République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université Larbi Tébessi - Tébessa
Faculté des Sciences Exactes et des Sciences de la Nature et de la vie

Département : Mathématiques et Informatique

Mémoire de fin d'étude

Pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Option : Système d'information

Thème

Apprentissage profond pour la classification d'actions dans les vidéos de football

Présenté Par :
Sadi Adnene

Devant le jury :

Mr A.Gattal	*MCA*	*Université Larbi Tébessi*	*Président*
Mr I.Bendib	*MCB*	*Université Larbi Tébessi*	*Examinateur*
Mr A. Sahraoui	*MCB*	*Université Larbi Tébessi*	*Encadreur*
Mr M. Gasmi	*MCB*	*Université Larbi Tébessi*	*Co-Encadreur*

Apprentissage profond pour la classification d'actions dans Les vidéos de football

Sadi Adnene

Université de Larbi Tébessi

Résumé

L’apport technologique et l’apparition des nouvelles techniques en intelligences artificielles à beaucoup aider les chercheurs des autres sciences à atteindre leurs objectifs ciblé dans leurs recherches. Parmi les domaines influencer cet apport technologique la science du sport et plus précisément le monde de football.

Les techniciens du domaine de football utilisent beaucoup plus l’analyse vidéo pour lire la façon du jeu que ce soit de leurs équipes afin de les mettre en bon emplacement ou des équipes adverses dans le but de les mettre en échec.

Le but de ce travail est la conception et l’implémentation d’une approche basée sur l’apprentissage profond et les réseaux de neurones convolutionnelles(CNN) pour la classification des tactiques des jeux à partir de séquences vidéo. Les démarche utilisé compte sur deux étapes, la première concerne la détection des joueurs et leurs emplacements dans le terrain a partir des frames, et la deuxième la classification de ces frames afin de prédire le tactique du jeu appliqué par l’équipe choisie.

Mots clé : vidéo, traitement vidéo, football, analyse des vidéos, apprentissage profond, CNN, détection des objets, classification.

Abstract

The technological contribution and the appearance of new techniques in artificial intelligence greatly help researchers from other sciences to achieve their targeted objectives in their research. Among the fields influencing this technological contribution is the science of sport and more precisely the world of football.

Football technicians use video analytics a lot more to read the way the game is played by their teams to put them in the right place or to opposing teams in order to check them.
The aim of this work is the design and implementation of an approach based on deep learning and convolutional neural networks (CNN) for the classification of game tactics from video sequences. The approach used has two stages, the first concerns the detection of players and their locations in the field from the frames, and the second the classification of these frames in order to predict the tactics of the game applied by the chosen team.

Keywords: video, video processing, football, video analysis, deep learning, CNN, object detection, classification.

مـلـخـص

تساعد المساهمة التكنولوجية وظهور التقنيات الجديدة في الذكاء الاصطناعي بشكل كبير الباحثين من العلوم الأخرى على تحقيق أهدافهم المستهدفة في أبحاثهم. من بين المجالات التي تؤثر على هذه المساهمة التكنولوجية ، علم الرياضة وبشكل أدق عالم كرة القدم يستخدم فنيو كرة القدم تحليلات الفيديو كثيرًا لقراءة الطريقة التي تلعب بها فرقهم اللعبة لوضعها في المكان المناسب أو للفرق المنافسة للتحقق منها

الهدف من هذا العمل هو تصميم وتنفيذ نهج قائم على التعلم العميق والشبكات العصبية التلافيفية لتصنيف تكتيكات اللعبة من تسلسلات الفيديو. يتألف المنهج المستخدم من مرحلتين، الأولى تتعلق بالكشف عن اللاعبين ومواقعهم في الميدان من الإطارات، والثانية تصنيف هذه الإطارات من أجل التنبؤ بأساليب اللعبة التي يطبقها الفريق المختار

الكلمات المفتاحية : الفيديو ، معالجة الفيديو ، كرة القدم ، تحليل الفيديو ، التعلم العميق ، الشبكات العصبية التلافيفية ، كشف الأشياء ، التصنيف.

Dédicace

Je dédie mon modeste travail

A mes chers parents Pour leur grand soutien indéfini, Pour leur amour et leur encouragement.

Leurs sacrifices que rien au monde ne peut les récompenser et sans lesquels je n'aurais jamais pu parcourir ce chemin.

Que Dieu vous protège.

A mes sœurs " Nourhene" et " Sara "
Et toute ma grande famille maternelle et paternelle.

A mes chers frères et sœurs Pour leur amour et leur incontestable appui.
A mes chers ami(e)s.surtout " Zied & Mohamed ".

Sans oublier tous les enseignants que ce soit du primaire, du moyen, du secondaire ou de l'enseignement supérieur

A toutes les personnes chères à mon cœur je dédie ce travail.

Sadi Adnene

Remerciement

Mes remerciements vont tout premièrement à dieu tout puissant pour la volonté, la santé et la patience qui m'a donné durant tous ces années d'études.

Je remercie vivement les membres de jury qui ont accepté d'examiner et pour l'honneur qu'ils me font en jugeant mes travaux et pour avoir accepté de participer au jury de ce travail.

Je tiens à présenter mes reconnaissances à mes encadrantes Dr.A. SAHRAOUI, Dr.G.MOHAMED pour leurs conseils, leurs orientations, et la disponibilité qu'il m'a accordées pour faire réussir ce travail, tout au long de la période.

Je profite cette occasion pour exprimer ma gratitude à ces personnes pour leurs contributions comme mes collègues, ma famille et mes amis.

Je voudrai exprimer tout d'abord ma gratitude à ma belle-famille et sur tout, à ma mère, mon père, mes sœurs. Ils sont toujours intéressées à mes études avec le succès et m'ont aidé beaucoup plus qu'elles ne peuvent le croire.

Je voudrais également adresser mes plus sincères remerciements à tous mes proches et amis qui m'ont toujours soutenue et encouragée au cours de la réalisation de cette mémoire. Vraiment merci.

Merci à vous tous !

A tous qui j'aime…… !

A tous qui m'aiment…… !

TABLEAU DES MATIERES

CHAPITRE 01 Introduction au Traitement vidéo

CHAPITRE 02 Les Architectures d'apprentissage Profond Pour les Tâches de Traitement Vidéo

CHAPITRE 03 Conception Et Implémentation d'un Modèle de Détection Et De Classification

LISTE DES Tables

Page

LISTE DES FIGURES

Page

Introduction Générale

Au cours des dernières années, le football devenu le sport le plus populaire au monde attire généralement des millions de personnes pour les regarder au niveau international. Le football moderne se présente comme un système dynamique, où il existe une relation de compétition entre les équipes et la coopération des membres de la même équipe. Cette modalité est caractérisée comme un système complexe, dans lequel l'organisation se déroule à plusieurs niveaux connaît sous l'analyse tactique.

En pratique, après avoir fait un diagnostic a votre équipe et avoir lit le positionnement de l'équipe adverse, il est important de fournir à sa propre équipe "les médicaments" juste au moment opportun de façon à en améliorer " l'état de santé ".

La situation optimale pour une équipe de football se vérifie quand la supériorité en phase défensive et en phase offensive est nette. A cet effet l'entraîneur ne devra pas faire autre chose que vérifier constamment le maintien d'une telle situation (idéale) et n'intervient habituellement que s'il y a une obligation.

L'analyste vidéo dans le football devient indispensable au sein d'un club. Les outils numériques permettent d'être proactif dans l'analyse de votre équipe ou des adversaires. L'analyste vidéo visionne les matchs et découpe ces derniers en plusieurs de jeu. Un échange a alors lieu entre l'analyste vidéo et le coach. Il permet au coach de mettre en place des ajustements tactiques. Cette interprétation réalisée par l'analyste vidéo peut concerner les matchs des équipes adverses, afin de les mettre en échec.

L'apport technologique et l'apparition des nouvelles techniques en intelligences artificielles à beaucoup aider les équipes et les coaches dans son rôle. Le but de notre travail est de donner plus d'aide aux entraîneurs en se basant sur les techniques de l'apprentissage profond.

Notre travail focalise sur la lecture des tactiques du jeu appliqué par une équipe de football, la chose qui va donner plus d'information aux entraîneurs que ce soit pour mettre à jour l'emplacement de son équipe ou pour mieux affronter l'équipe adverse.

Notre idée consiste à développer deux approches pour 3 buts complémentaires. La première approche consiste à la conception et l'implémentation d'un modèle de détection des joueurs et lors emplacement dans le terrain à partir de séquences vidéo. Le premier but est de générer une Dataset pour l'apprentissage et le deuxième pour générer des frames pour la classification des tactiques du jeu. La deuxième approche est de définir un modèle de classification à base de l'apprentissage profond dans le but est de prédire les tactiques du jeu à partir des séquences vidéos.

Ce travail est très difficile par rapport l'entraîneur et prend beaucoup de temps pour traiter les problèmes des plans défensifs offensifs (L'analyse des systèmes de jeu) ; mais avec naissances de l'intelligence artificielle, La recherche sur les techniques de jeu dans une vidéo a attiré beaucoup d'attention sur les techniques d'apprentissage profond pour faciliter et augmenter l'analyse des vidéos de footballs pour les entraîneurs des équipes ;pour cela , La classification des vidéos de foot a une importance considérable pour le traitement vidéo et devient de plus en plus vaste, diversifiée et partagée. Il attire de nombreux chercheurs ces dernières années par ce qu'il représente une part importante du contenu multimédia disponible dans le cyberespace et explorée largement en raison des avantages commerciaux potentiels et de l'audience massive dans le monde entier mais il devient toujours une tâche difficile.
L'objet de ce mémoire est organisé en trois chapitres :

- Dans le premier chapitre nous allons présenter les notions de base sur les vidéos, le processus de numération, la compression vidéo etc. Ensuite, nous allons présenter la classification vidéo dans l'intelligence artificielle, après nous allons mettre en évident un aperçu sur l'apprentissage profond et leurs techniques.

- Le deuxième chapitre a été consacré à la description des modèles de base de l'apprentissage profond ainsi que ces principales fonctionnalités dans la classification des vidéos. Ce chapitre finira par une étude comparative entre les architectures de classification des vidéos.

- Dans le troisième chapitre on a le consacré pour parler du domaine d'application ou on a donné des notions de bases sur les systèmes et les tactiques du jeu en football, puis on a entamé la conception et l'implémentation de nos deux approches de détection et de classification.

CHAPITRE 01

INTRODUCTION AU TRAITEMENT VIDÉO

1. La modalité vidéo.
2. La classification des vidéos.
3. Problèmes de traitement vidéo.

INTRODUCTION

Aujourd'hui, le traitement multimédia se concentre principalement sur trois types de données, à savoir les images, la parole et le texte. Ces types de données sont valorisés pour l'analyse, la classification, la reconnaissance et la détection de contenu. En raison de ses caractéristiques particulières, la vidéo est devenue une modalité plus demandée et nécessairement omniprésente sur les réseaux sociaux et les plateformes de téléchargement ces dernières années. Les données vidéo disponibles en ligne contiennent une grande quantité de données, ce qui rend leur traitement et leur partage un gros problème. Cela encourage le développement d'algorithmes et de plateformes de traitement capables d'analyser leur contenu sémantique pour diverses applications.

Dans ce chapitre, nous allons mettre en évidence le type de données vidéo en présentant ces définitions, son processus de numérisation, la compression vidéo et ses caractéristiques fondamentales. Ensuite, nous allons présenter les techniques modernes la de classification vidéo. Après, nous allons discuter les problèmes de la classification vidéo.

1. La modalité vidéo

1.1. Définition de la vidéo

D'après Christophe Savariaux **[1]**, la vidéo est considérée comme une source visuelle qui combine une succession d'images animée qui défilent à un rythme fixe (cadence) pour donner l'illusion du mouvement (25 image/sec pour le PAL). En d'autres termes, la vidéo transmet un signal à un écran et traite l'ordre dans lequel les captures d'écran doivent être affichées. Les vidéos ont généralement des composants audio qui correspondent aux images affichées à l'écran. De plus, c'est la technologie de capture électronique des successions d'images accompagnées par des textes, son ou mouvements à une certaine cadence. C'est un ensemble de scènes animées tel que : Les documentaires, match de sport, Films, télésurveillance, etc.

1.2. La séquence de vidéo

La séquence vidéo est une source d'informations visuelles sous la forme d'arrangements de clips vidéo, audio, texte et graphiques dans un ordre chronologique. Les informations contenues dans un clip vidéo sont beaucoup plus riches que les informations incluses dans une seule image. Une séquence peut enregistrer plusieurs vues du mouvement, dans lesquelles la dynamique est sauvegardée en raison de l'apparition de certains objets et de la disparition d'autres, tandis qu'une seule image ne fournit qu'une seule scène de mouvement.

