadalawada. Real Time Object Detection and Recognition Using Deep Learning Methods, mémoire de Master professionnel en Sciences Informatique, Institut de technologie de Blekinge, 371 79 Karlskrona, Suède, 2020.

[13] “Machine Learning.” 2020. IBM Cloud Education; Accès: https://www.ibm.com/cloud/learn/machine-learning

[14] E. S. Armengol. Project of implementing an intelligent system into a Raspberry Pi based on deep learning for face detection and recognition in real-time, mémoire de Master professionnel en Sciences Informatique, Université Polytechnique de Catalogne, Bercelone, 2019.

[15] C. Deluzarche. “Deep Learning : qu'est-ce que c'est ?”

[16] “Deep Learning.” 2020. IBM Cloud Education; Accès: https://www.ibm.com/cloud/learn/deep-learning

[17] H. Mujtaba. “Face Recognition with Python and OpenCV.” 2021.

[18]“What Is Computer Vision?.” Intel Corporation; Accès: https://www.intel.com/content/www/us/en/internet-of-things/computer-vision/overview.html

[19] “What is computer vision?.” IBM Cloud Education; Accès: https://www.ibm.com/topics/computer-vision

[20] F. Shaikh. “Understanding and Building an Object Detection Model from Scratch in Python.” 2018.

[21] N. Babich. “What Is Computer Vision & How Does it Work? An Introduction.” 2020.

[22] G. Boesch. “What is Computer Vision? The Complete Technology Guide for 2021.” 2021.

[23] “An Introduction to Deep Learning.” 2016. Algorithmia; Accès: https://algorithmia.com/blog/introduction-to-deep-learning

[24] D. Y. Moualek. Deep Learning pour la classification des images, mémoire de Master professionnel en Modèle Intelligent et Décision(M.I.D), Université Abou Bakr Belkaid, Tlemcen, 2017.

[25] Y. Xia, B. Zhang, & F. Coenen. "Face Occlusion Detection Using Deep Convolutional Neural Networks." *Int. J. Pattern Recognit. Artif. Intell.* 30 (2016): 1660010:1-1660010:24.

[26] "Deep Learning for Computer Vision." Run.ai. ; Accès: https://www.run.ai/guides/deep-learning-for-computer-vision/

[27] V. Meel. "ANN and CNN: Analyzing Differences and Similarities." 2021.

[28] G. Boesch. "Deep Neural Network: The 3 Popular Types (MLP, CNN and RNN)." 2021.

[29] R. Bhatia. "Why Convolutional Neural Networks Are The Go-To Models In Deep Learning." 2018.

[30] J. Dejasmin. "Les réseaux de neurones convolutifs." 2018.

[31] A. Jain. "Deep Learning for Computer Vision – Introduction to Convolution Neural Networks." 2016.

[32] J. Cowley. "Convolutional neural networks." 2018. IBM Developer.

[33] A. B. LOUAM. Deep Learning basé sur les méthodes de réduction pour la reconnaissance de visage, mémoire de Master professionnel en Réseaux et Télécommunication, Université Mohamed Khider, Biskra, 2019.

[34] P. Blanc-Durand. "Réseaux de neurones convolutifs en médecine nucléaire : Applications à la segmentation automatique des tumeurs gliales et à la correction d'atténuation en TEP/IRM." (2018), 10.13140/RG.2.2.23231.97441.

[35] "Convolutional Neural Networks." 2020. IBM Cloud Education; Accès: https://www.ibm.com/cloud/learn/convolutional-neural-networks

[36] N. Tajbakhsh, J. Y. Shin, R. T. Hurst, C. B. Kendall & Liang J. (2017). "Automatic Interpretation of Carotid Intima–Media Thickness Videos Using Convolutional Neural Networks." Deep Learning for Medical Image Analysis, 105–131. doi:10.1016/b978-0-12-810408-8.00007-9.

[37] “Convolutional neural network.” 2020. DataScientest.

[38] Mohana & R. H V. “Object Detection and Tracking using Deep Learning and Artificial Intelligence for Video Surveillance Applications.” *International Journal of Advanced Computer Science and Applications* 10 (2019): n. pag.

[39] “Object Detection using Single Shot MultiBox Detection (SSD) and Deep Neural Network (DNN).” Accès: https://medium.com/featurepreneur/object-detection-using-single-shot-multibox-detection-ssd-and-opencvs-deep-neural-network-dnn-d983e9d52652

[40] D. Gutierrez. “Overview of the YOLO Object Detection Algorithm.” 2018. ODSC.