Le cerveau de l'être humain peut traiter la dynamique des scènes afin de reconnaître les objets dès qu'ils se déplacent, classer le type de leur mobilité, même s'ils ne sont pas très visibles, décidé de l'interaction faciale, prédire une sensation, etc. Cette dynamique se traduit donc par un signal très fort dans l'intervalle visuel et acoustique de l'être humain. La science du traitement de la dynamique des scènes est donc très utile pour le traitement vidéo. Cette convergence est utile pour deux raisons : la relation entre les objets d'une même image et les caractéristiques des images. D'une part, la dynamique des scènes comprend de nombreuses informations sur les relations spatiales et temporelles entre les objets. Ceci est important pour la détection, l'identification, comme les applications de sécurité et la surveillance. D'un autre côté, les caractéristiques de l'image incluent l'intensité (c'est-à-dire la couleur), sa taille (définition) en points ou pixels, ses dimensions réelles (en centimètres ou en pouces) et sa résolution (pixel / pouce). Ces caractéristiques ont une relation très forte avec la direction de la dynamique de la scène. En d'autres termes, pour chaque objet image, ces caractéristiques ne changent pas pendant la scène. Ceci est utile pour supprimer les images en double et conserver une taille optimale pour la vidéo. Une scène, qui contient des images individuelles dans un plan, a la hiérarchie suivante :

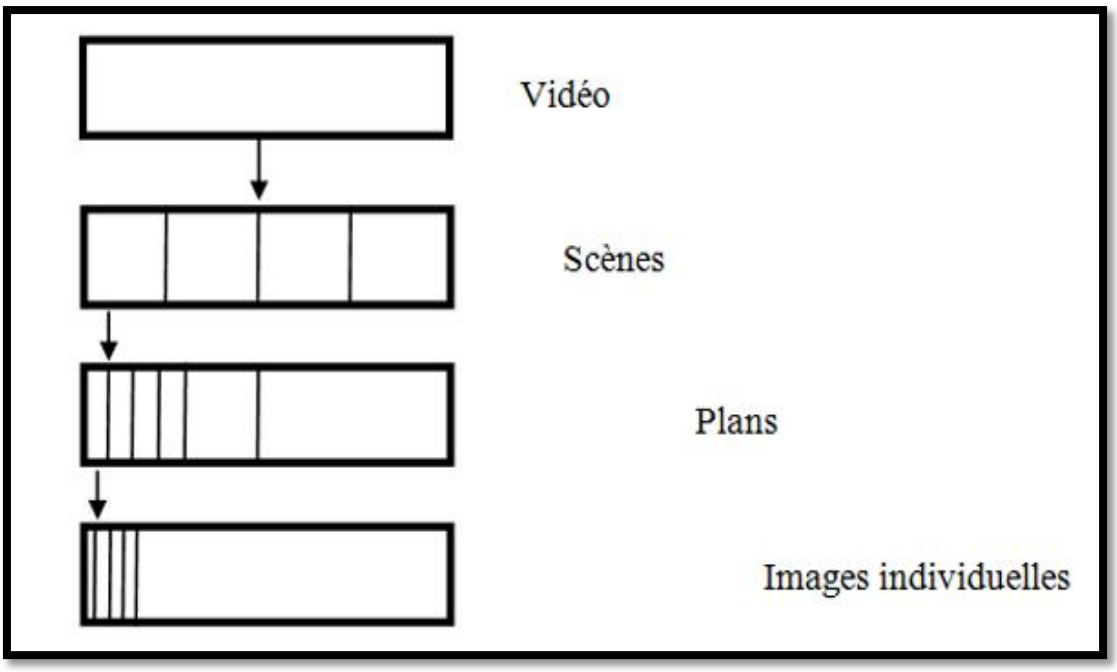

FIGURE 1.1- La hiérarchique d'une vidéo **[48]**

La séquence vidéo est caractérisée par trois paramètres suivants :

- **Résolution en luminance**

  C'est le nombre de nuances ou de couleurs possibles pour un pixel ("8 bits pour les niveaux de gris et de 24 bits pour les séquences en couleurs").

- **Résolution spatiale**

  La résolution spatiale est le nombre de lignes et de colonnes de la matrice de pixels dans la vidéo.

- **Résolution temporelle**

  Se distingue la fluidité (vitesse) d'une vidéo par le nombre d'images par secondes (en anglais frame rate), exprimé en FPS (Frames per second).

La valeur de ces trois paramètres détermine l'espace mémoire nécessaire pour stocker séquence vidéo. Ce tableau résume l'hiérarchique de séquence de vidéo :

TABLEAU 1.1 – La hiérarchique de la vidéo.

Structure	Définition
Scène	Un ensemble d'événement ou une partie d'une pièce de théâtre, film programme télévisé, etc. Ex : Scène de crime.
Plan	Un ensemble d'images qui représente le même objet ou appartient au même intervalle spatio-temporelle
Image	Unité de Base d'un plan. Chaque plan est composé d'un ensemble d'images individuelles.

1.3. Le processus de numérisation d'une vidéo

La numérisation fait aujourd'hui partie intégrante de notre vie quotidienne, alors que le contenu sur Internet continue de progresser et de s'améliorer. L'environnement numérique est l'avenir des médias qui évoluent progressivement vers ce monde. Les supports analogiques tels que les bandes vidéo, les disquettes, les enregistrements audio / vidéo analogiques, convergent lentement vers les formats numériques et seront disponibles sur les supports numériques. L'avantage de la numérisation est la capacité de manipuler et de traiter les données stockées sur des supports numériques, en utilisant une intervention humaine minimale et un temps de traitement réduit au niveau de la machine.

Par exemple, avec l'évolution de la technologie numérique, le temps nécessaire au montage d'un film cinématographique d'une heure sous forme analogique est passé de quelques semaines à quelques heures. De plus, la créativité peut être étendue à n'importe quelle limite en un rien de temps pour une sortie de meilleure qualité qui peut ne pas être possible tout en travaillant avec des médias analogiques. Les derniers progiciels et systèmes de montage vidéo ont rendu le montage vidéo complexe et compliqué simple et facile **[2]**.

La transition de l'analogique au numérique est la grande révolution technologique qui a eu lieu au milieu du dernier siècle. Avant la révolution, les signaux du monde qui nous entourait, tels que le son, la vidéo, les ondes radio, les appels téléphoniques, la communication, l'écriture, étaient analogues. Au cours des années 1960, il est obligatoire pour l'être humain de pouvoir mettre en évidence les avantages de la transformation de signaux analogiques en signaux numériques **[3]**. Après la révolution, ces signaux sont interprétés au format numérique (notation

binaire) et même pour les processus de transmission et de communication. Cette transformation se décompose en trois étapes : *l'échantillonnage*, La *quantification* et le *codage.*

1.3.1. L'échantillonnage

L'échantillonnage est un processus qui se produit lors de la numérisation et qui consiste à la transformation d'un signal analogique en un signal numérique. Autrement dit, un convertisseur transforme le signal électrique qui est un signal continu, en petites successions d'événements qui s'enchaînent à intervalles (des instantanés) de temps réguliers. Lorsqu'une source génère un signal analogique et ce signal doit être transformé en signal numérique (notation binaire ; 0 ou 1), ce signal doit être discrétisé dans le temps. Cette discrétisation du signal analogique est appelée échantillonnage.

La figure suivante indique un signal à temps continu x(t) et un signal échantillonné *x*. Lorsque *x* est multiplié par un train d'impulsions périodique, le signal échantillonné x_s est obtenu.

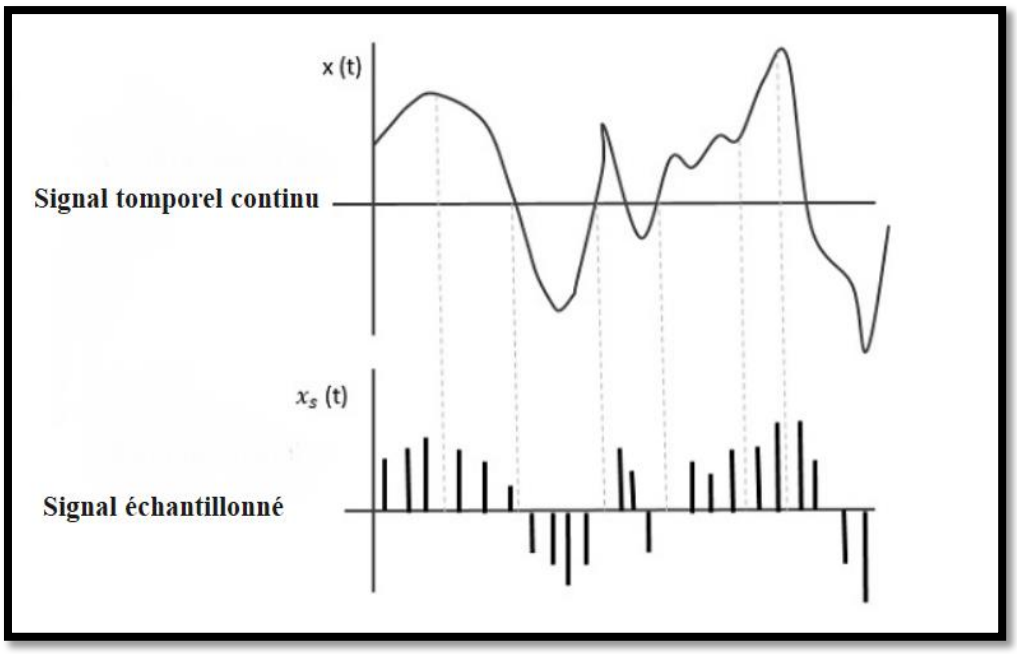

FIGURE 1.2- Le processus d'échantillonnage.

Selon le *théorème de Shannon*, il faut double (égale ou supérieur) a la fréquence de signal analogique que l'on échantillonné Pour pouvoir numériser correctement un signal. On a :

Fe ≥ 2fmax.

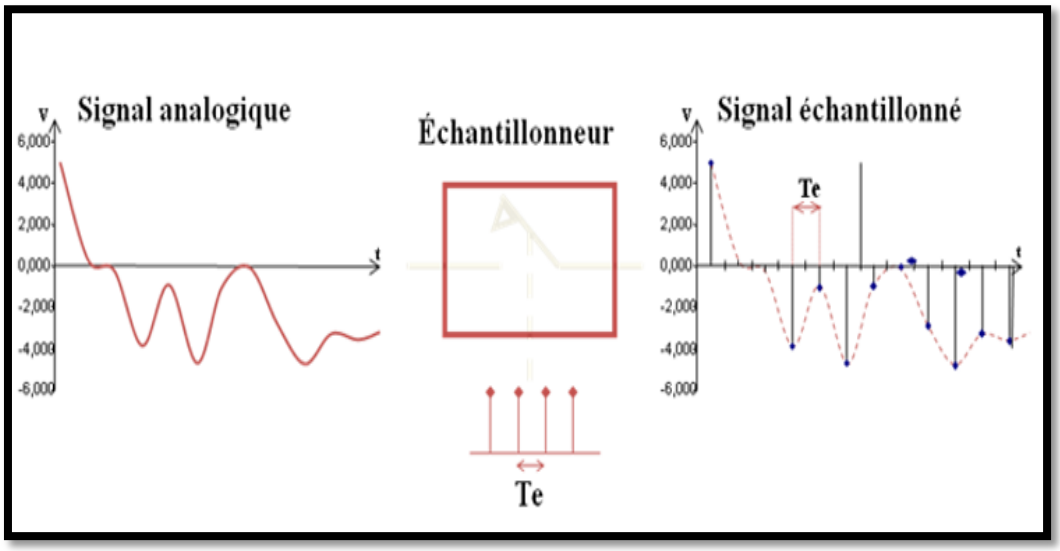

FIGURE 1.3- L'échantillonnage d'une vidéo. **Te :** période d'échantillonnage, intervalle de temps séparant deux échantillons. **Fe :** fréquence d'échantillonnage, nombre d'échantillons pendant une seconde.

1.3.2. La Quantification

La quantification est une opération réalisée après le processus d'échantillonnage d'un signal vidéo. Elle consiste à attribuer et affecter un nombre binaire à chaque échantillon prélevé au signal lors l'échantillonnage.

Elle s'exprime en bits (la plus petite unité de numérisation). Cette quantification est assurée par un Convertisseur Analogique/Numérique (CAN). Les niveaux de tension possibles ne dépendent que du nombre de bits **n** utilisé par la quantification. Chaque bit pouvant prendre deux valeurs (0 ou 1). Donc un convertisseur à **n** *bits* possède 2^n niveaux de quantification, comme montre la figure suivante :

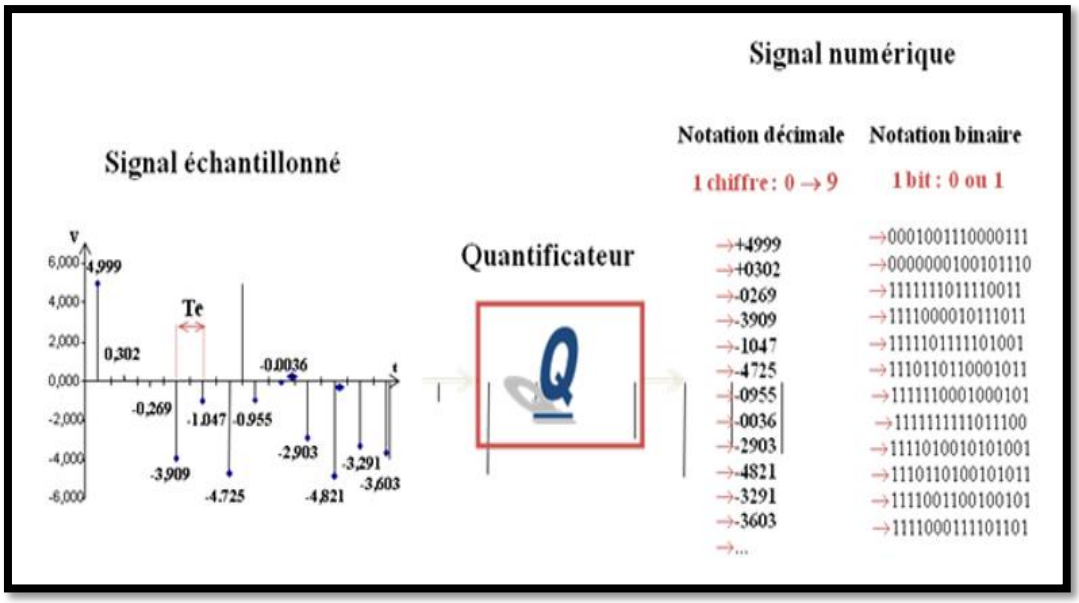

FIGURE 1.4- Quantification d'un signal échantillonné.