[41] H. Mujtaba. “ Real-Time Object Detection Using TensorFlow.” 2020.

[42] A. Neelopant, Dr. S. V. Viraktamath, P. Navalgi. “Comparison of YOLOv3 and SSD Algorithms.” 2021. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 02 (February 2021).

[43] S. Tsang. “Review: SSD — Single Shot Detector (Object Detection).” 2018.

[44] G. Boesch. “Object Detection in 2021: The Definitive Guide.” 2021.

[45] C. Kyrkou. “YOLOpeds: Efficient Real-Time Single-Shot Pedestrian Detection for Smart Camera Applications.” *IET Comput. Vis.* 14 (2020): 417-425.

[46] Y. Bouafia & L. Guezouli. “An Overview of Deep Learning-Based Object Detection Methods.” 2019.

[47] A. MAHI. Détection de visage par l’algorithme de boosting, mémoire de Master professionnel en Réseau et Système de Télécommunication, Université Abou Bakr Belkaid, Tlemcen, 2018.

[48] H. Mujtaba. “Real-time Face detection | Face Mask Detection using OpenCV.” 2021.

[49] G. Boesch. “Face Detection in 2021: Real-time applications with deep learning.” 2021.

[50]C. Bernstein. “Face detection.” 2020. TeckTarget; Accès: https://searchenterpriseai.techtarget.com/definition/face-detection

[51] H. Wang. “Real-time Face Detection and Recognition Based on Deep Learning.” 2018.

[52] F. Ye, M. Ding, E. Gong, X. Zhao & L. Hang. “Tiny Face Detection Based on Deep Learning.” 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2019, pp. 407-412, doi: 10.1109/ICIEA.2019.8834282.

[53] R. Khurana. “Face detection technology.”

[54] M. Vineetha, G. Varalakshmi, G. Bala Kumar & J. Prasad. “Face recognition system with face detection.” (2013-2017). Department of electronics and communication engineering.

[55]“YOLO v3.” PaperswithCode; Accès: https://paperswithcode.com/method/yolov3

[56]A. Chakure. “All you need to know about YOLO v3 (You Only Look Once).” 2021.

[57] R. De Palma. “YOLOv3 Architecture: Best Model in Object Detection.”

[58] “A Quick & Easy Way Of Knowing Deep Learning In Python Especially For Beginners.” 2019. Technostacks; Accès: https://technostacks.com/blog/deep-learning-with-python

[59]“Why did we build Visual Studio Code?” 2021. Visual Studio Code; Accès: https://code.visualstudio.com/docs/editor/whyvscode

[60] D. Johnson. “What is TensorFlow? How it Works? Introduction & Architecture.” 2021.

[61] “OpenCV Python Tutorial.” GeeksforGeeks. 2021.

[62] “Introduction to OpenCV-Python Tutorials.” ; Accès: https://docs.opencv.org/master/d0/de3/tutorial_py_intro.html

[63]“Numpy, Matplotlib & Scipy Tutorial.” Python-Course; Accès: https://www.python-course.eu/numpy.php

[64] J. Brownlee. “Your First Deep Learning Project in Python with Keras Step-By-Step.” 2021.

[65] “Keras: Deep Learning library for Theano and TensorFlow.” ; Accès: https://faroit.com/keras-docs/1.2.0/

[66] A. Johari. “Matplotlib Tutorial – Python Matplotlib Library with Examples.” 2021.

[67] K. Meghna. “Python | Introduction to Matplotlib.” GeeksforGeeks. 2018.

[68] “Python Pillow - Quick Guide.” Accès: https://www.tutorialspoint.com/python_pillow/python_pillow_quick_guide.htm

[69] “Python Imaging Library (PIL).” Accès: https://he-arc.github.io/livre-python/pillow/index.html#qv

[70] “WIDER FACE: A Face Detection Benchmark.” Accès: http://shuoyang1213.me/WIDERFACE/

[71] S. Wang & K. Wang. “Real-Time and Accurate Face Detection Networks Based on Deep Learning.” 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), 2019, pp. 1541-1545, doi: 10.1109/EITCE47263.2019.9094843.

[72] J. Solawetz. “YOLOv5 New Version - Improvements And Evaluation.” 2020.

[73] D. Qi, W. Tan, Q. Yao & J. Liu. “YOLO5Face: Why Reinventing a Face Detector.” (2021). ArXiv, abs/2105.12931.