La qualité d'un signal numérique dépend de deux facteurs :

- **La fréquence d'échantillonnage (Fe)** : Elle est exprimée en *Hertz* [*Hz*]. Plus que les nombres échantillons relevés de petit intervalle de temps, plus le signale numérique est en bon qualité
- **Nombre de bit dans lequel ont codé les valeurs (la résolution)** : Plus la valeur de **n** bit est élevée, moins de perte d'information (qualité de signal) et recto versa.

1.3.3. Codage

Dans cette étape Chaque échantillon sera codé sur un ensemble de bit. Comme dans l'exemple suivant :

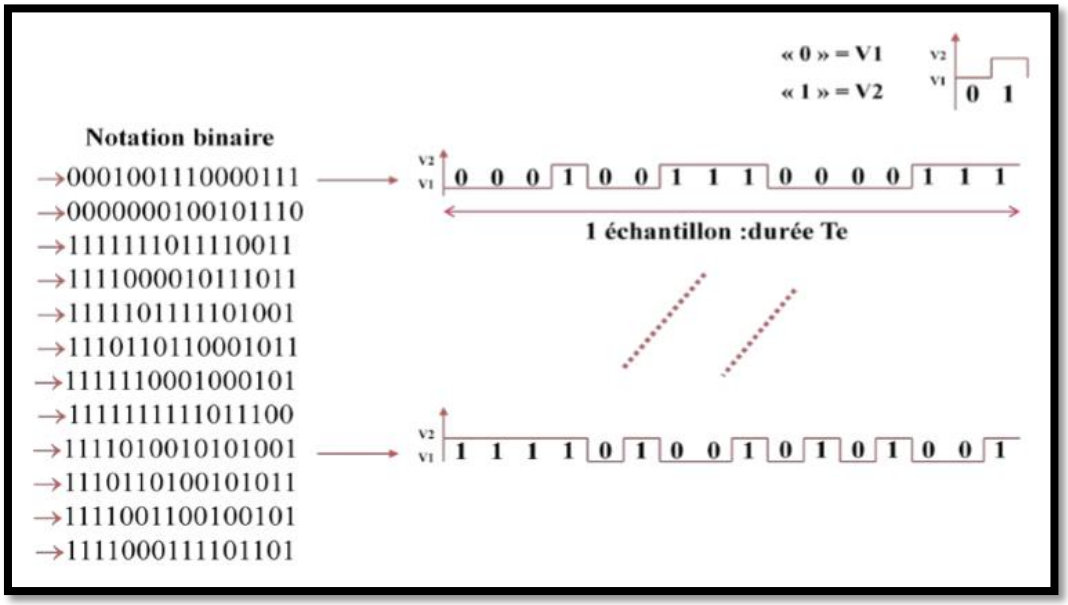

FIGURE 1.5- Codage d'un signal vidéo numérique

1.4. La compression vidéo

Comme nous l'avons vu dans les sections précédentes, les sources d'informations visuelles sur Internet sont multiples, la vidéo est la plus intéressante en particulier. Selon les prévisions du Cisco Global Cloud Index (CGCI), le contenu vidéo contribue à plus de 80% du trafic Internet aujourd'hui **[4]**, et ce pourcentage est en constante augmentation. En effet, il est impératif de construire des protocoles de compression vidéo standard efficaces pour générer des flux vidéo préservant la bonne qualité des informations avec un coût financier minimal et prenant en charge des bandes passantes fixes.

De plus, les connaissances extraites par vision artificielle sont basées sur un traitement efficace de vidéos comme la détection d'objets ou la détection de mouvement **[5]**. La qualité de ces connaissances dépend de la qualité des vidéos compressées ou du protocole de compression choisi. Une bonne qualité de compression peut offrir des avantages pour d'autres expériences

futures de la vision artificielle.

Ainsi, les techniques de compression sont également utiles pour la reconnaissance d'actions dans la vidéo **[6]**.

Par conséquent, la compression vidéo est une étape essentielle pour le traitement vidéo, c'est une opération qui consiste à réduire la quantité initiale de données d'un fichier numérique (son, vidéo, image); en minimisant également l'impact sur la qualité vidéo de la vidéo; sur la suppression de la redondance tout en préservant la qualité perceptible, le fichier compressé peut être envoyé efficacement sur le réseau en respectant la bande passante fixé ou stocké efficacement sur des disques informatiques.

1.4.1. Les types de vidéo

Dans la littérature, il y a deux domaines pour transporter et stocker des données, c'est l'analogique et le numérique. L'analogique est né avec le début de l'électricité tandis que le numérique est apparu plus récemment avec l'ère de l'informatique [7].

- **La vidéo analogique**

  La vidéo analogique est un signal vidéo qui peut prendre n'importe quelle valeur entre deux extrémités. Le principe de l'analogique est de reproduire le signal à enregistrer sous une forme similaire sur un support magnétique. Le signal est en forme d'onde et est très sensible aux perturbations externes. Ces perturbations peuvent entraîner une modification importante du signal **[8]**. Il existe plusieurs normes pour la vidéo analogique (PAL, NTSC, SECAM, DIGITAL8).

- **La vidéo numérique**

  La vidéo numérique consiste en une succession d'images numériques. Il contient une technique d'enregistrement, de traitement et de reproduction d'images vidéo sous forme de données numériques. La vidéo numérique utilise un signal de type binaire. Cela signifie que le signal vidéo qui, à un moment donné, ne peut prendre qu'une des deux valeurs 0 et 1. Ainsi, chaque image de la vidéo est traduite en une succession de 0 et 1.

1.4.2. Les caractéristiques de la vidéo numérique

- **La Cadence**

  C'est le nombre d'images par seconde (FPS). L'œil humain a la particularité de pouvoir distinguer environ 20 images par seconde. Ainsi, en affichant plus de 20 images par seconde (vidéo), il est possible de tromper l'œil et de lui faire croire en une image animée.

- **Résolution de l'image**

Est le nombre de pixels qui composent l'image en hauteur (axe Vertical) et en largeur (axe horizontal).

La résolution définit la netteté et la qualité d'une image, c'est-à-dire. Plus la résolution est élevée, plus l'image est précise en détail.

- **Pixels (Picture elements- : en anglais)**

Est l'élément de base d'une image ou d'un écran, c'est-à-dire le plus petit élément que l'on puisse trouver dans une image. Chaque pixel a ses propres caractéristiques (couleur, luminosité, luminosité, position).

- **Format D'image**

Est une représentation informatique sur la façon dont l'image est codée (jpg, png, etc.).

- **Espace colorimétrique**

Est une gamme définie de couleurs. Les espaces colorimétriques connus qui sont RGB, Adobe RGB et ProPhotoRGB.

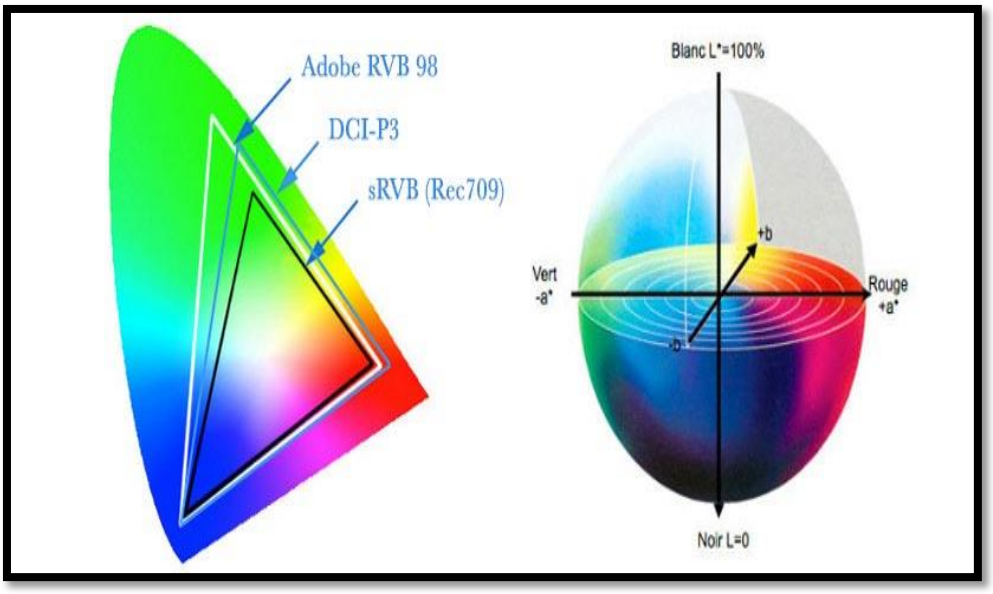

FIGURE 1.6- L'espace colorimétrique. **[49]**

- **Conteneur (extension de fichier)**

Est un format de fichier qui contient les flux vidéo et audio ainsi que tous les autres éléments associés. Parmi les Conteneurs vidéo les plus courants on peut citer : .mp4, .mov, .avi, .mkv, .flv, .mpeg, etc.

- **Codec (codage--décodage)**

Est un procédé (un circuit imprimé ou un logiciel) de compression et/ou de décompression d'un signal numérique **[9]**. Il s'agit d'applications qui interviennent dans le traitement de la vidéo numérique afin de réduire le volume de mémoire occupé par celle-ci suivant la qualité d'image que l'on veut obtenir.

- **Débit**

C'est la quantité d'informations qui "passe" chaque seconde dans notre vidéo, par exemple dans l'extension DV la quantité qui transmis est 25mbps (Mēga bit par seconde) ça veut dire dans 1h = 12,5 go.

- **Support**

C'est l'enregistrement et diffusion de la vidéo. Le support est le seul garant de la sécurité des données**.**

- **Entrelacement vidéo (balayage entrelacé)**

C'est un mode de capture et de diffusion de vidéo où chaque image est constituée par deux trames (Field), capturés à l'origine à 2 moments différents t et t + 20ms.

Un signal vidéo est entrelacé : signifie que chaque écran d'information est en fait composé de deux champs différents : le champ impair et le champ pair. Les lignes impaires sont tracées en premier sur l'écran. Ensuite, les lignes paires sont tracées entre les lignes impaires avant d'être effacées. Cela se produit à une vitesse beaucoup plus élevée que celle de la perception de l'œil humain **[10]**.

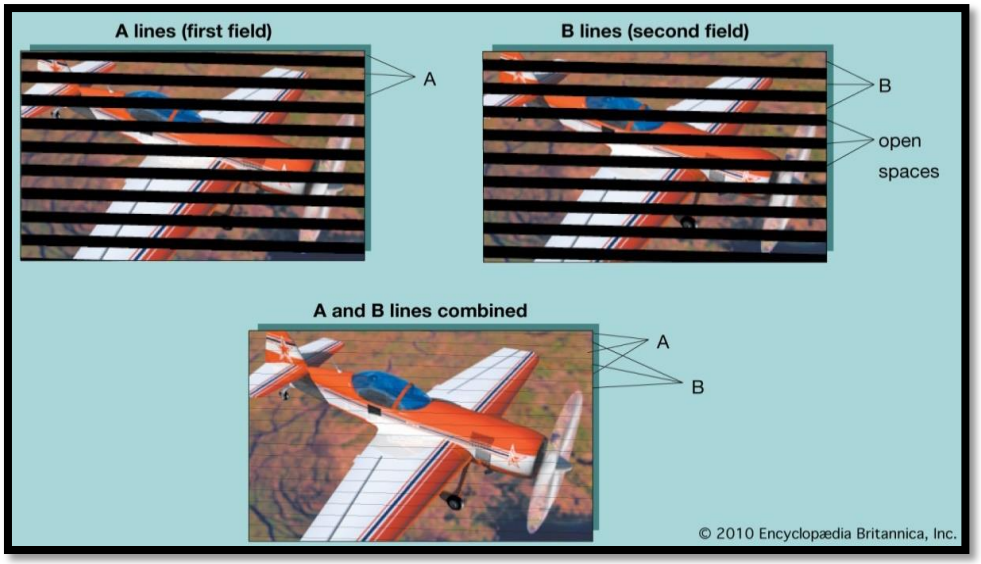

FIGURE 1.7- L'entrelacement d'une vidéo **[50]**.

1.4.3. Les formats de fichier vidéo

Un format de fichier vidéo est un type de fichier pour stocker des données vidéo numériques sur un système informatique. La vidéo est presque toujours stockée à l'aide d'une Compression avec perte pour réduire la taille du fichier.

1.4.3.1.Les codecs vidéo

TABLEAU 1.2 – Les codecs vidéo

Codecs	Définition	Avantages
MPEG-PS (MPEG Program Stream) **Extension :** .mpg .mpeg .m2p .ps	Est un format Développé par le Moving Picture Experts Group. Il est spécifié dans la norme MPEG-1 et MPEG-2	▪ Utilisé pour diffusion, surtout pour la télévision, et pour le stockage sur DVD ▪ Ce format comprend des fonctions de correction d'erreurs.
MPEG-1 **Extension :** .mpeg, .mpe .mpg .mpv .dat	Crée et élaborer en 1991 par le groupe MPEG (Motion Picture Expert Group)	▪ Une norme standard pour la compression des données vidéo et des canaux audio.
MPEG-2 **Extension:** .mpeg .mpe .mpg .mpv .mp2 .m2p .vob.	C'est un format très conçu remplacé le format MPEG-1	▪ Utiliser par des nombreux lecteurs multimédias sur ordinateur ▪ Utiliser pour la diffusion de programmes de télévision numérique
H.**264 ou MPG4** **Extensions :** .MP4 . M4P .M4B .M4R .M4V .M4A .DIVX .F4V .FLV	Conçu en 1998 par le groupe MPEG	▪ Format populaire utilisé pour enregistrer des vidéos au format HD ou SD ▪ Pouvant être compressées pour ne pas occuper trop de mémoire ou d'espace de stockage

1.4.3.2.Les conteneurs vidéo

TABLEAU 1.3 – Les conteneurs vidéo

Conteneurs	Définition	Avantages
AVI (Audio Vidéo Interleave) **Extension :** .AVI	Est l'un des plus anciens formats conteneur, créé par Microsoft pour Windows.	▪ Format très utilisé pour tous les appareils fonctionnant sous Android et compatible pour Mac ou Linux. ▪ Ce format permet de compressé par n'importe quel codec.
MP4 (Moving Picture 4) **Extension:** .MP4	C'est un format de vidéo créé en 2004 il utilise par un nombre croissant de caméras.	▪ Utilisé et accepté par la majorité des ordinateurs, Smartphones, télévisions ou lecteurs de DVD. ▪ Son système de compression permet d'obtenir des fichiers plus légers en supprimant les images fixes.
MKV (Matroska Vidéo) **Extensions :** .Mkv .mka .mks .mk3d	Format conteneur russe, libre, créé en 2003, pouvant contenir de très nombreux codecs	▪ Réaliser des fonctions de chapitrale ▪ Créer des menus ▪ Faire des recherches dans le fichier et sélectionner une source sonore ▪ Compatible avec des lecteurs matériels, mais aussi avec des lecteurs logiciels.
WMV (Windows Media Player) **Extensions:** .wmv	Développé par Microsoft.	▪ C'est un format de codage des données multimédia ▪ Compatible avec le Blu-ray mais essentiellement pour être lu par Windows Media Player
FLV (Flash vidéo) **Extensions :** .flv	C'est l'un des formats conteneur populaire de partage sur le web utilisé par Adobe Flash Player.	▪ Conserver un haut degré de qualité de diffuser des vidéos Internet ▪ Permet d'obtenir des fichiers peu volumineux
MOV (QuickTime) **Extensions :** .mov .qt, .qtx .qtr.	Format conteneur créé en 1989 par Apple et mis sur le marché en 1991.	▪ D'encoder plusieurs pistes de plusieurs types comme : audio, vidéo, texte (pour les sous-titres). ▪ Compatible aussi bien pour Mac que pour PC, et fonctionne sur les Smartphones iOS ▪ Utilisé par son logiciel de lecture de vidéo, QuickTime Player ▪ Les fichiers dans ce format sont légers

1.4.4. Les applications de traitement vidéo

1.4.4.1.Le suivie de mouvement des objets :

Le suivi d'objets dans des séquences vidéo est une piste de recherche de plus en plus importante en traitement vidéo et dans de nombreux domaines de la vision artificielle. L'objectif du suivi de flux vidéo est de localiser la position et le mouvement d'objets ou de personnes au fil du temps en reliant les instances du même objet à celles détectées dans les images précédentes de manière continue et fiable dans les images successives. Les objets sont souvent des personnes, mais peuvent également être des animaux, des véhicules ou d'autres objets d'intérêt, comme le ballon dans un match de football **[11]**.

Le suivi permet non seulement la détection des Objets mais pourraient produire autres informations telles que la trajectoire, la vitesse et la direction des objets à suivre. En effet, il existe deux principaux types de suivi d'objets :

- **Suivi d'objets hors ligne** : suivi d'objets sur une vidéo enregistrée
- **Suivi d'objets en ligne** : suivi d'objets effectué sur un flux vidéo en direct, par exemple, une caméra de surveillance.

1.4.4.2. La détection d'objet/Personne :

La détection d'objets a considérablement évolué au cours des deux dernières décennies, avec la transition des approches traditionnelles d'apprentissage statistique ou automatique vers les approches d'apprentissage approfondi **[12]**. La détection joue un rôle très important pour de nombreuses applications de traitement d'image et de vision par ordinateur, elle dépend généralement de la nature de la scène et de l'application correspondante car elle consiste à percevoir une scène statique ou dynamique. Les techniques de détection les plus utilisées sont :

- Méthodes de soustraction de fond
- Détection d'objet basée sur la segmentation de l'image
- Détection d'objet basée sur une reconnaissance de forme Reconnaissance des objets

1.4.4.3. La reconnaissance d'objet :

La reconnaissance d'objets est un terme général pour décrire un ensemble de tâches de vision par ordinateur connexes qui impliquent l'identification d'objets dans des images numériques après la détection et la localisation de cet objet **[13]**. L'emplacement fait référence à l'identification de l'emplacement d'un ou plusieurs objets dans des images successives. Les objets sont souvent des personnes, mais peuvent également être des animaux, des véhicules ou d'autres objets d'intérêt, comme le ballon dans un match de football.

2. La classification vidéo

Les utilisateurs d'Internet regardent et partage des vidéos sur les réseaux sociaux ou les plateformes de téléchargement comme You Tube, leurs besoins ont généré une énorme quantité de données. Ainsi, l'application avancée de traitement de vidéo en profondeur est encouragée par la présence d'un grand nombre de vidéos étiquetés.

La classification vidéo se concentre sur un apprentissage automatique supervisé des frames en fonction du contenu et des séquences vidéo, tandis que le « *sous-titrage vidéo* » consiste à générer de courtes descriptions pour les vidéos et à dégager des informations dynamiques telles que les actions humaines et les trajectoires des voitures. Dans la classification vidéo, les chercheurs proposent des méthodes en trois dimensions : le regroupement de caractéristiques temporelles (TFP) **[14]** et le réseau de convolution 3D (C3D) **[15]**, tandis que le modèle séquence à séquence est appliqué au sous-titrage vidéo.

Par conséquent, la classification des vidéos est un grand défi qui a reçu beaucoup d'attention dans la communauté des chercheurs. De nos jours, l'approche du l'intelligence artificiel et le réseau neuronal prend une place de choix dans le traitement de cette tâche.

L'Intelligence Artificielle (IA) est un vaste domaine, où nous essayons d'imiter le comportement humain dans le but de rendre les machines si puissantes pour accomplir de nombreux types de tâches telles que la résolution de problèmes, la représentation des connaissances, la reconnaissance vocale, et autres. L'idée de base est de mettre les connaissances dans la machine. Grâce à ces deux domaines (IA, DL). Il existe des systèmes sophistiqués capable de changer leur comportement sans qu'il soit nécessaire d'apporter des modifications à leur code, mais uniquement sur leurs données d'entraînement. Ainsi, avec cette vague de techniques avancées d'apprentissage machine qui donne à l'IA un pas en avant, où est la place de DL ou qu'est-ce que le DL apporte dans ce domaine.

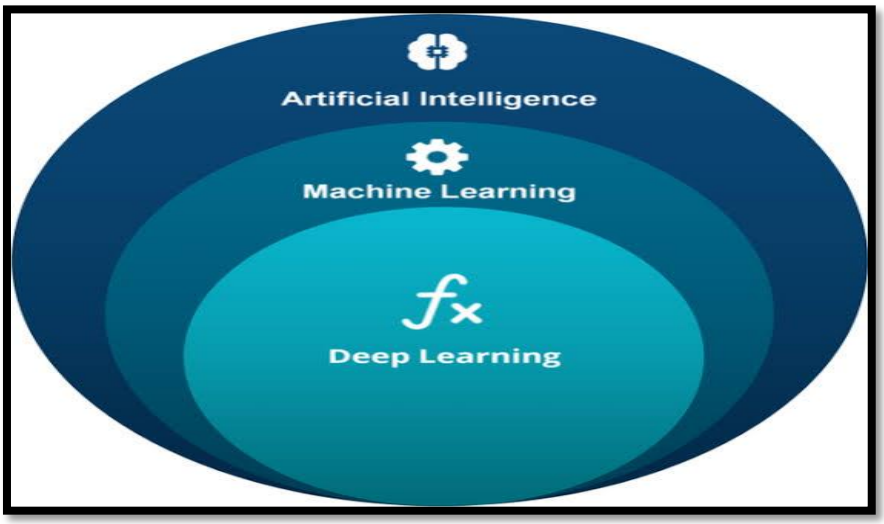

FIGURE 1.8- Affiliation de Deep Learning (relation entre IA et ML et Deep Learning). **[51]**

En d'autres termes, DL est un sous-ensemble des méthodologies et techniques de ML qui utilisent le réseau neuronal artificiel (ANN). C'est l'adaptation des réseaux neuronaux qui imite la structure du cerveau humain. La force de DL réside dans le fait que la machine peut extraire des caractéristiques et apprendre toute seule, indépendamment de l'intervention d'un expert. Il a été appliqué dans de nombreux domaines différents (traitement des images, textes, paroles et vidéos). Le succès de DL appartient à la disponibilité de plus de données d'entraînement. Google, Facebook et Amazon a déjà commencé à l'utiliser pour faire l'analyse de leurs énormes quantités de données. Dans ce qui suit, nous détaillerons certaines approches qui ont été proposées pour ce domaine.

2.1. L'apprentissage automatique (Machine Learning)

D'après L. Samuel (IBM, 1959) : « L'apprentissage automatique (machine Learning) est le champ de l'IA qui permet à une machine (au sens large) d'apprendre. C'est-à-dire, d'améliorer progressivement ses performances sur une tâche spécifique en se basant sur des données, le tout sans être explicitement programmé pour résoudre cette tâche. » **[47]**. En particulier, l'apprentissage automatique est le concept selon lequel une machine explore la construction et l'étude d'algorithmes qui peuvent apprendre et faire des prédictions sur différents ensembles de données. En fonction du résultat de l'apprentissage automatique, nous pouvons distinguer différentes tâches d'apprentissage automatique telles que : classification, clustering, régression, etc.

2.1.1. Technique d'apprentissage automatique

Les techniques d'apprentissage automatique sont appliquées dans un large éventail d'applications afin de résoudre un certain nombre de problèmes fascinants. Parmi les techniques, nous pouvons cités les plus populaires :

2.1.1.1.K plus proche voisins (K-PPV)

L'algorithme *K plus proche voisins (K-PPV)* est un algorithme supervisé non paramétrique pour la classification **[16]**. Cet algorithme est largement utilisé pour plusieurs tâches d'intelligence en classification, segmentation d'images et vidéo. Le principe de l'algorithme est de regrouper les échantillons selon leur voisinage, c'est-à-dire que chaque point est affecté à la classe la plus représentée parmi ses k voisins les plus proches. Cet algorithme est basé sur les deux éléments principaux suivants :

- (k) : le nombre de cas les plus proches à utiliser
- Une *métrique* pour mesurer le plus proche voisin(v).

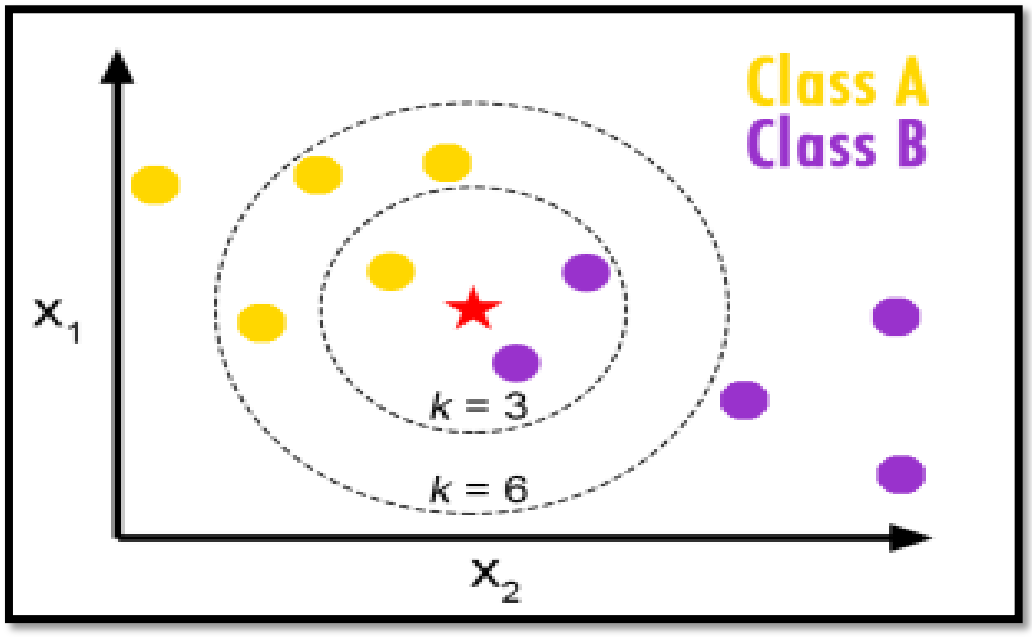

FIGURE 1.9-La tâche de la classification en utilisant l'algorithme K plus proche voisins (K-PPV). *Pour k = 3 la classe majoritaire du point central est la classe B et pour k = 6 la classe majoritaire devient la classe A.* **[52]**

2.1.1.2.Support Vector Machine(SVM)

La technique « *Support Vector Machine (SVM)* » est développé par Vapnik en 1995 **[17]**, est l'une des méthodes les plus populaires dans la famille des approches supervisé. La technique est destinée à résoudre des problèmes de classification ou de régression. Son but est de sépare les données d'apprentissage et maximiser la distance entre deux classes à partir d'un classificateur **[18]**. Cette technique est appliquée dans divers domaines, à savoir la reconnaissance des caractères, la reconnaissance d'images et le diagnostic médical, etc.

2.1.1.3.Decision Tree (Les arbres de décision)

Les arbres de décision sont des modèles de ML supervisés, pouvant être utilisés pour la classification que pour la régression. Leur principe repose sur la construction d'un arbre de taille limitée. La racine constitue le point de départ de l'arbre et représente l'ensemble des données d'apprentissage. Les ensembles de données sont segmentées en plusieurs possibles groupes qui situées aux extrémités des branches (les « feuilles » de l'arbre), en fonction d'une variable discriminante (un des attributs).

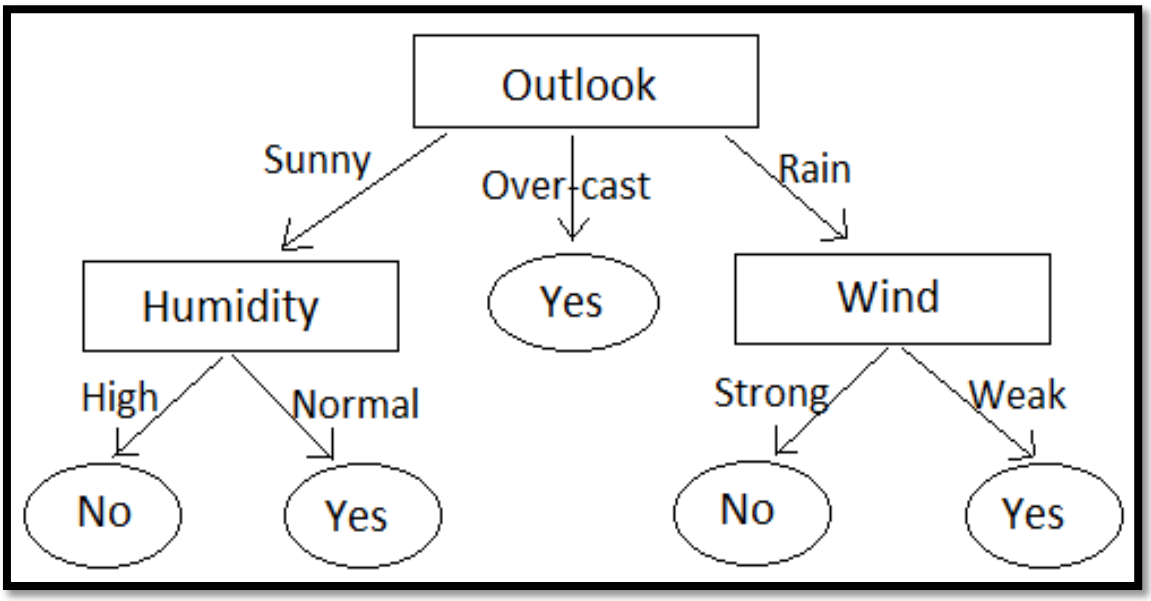

FIGURE 1.10- La classification en utilisant les arbres de décision **[53]**

L'arbre de décision est un outil utilisé dans des domaines variés tels que la sécurité, la fouille de données, la médecine, etc. **[16]**

2.2. La classification

La classification consiste à déterminer une catégorie des nouvelles observations à partir des données existent (training data). La classification est appliquée dans plusieurs domaines tel que : le diagnostic médical, le marketing ciblé, etc. cela nécessite plus de détails, dans autre terme, La classification est une méthode mathématique d'analyse de données, il est appliqué sur des données numériques (points, tableaux, images, sons, vidéo etc.), pour faciliter l'étude d'une population d'effectif important

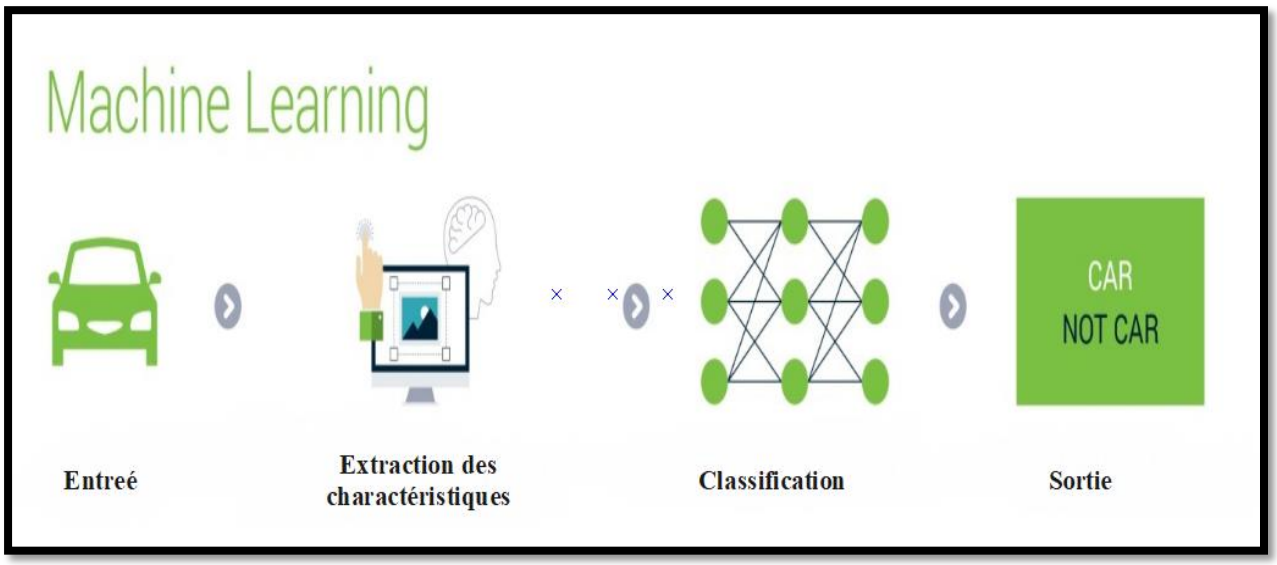

FIGURE 1.11- Le principe de la classification. **[54]**

Comme son nom l'indique, la classification est obtenue en classant les objets dans des sous catégories à partir d'un groupe général, mais la machine peut-elle distinguer les personnes ou les animaux à l'aide des images, ou peut-elle distinguer et classer les textes en fonction de leur contenu

Oui, la machine peut effectuer toutes les classifications à l'aide des données et les algorithmes appropriés, et donner des résultats précis en peu de temps.

Mais dans l'apprentissage automatique, la classification donner un problème d'identification une nouvelle observation à quel groupe de catégories (sous-populations) appartient, sur la base d'un ensemble de données de formation contenant des observations et dont les catégories sont connues

2.2.1. Les types d'apprentissages

En effet, le problème de la classification peut être résolu en fonction de l'approche suivante :

- **L'apprentissage supervisé**

  Est une approche dans laquelle le programme informatique apprend les données qui lui sont transmises par l'utilisateur, puis utilise cet apprentissage pour classer une nouvelle observation. Cet ensemble de données peut simplement être bi-classe (comme identifier si la personne est un homme ou une femme ou si le courrier est du spam ou non) ou il peut être multi-classe (reconnaissance vocale, reconnaissance de l'écriture manuscrite, identification biométrique, classification des documents, etc.).

- **L'apprentissage non supervisé**

  Est une technique d'apprentissage, ou le système ne dispose que d'exemples, et que le nombre de classes et leur nature n'ont pas été prédéterminés. L'objectif de l'apprentissage non supervisé ou clustring est de découvrir des modèles cachés dans un ensemble de données pour classer les données brutes. L'apprentissage non supervisé est utiliser pour a détection d'anomalies, y compris pour les achats frauduleux de cartes de crédit et les systèmes de recommandation qui conseille sur les produits à acheter ensuite

- **L'apprentissage semi supervisée**

  Est une approche se situe ainsi entre l'apprentissage supervisé (utilise des données étiquetées) et l'apprentissage non-supervisé (utilise des données non-étiquetées) il permet d'améliorer significativement la qualité de l'apprentissage avec la combinaison entre les données (étiqueté et non étiqueté)

- **L'apprentissage par renforcement**

Consiste à apprendre par interaction avec l'environnement dans lequel il doit atteindre un certain but à suivre dans une situation donnée par exemple conduire un véhicule ou connais le chemin dans Labyrinthe et, en observant le résultat de certaines actions avec feedback sous forme de « récompenses » et de « punitions » pendant qu'il navigue dans l'espace du problème. Apprend à identifier le comportement le plus efficace dans le contexte considéré. Ceci est appelé le signal de renforcement**.**

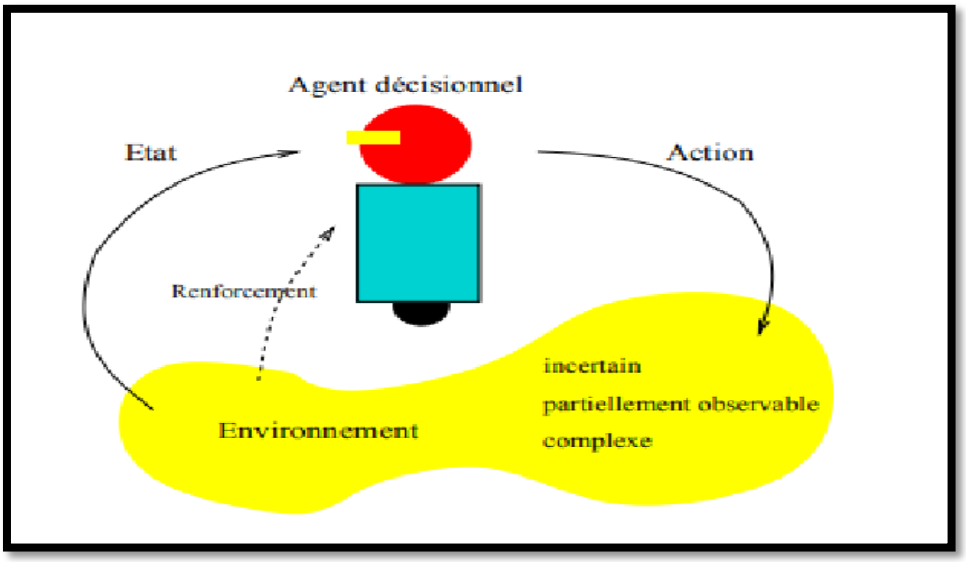

FIGURE 1.12- Exemple d'un agent utilise l'apprentissage par renforcement.

- **L'apprentissage en Profondeur (DeepLeanring) :**

Est un type d'intelligence artificielle et fait partie d'une famille plus large de méthodes d'apprentissage automatique (Machine Learning) ; basées sur l'apprentissage des représentations de données, où la machine est capable d'apprendre par elle-même.

Deep Learning découvre la structure complexe dans de grands ensembles de données en utilisant l'algorithme de rétropropagmentation pour expliquer comment une machine doit modifier ses paramètres internes utilisés pour calculer la représentation de chaque couche à partir de la représentation de la couche précédente. En va détailler dans le point suivant ce type d'apprentissage.

2.2.2. Type de classification

- **Classification binaire**

Les données entrantes sont classées dans l'une des deux catégories possibles. Par exemple pour le diagnostic du malade de personne on a deux classes : malade {oui ou non}.

- **Classification multi- classe**

Les données sont divisées en plusieurs catégories possibles (supérieurs à 2). Ce type est très utile pour la grande masse de données. Exemple : la détection d'émotions selon le visage {heureux, fatigué, triste, etc.}

2.3. Apprentissage profond

L'apprentissage profond est un type d'apprentissage automatique qui vise à former et à enseigner à l'ordinateur à exécuter des fonctions humaines telles que la distinction d'objets visuels et l'identification du son et de l'image. Au lieu d'organiser les données, le deep Learning définit ses propres paramètres de base qui permettent à la machine d'apprendre de manière indépendante, et l'intérêt actuel pour le deep Learning est en partie dû à l'enthousiasme pour l'intelligence artificielle. Les techniques d'apprentissage approfondi ont amélioré la capacité de classer, de reconnaître, etc.

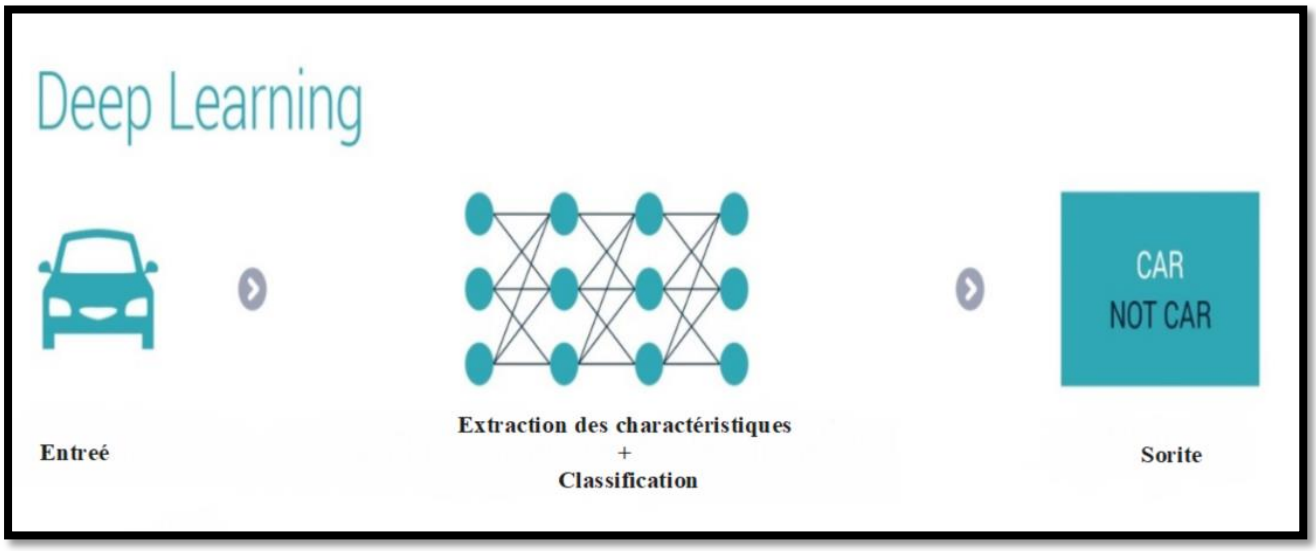

FIGURE 1.13-La classification dans l'apprentissage profond **[54]**

Le deep Learning est une sous-catégorie du machine Learning, il concerne l'utilisation des réseaux de neurones pour améliorer des choses telles que la reconnaissance vocale, la vision par ordinateur et le traitement du langage naturel (tous des domaines particulièrement complexes pour les chercheurs en IA

2.3.1. Historique Machine Learning et Deep Learning : ce tableau résume l'histoire

TABLEAU 1.4.Historique deep Learning

L'année	L'inventeur	Contribution
1873	Alexander Bain	▪ Introduction du Neural Groupings comme les premiers modèles de réseaux de neurones
1943	Warren McCulloch Walter Pitts	▪ écrit un article sur le fonctionnement des neurones. Ils ont modélisé un réseau neuronal simple avec des circuits électriques
1949	Donald Hebb	▪ introduit la règle d'apprentissage de Hebb qui servira de fondation pour les réseaux de neurones modernes
1950	Alan Turing	▪ crée le mondialement connu (Turing Test). ▪ est un test en intelligence artificielle (IA)pour déterminer si un ordinateur est capable de penser comme un être humain
1952	Arthur samuel	▪ Samuel créé vit le premier programme informatique qui pouvait apprendre en cours d'exécution. C'était une machine qui jouait aux dames
1958	Frank rosenblatt	▪ conçu le premier réseau de neurones artificiels, appelé Perceptron,. L'objectif principal était la reconnaissance des formes
1959	Bernard Windrow Marcian Hoffcréé	▪ Introduire deux modèle de réseaux neurone : ▪ **ADELINE**, il pourrait détecter des modèles binaires. Par exemple, dans un flux de bits, il pourrait prédire ce que serait le suivant ▪ **MADELINE**, elle pourrait éliminer l'écho sur les lignes téléphonique
1965	Alexy,Grioryevich Ivakhnenko	▪ Les premiers réseaux neurone en profondeur ▪ Travailler sur des algorithmes d'apprentissage en profondeur et mis au point une méthode de traitement de données.
1974	Paul Werbos	▪ Introduction de la rétro propagation.

1979-1980	Fukushima	▪ crée un réseau neuronales Neocognitron, capable d'apprendre d'identifier des modèles visuels et d'être utilisé dans la reconnaissance de caractères manuscrits et d'autres modèles.
1982	John Hopfield	▪ Introduction des réseaux de Hopfield(des réseaux qui avait des lignes bidirectionnelles)
1985	Hilton et Sejnowski	▪ Introduction des machines de Boltzmann
1985	Terry Segnovsky	▪ crée le programme « NETtalk » qui permet la prononciation des mots anglais et parvient à s'améliorer avec le temps
1986	Michael I. Jordan	▪ Définition et introduction des réseaux de neurones récurrents
1990	Yann LeCun	▪ Introduction de LeNet et montra la capacité des réseaux de neurones profond ▪ identifié la première présentation pratique de la dépression inverse, combinant neurotransmetteurs et propagation des chiffres « manuscrits ».
1997	Schuster et Paliwal	▪ Introduction des réseaux de neurones récurrents bidirectionnels.
1997	Hochreiteret Schmidhuber	▪ Introduction de LSTM, qui a résolu le problème du vanishing gradient dans les réseaux de neurones récurrent.
2006	Geoffrey Hinton	▪ Introduction des Deep Belief Network
2009	Fei Fei Lee	▪ Lancer l'image Net
2012	Geoffrey Hinton	▪ Introduction de AlexNet qui remporte le challenge ImageNet par une large marge
2012	Jeff Dean	▪ Crée (Google Brain) ▪ Un réseau neuronal profond, qui se concentrait sur la détection de motifs dans les images et les vidéos.et utilisé plus tard pour détecter des objets dans des vidéos YouTube
2014	Facebook	▪ Crée un réseau neurone profond (Deep Face) peut connaître les gens avec même précision qu'un être humain

2.3.2. Fonctionnement d'apprentissage profond

Le deep Learning est basé sur un réseau neuronal artificiel inspiré du cerveau humain. Ce réseau est composé de plusieurs «couches» de neurones, chacune recevant et interprétant des informations de la couche précédente. Par exemple, le système apprendra à déterminer s'il y a un visage sur une photo avant de découvrir quelle personne.

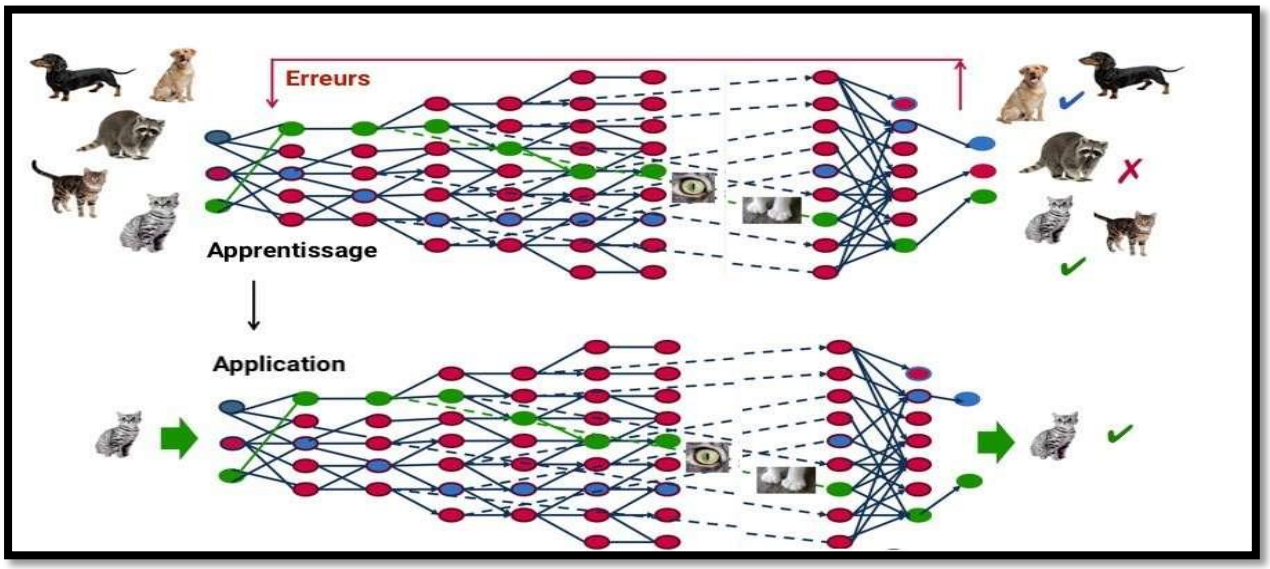

FIGURE 1.14- exemple de fonctionnement de Deep Learning **[55]**

Le Deep Learning découvre une structure complexe dans de grands ensembles de données en utilisant l'algorithme de propagation inverse pour indiquer comment une machine doit modifier ses paramètres internes qui sont utilisés pour calculer la représentation dans chaque couche à partir de la représentation dans la couche précédente.

À chaque étape, la « mauvaise» réponse est éliminée et renvoyée au niveau amont pour ajuster le modèle mathématique. Au fil du temps, le programme a réorganisé l'information en éléments plus complexes. Lorsque ce modèle est ensuite appliqué à d'autres situations, il est généralement capable de reconnaître un chat sans qu'on lui dise qu'il n'a jamais appris le concept de chat. Les données de base sont essentielles : plus le système possède d'expérience, plus il sera efficace.

La base de l'apprentissage profond est la représentation distribuée dans l'apprentissage automatique. "Distribué" signifie l'hypothèse que l'observation est le résultat d'une interaction entre différents facteurs. L'apprentissage en profondeur nécessite également qu'une telle interaction puisse être divisée en plusieurs couches, ce qui signifie l'abstraction multiple de la valeur observée.

2.3.3. Application d'apprentissage profond

L'apprentissage profond est utilisé dans de nombreux domaines :

- La reconnaissance d'image
- La traduction automatique
- La conduite autonome
- Le diagnostic médical
- La recommandation personnalisée
- La modération automatique des réseaux sociaux
- La prédiction financière et trading automatisé
- L'identification de pièces défectueuses
- La détection de malwares ou de fraudes
- *Le chatbots* (agents conversationnels)
- L'exploration spatiale
- Les robots intelligents. [**36**]

2.3.4. Architectures d'apprentissage profond

Il existe plusieurs cadres d'apprentissage en profondeur déjà largement utilisés, tels que le réseau neuronal profond, le réseau neuronal convolutionnel (CNN) et le réseau neuronal récursif (RNN), et le LSTM.

2.3.4.1.Les réseaux Long Short Memory (LSTM) :

Les réseaux de mémoire à court terme sont un type de réseau neuronal récurrent (RNN) introduit par Hochreiter et Schmidhuber en 1997, une technique capable d'apprendre et de se souvenir sur de longues séquences de données d'entrée.

Les réseaux LSTM contiennent leurs informations dans une mémoire (qui est très similaire à la mémoire d'un ordinateur). Cette mémoire peut être considérée comme une cellule fermée, où fermée signifie que la cellule décide de stocker ou de supprimer des informations de sa mémoire.

Dans un LSTM vous avez trois portes :*

- **Porte d'entrée (input gate)** : pour rôle d'extraire l'information de la donnée courante
- **Porte d'oubli (forget gate):** Cette porte décide de quelle information doit être conservée ou jetée
- **Porte de sortie (output gate) :** décider de quel sera le prochain état caché, qui contient des informations sur les entrées précédentes du réseau et sert aux prédictions.

Les portes d'un LSTM sont analogiques, modélisées par une fonction qui est généralement un

sigmoïde. Ce sigmoïde (signifie qu'ils vont de 0 à 1) est appliqué à la somme pondérée des entrées (en bleu), des sorties (en vert) et de la cellule (en orange), par des poids. Vous pouvez voir une illustration d'un LSTM avec ses trois portes ci-dessous:

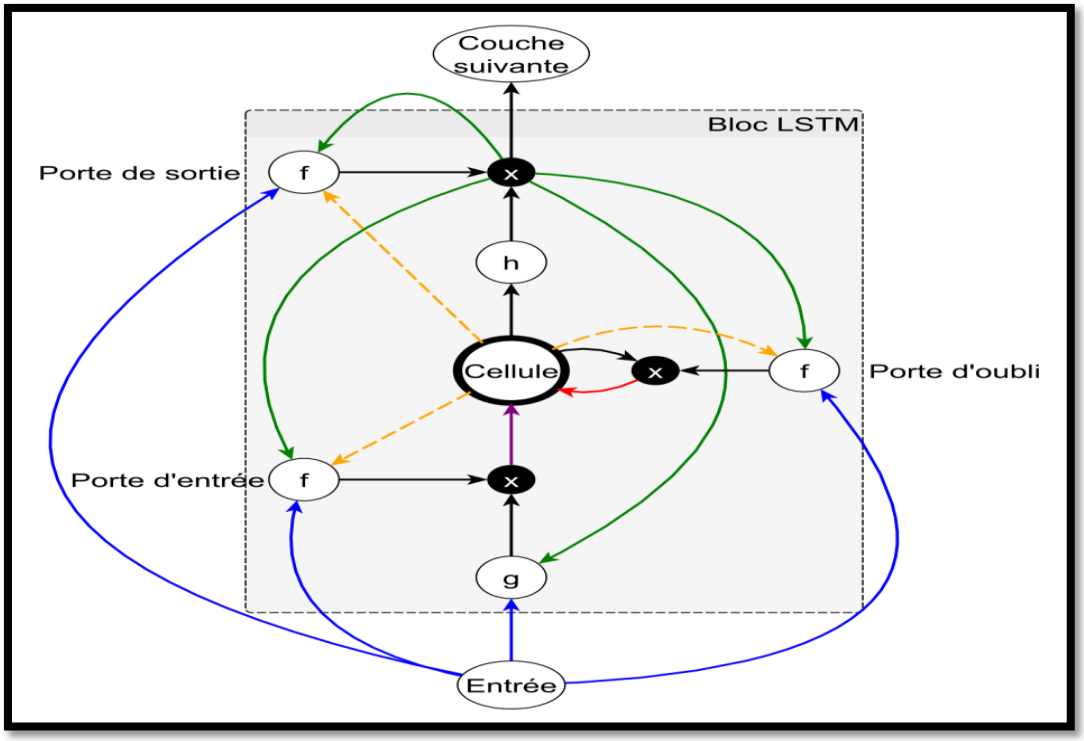

FIGURE 1.15-Fonctionnement d'un réseau LSTM **[56]**

2.3.4.2.Les réseaux convolutifs (CNN)

Un CNN (réseau neuronal convolutionnel) est très similaire aux réseaux neuronaux de base. Il est également composé de neurones avec des poids et des biais.

Il se compose de plusieurs couches permettant à notre système de reconnaître des objets sur une image:

- **La couche convolution** : qui tire d'une image ses caractéristiques principales
- **La couche ReLU** : qui créer des modèles non linéaires et augmente la vitesse de formations du modèle.
- **La couche Pooling** : qui permet de réduire la taille de l'image et donc le nombre d'informations pour ne garder que le plus important
- **La couche Entièrement connectée *(fully-connected)*** : permet de classifier l'image en entrée d'un réseau (qui représente sous forme d'un vecteur), pour indiquer la probabilité de chaque neurone dans notre image.

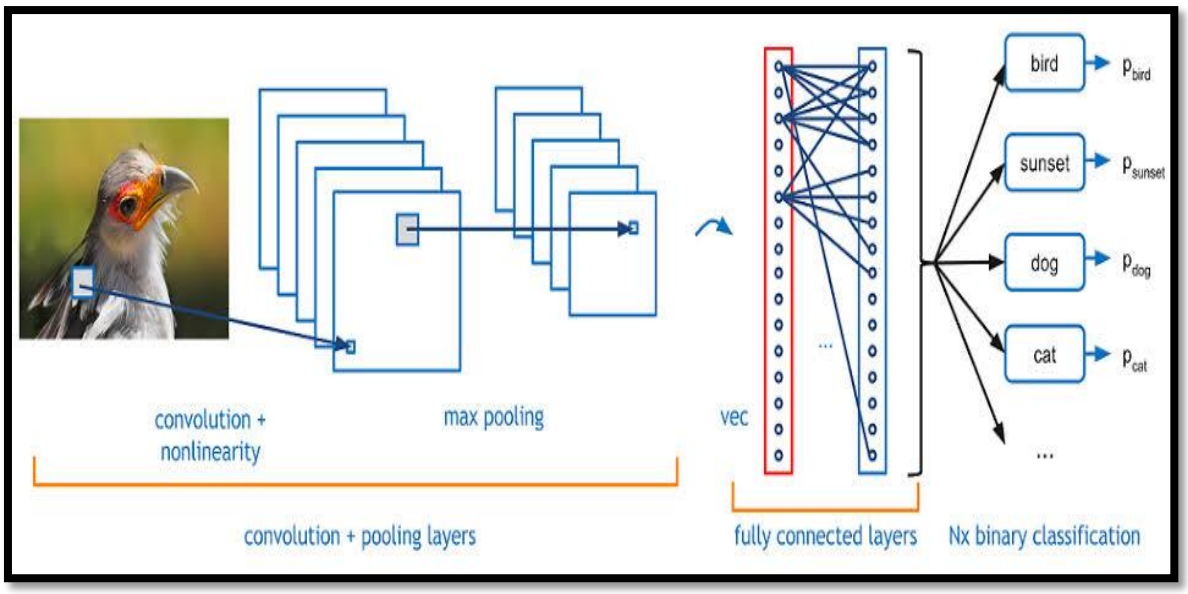

FIGURE 1.16-Fonctionnement d'un réseau convolutif (CNN) **[57]**

3. Problèmes de traitement vidéo :

- L'extraction des objets en mouvement par rapport à l'arrière-plan.
- La Complexité de positionnement et la calibration de l'ensemble des caméras
- La détection peut être sur des images animées très complexes avec plusieurs personnes dans la même image
- Les conditions d'éclairage et d'illumination : Dans toute action de détection, la lumière est un facteur important et c'est le problème le plus délicat à résoudre. C'est à dire Le changement dans la luminosité entre l'ensemble des caméras où chaque caméra est positionnée dans un endroit différent avec des conditions différentes
- Le plus grand problème est assuré un temps de calcul compatible au temps réel (vitesse de traitement)

La vidéo est un problème de classification intéressant car elle comprend à la fois des caractéristiques temporelles et spatiales. C'est-à-dire qu'à chaque image d'une vidéo, l'image elle-même contient des informations importantes (spatiales), tout comme le contexte de cette image par rapport aux images qui la précèdent dans le temps (temporel).

Conclusion

Dans ce chapitre, nous avons dérivé la modalité vidéo de manière générale, tout d'abord, nous présentons la vidéo, le processus d'enregistrement d'une vidéo, les formats vidéo, la compression vidéo et les différents caractéristiques dans ce contexte comme le traitement vidéo applications, etc. Nous avons également parlé de la classification des vidéos dans le l'apprentissage automatique en essayant de clarifier les concepts qui nous intéressent dans ce domaine, y compris la classification et les implications de la relation entre l'apprentissage existant et l'apprentissage profond, ainsi que l'évolution de l'histoire de l'apprentissage profond et de leur fonctionnement. Nous avons également présenté quelques algorithmes ou méthodes de classification (kPPV, Support Vector Machine, Arbre de décision). Enfin, nous avons mentionné quelques architectures qui ont été proposé en ce domaine

Dans le chapitre suivant, nous essayons décrire les architectures d'apprentissage profond de base et plus précisément les architectures de classification des vidéos, puis nous allons consacrer une partie pour une étude comparative entre ces architectures

CHAPITRE 02

LES ARCHITECTURES D'APPRENTISSAGE PROFOND POUR LES TÂCHES DE TRAITEMENT VIDÉO

1. Les modèles d'apprentissages profonds de base.
2. Les architectures de classification vidéo.
3. Étude comparative Entre Les architectures de classification des vidéos

INTRODUCTION

Le contenu numérique est aujourd'hui intrinsèquement multimédia : texte, audio, image, vidéo, etc. En particulier, la vidéo est devenue un nouveau moyen de communication entre les internautes avec la prolifération d'appareils mobiles riches en capteurs. Accélérées par l'augmentation spectaculaire de la bande passante Internet et de l'espace de stockage, les données vidéo ont été générées, publiées et diffusées de manière explosive, devenant un élément indispensable des mégadonnées d'aujourd'hui. Cela a encouragé le développement de techniques avancées pour un large éventail d'applications de compréhension vidéo, y compris la publicité en ligne, la récupération vidéo, la vidéosurveillance, etc. Un problème fondamental qui sous-tend le succès de ces avancées technologiques est la compréhension du contenu vidéo. Progrès récents dans l'apprentissage d'images profondes, les domaines ont motivé des techniques pour apprendre des représentations de fonctionnalités vidéo robustes afin d'exploiter efficacement les indices multimodaux abondants dans les données vidéo.

Dans ce chapitre, nous passons en revue deux tâches d'apprentissage automatique visant à stimuler la compréhension des architectures d'apprentissage profond des vidéos : la classification vidéo et la détection des objets dans la vidéo. Alors que la classification vidéo se concentre sur l'étiquetage automatique des séquences vidéo en fonction de leur contenu fixé par l'utilisateur, comme les actions dans une vidéo de football ou les événements de trafic. La détection tente à identifier les objets, les actions, les événements contenant dans un vidéo, enrichissant le libellé unique de la classification vidéo pour capturer la dynamique la plus informative des vidéos.

1. Les modèles d'apprentissages profonds de base

Dans cette section, nous présentons brièvement les modèles d'apprentissage profond de base qui ont été largement adoptés dans la littérature pour la classification et la détection vidéo

1.1. Les réseaux récurrents (RNN)

Les RNN sont utilisés dans l'apprentissage profond et dans le développement de modèles qui simulent l'activité des neurones dans le cerveau humain. Ce réseau résoudre de nombreux problèmes réels auxquels les industries sont confrontées comme : la Prédiction du mot suivant/ Composition musicale/ Sous-titrage d'image/ Reconnaissance de la parole/ Détection d'anomalies de séries chronologiques/ Prédiction boursière/ Marquage vidéo, etc.

Les RNN sont également connu sous le nom d'Auto Associative ou Feedback Network (AAFN), ce dernier est un type de réseau neuronal qui utilise des boucles de rétroaction pour traiter une séquence de données ; il utilise leur raisonnement des expériences précédentes pour informer les événements à venir (dans la sortie finale), qui peut également être une séquence de données. Ces boucles de rétroaction permettent aux informations de persister avec les connexions entre eux parce qu'ils forment un cycle dirigé. Cela crée un état interne du réseau qui lui permet de présenter un comportement temporel dynamique.

La figure 2.1 ci-dessous, montré que dans RNN, toutes les entrées sont liées les unes aux autres. Supposons qu'un objectif de prédire le mot suivant dans une phrase donnée, dans ce cas, la relation entre tous les mots précédents aide à prévoir la meilleure sortie. Le RNN se souvient de toutes ces relations tout en s'entraînant. Pour y parvenir, le RNN crée des réseaux avec des boucles, ce qui lui permet de conserver les informations.

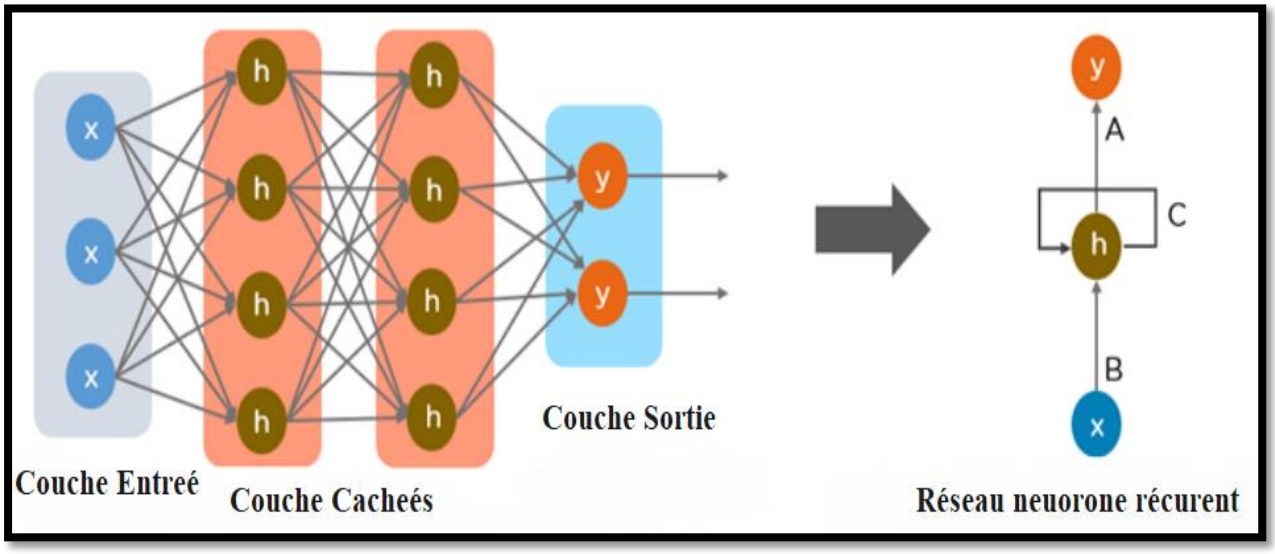

FIGURE 2.1- Réseaux neurone recurrent **[58]**

Dans la phase d'entrainement, le modèle prend l'input x (0) de la séquence d'entrée, puis il sort h (0) qui, avec x (1), est l'entrée pour l'étape suivante. Ainsi, h (0) et x (1) est l'entrée pour l'étape suivante. De même, h (1) de la suivante est l'entrée avec x (2) pour l'étape suivante et ainsi de suite. De cette façon, il garde en mémoire le contexte pendant l'entraînement. Les couches de RNN sont :

- **Input layer x (t)**

Est considéré comme l'entrée du réseau au pas de temps t.

- **Hidden layer h (t)**

Représente un état caché à l'instant t et agit comme «mémoire» du réseau. Dans laquelle **h(t)** = *f* (*B* **x(t)** + *C* **h (t−1)**) / f = La fonction f est considérée comme une transformation non linéaire telle que tanh, ReLU.

- **Output y(t)**

Illustre la sortie du réseau. Dans laquelle **y(t)= f (Ah(t))**

- **Poids**

Le RNN a des poids (B, A, C) partagés dans le temps ; input to hidden paramétré par une matrice de poids A, hidden to hidden paramétré par une matrice de poids C, hidden to output paramétré par une matrice de poids B.

1.2. Long Short-Term Memory (Lstm)

Les LSTM est une architecture de réseau neuronal récurrent artificiel (RNN), contrairement aux réseaux neuronaux à action directe standard, le LSTM a des connexions de rétroaction. Elle contient leurs informations dans une mémoire, (ce qui ressemble beaucoup à la mémoire d'un ordinateur). Cette mémoire peut être vue comme une cellule *gated*, où *gated* signifie que la cellule décide de stocker ou de supprimer des informations de sa mémoire. Cette cellule agit comme un transporteur d'information modulée par 3 portes configurables (par entraînement) qui interagissant d'une manière très spéciale. Généralement, on peut distinguer les gâtes suivantes :

- **Port Oubli (forgate gate)**

Pour optimiser les performances du réseau; la porte de l'oubli sera chargée de filtrer les informations contenues dans la cellule mémoire précédente, par la conservation que les éléments pertinents. Pour décider quelles valeurs vont être autorisées à passer la fonction sigmoïde **[19]** elle a le droit à écraser l'entrée entre [0,1]. Il est représenté comme suit :

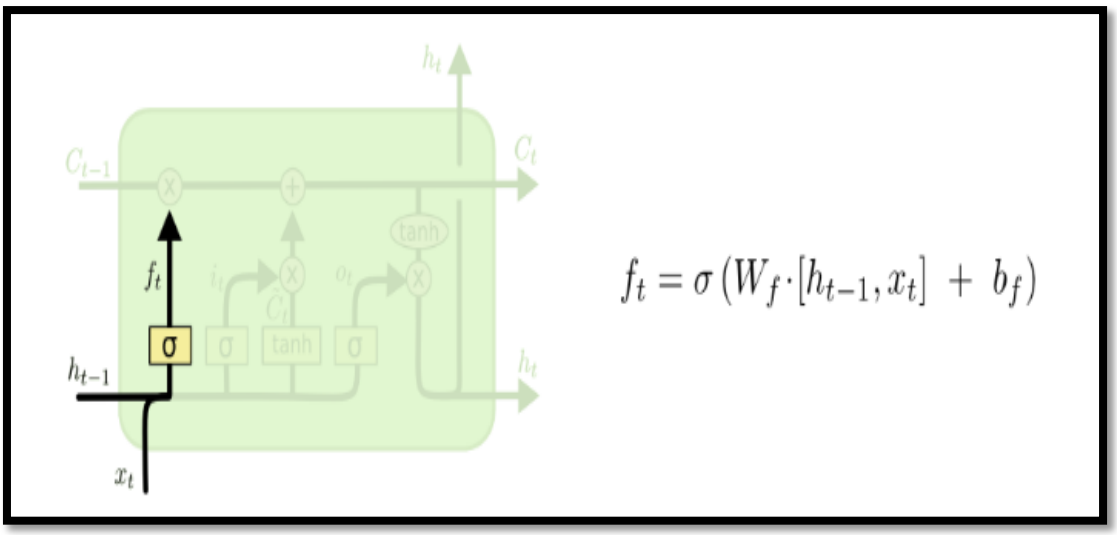

FIGURE 2.2-Porte oubli **[59]**

- **Port Entrée (input gate)**

La porte d'entrée responsable de l'ajout la quantité des informations que nous souhaitons les ajouter à notre état de cellule actuel ; pour cela la couche *tanh* crée un vecteur pour les nouveaux candidats à ajouter à l'état et la couche *sigmoïde* décide des valeurs à mettre à jour.

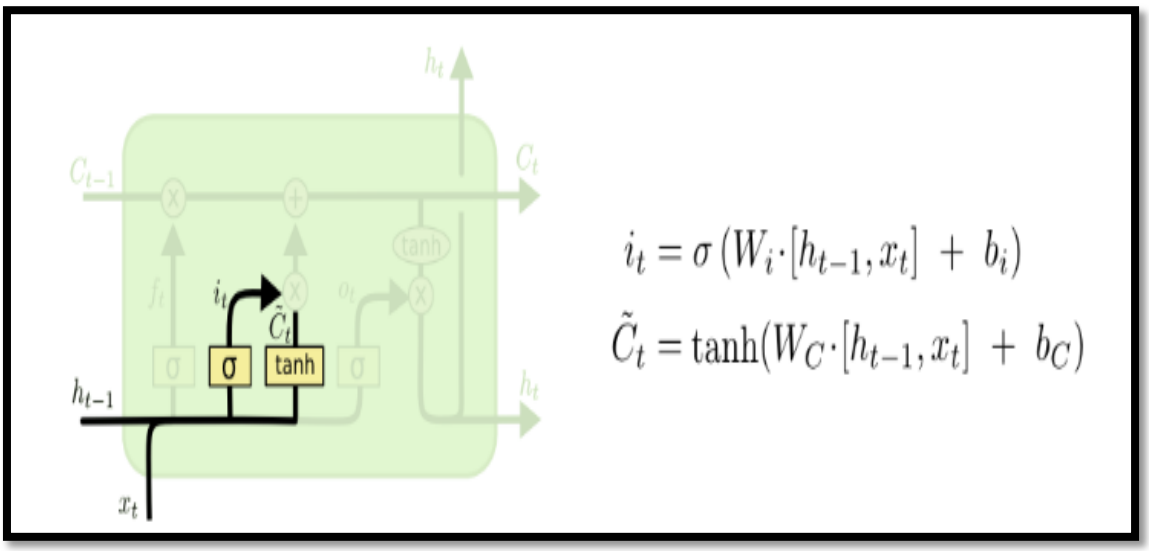

FIGURE 2.3-Porte Entrée **[59]**

- **Port sortie (output gate)** :

La port sortie « output » sélectionne et sort les informations nécessaires, de notre état de cellule qui sera effectuée par la fonction *sigmoïde* et *tanh*.

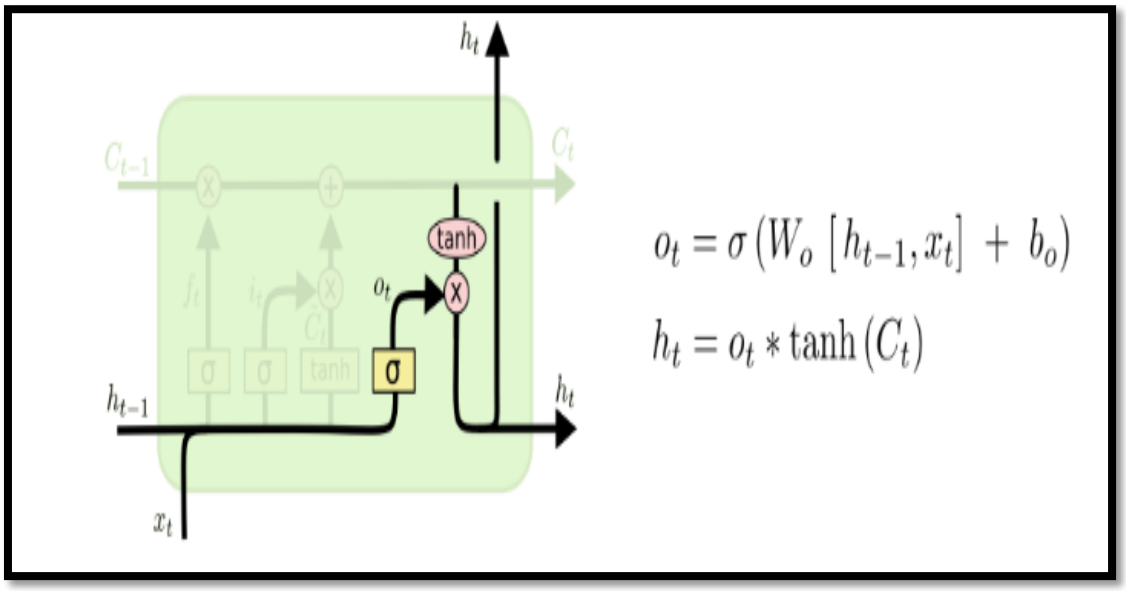

FIGURE 2.4- Porte Sortie **[59]**

- **xt** : vecteur d'entrée dans l'unité LSTM.
- **ft** : vecteur d'activation de la porte d'oubli.
- **It** : vecteur d'activation des portes d'entrée.
- **t** : vecteur d'activation des portes de sortie.
- **ht** : vecteur de sortie de l'unité LSTM.
- **ct** : vecteur d'état de cellule.
- **W, U, b**: les matrices des poids et des bais

Les réseaux LSTM est une topologie neuronale extrêmement utile et il appliqués à une variété de tâches d'apprentissage en profondeur comme la prédiction de texte et la prédiction de stock **[20]**. Ce réseau peut non seulement traiter une seule modalité de données telles que les images, mais également une multitude de données telles que le son et la vidéo. Par exemple, LSTM est applicable à des tâches telles que la reconnaissance des objets dans une image, la reconnaissance manuscrite, la reconnaissance vocale et la détection d'anomalies dans les réseaux ou les IDS (systèmes de détection d'intrusion).

1.3. Les réseaux de neurones convolutifs (CNN, Convnet)

Est une classe de réseaux de neurone artificiel profond à action directe (non récurrente) qui est appliquée à l'analyse de l'imagerie visuelle. Les CNN sont des Réseau de neurones avec une opération de convolution contient 4 couches, laquelle :

- **La couche de convolution (Convolutional layer)**

C'est l'élément clé des réseaux de CNN. Son but est d'identifier un ensemble de caractéristiques (*features*) d'une image reçue en entrée. Pour cela, nous effectuons un filtrage par convolution : dans le principe est de « *faire glisser* » la fenêtre représentant la feature (le filtre) sur l'image et de calculer le produit de convolution entre le filtre et chaque partie de l'image balayée. Donc cette couche reçoit plusieurs images en entrée et utilise chaque filtre pour calculer la convolution de chaque image. Les filtres correspondent exactement aux caractéristiques que nous voulons trouver dans l'image.

Nous obtenons une carte d'activation (feature map) pour chaque paire (image, filtre), qui nous indique où la feature se trouve dans l'image : plus la valeur est élevée, plus il y a de positions dans l'image qui correspondent au filtre.

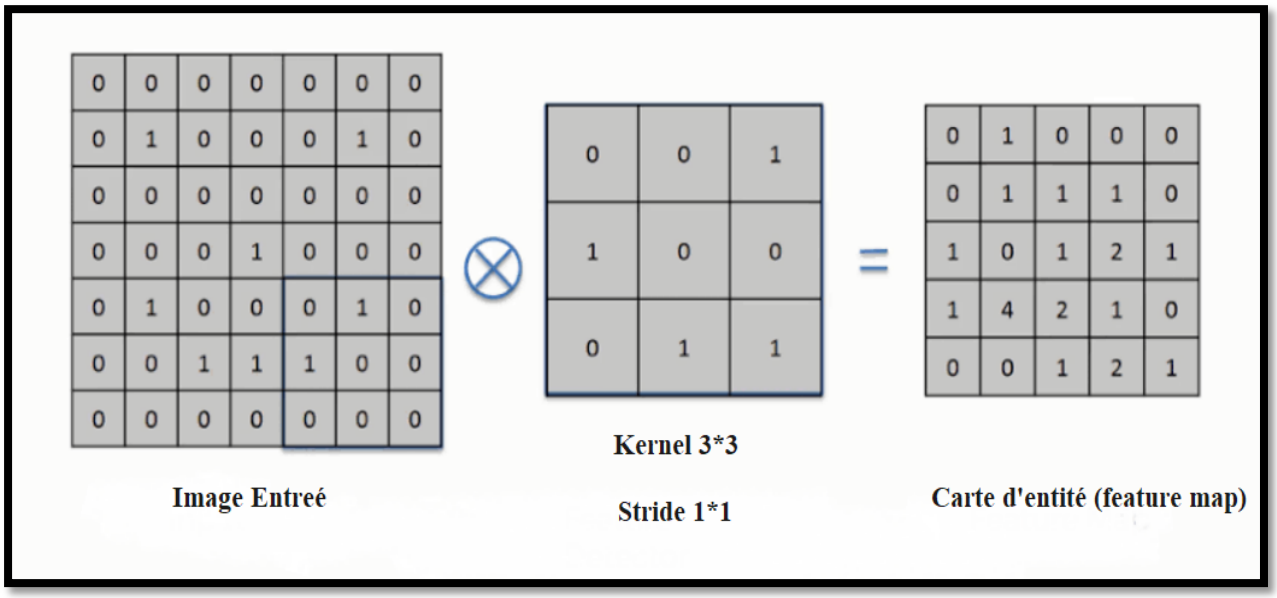

FIGURE 2.5- L'opération de convolution **[63]**

- **La couche de pooling (pooling layer)**

Il reçoit plusieurs *Feature Map* en entrée et applique une opération de pooling (subsampling) à chaque feature map. Elle permet la réduction de la taille de l'image en tenir

en compte ses caractéristiques importantes. Le principe est de couper l'image en cellules régulières, puis nous conservons la valeur maximale dans chaque cellule par rapport le filtre

utilisé (2 x 2 pixels, 3x3, etc.). Donc des petites cellules carrées sont souvent utilisées afin de ne pas perdre trop d'informations.

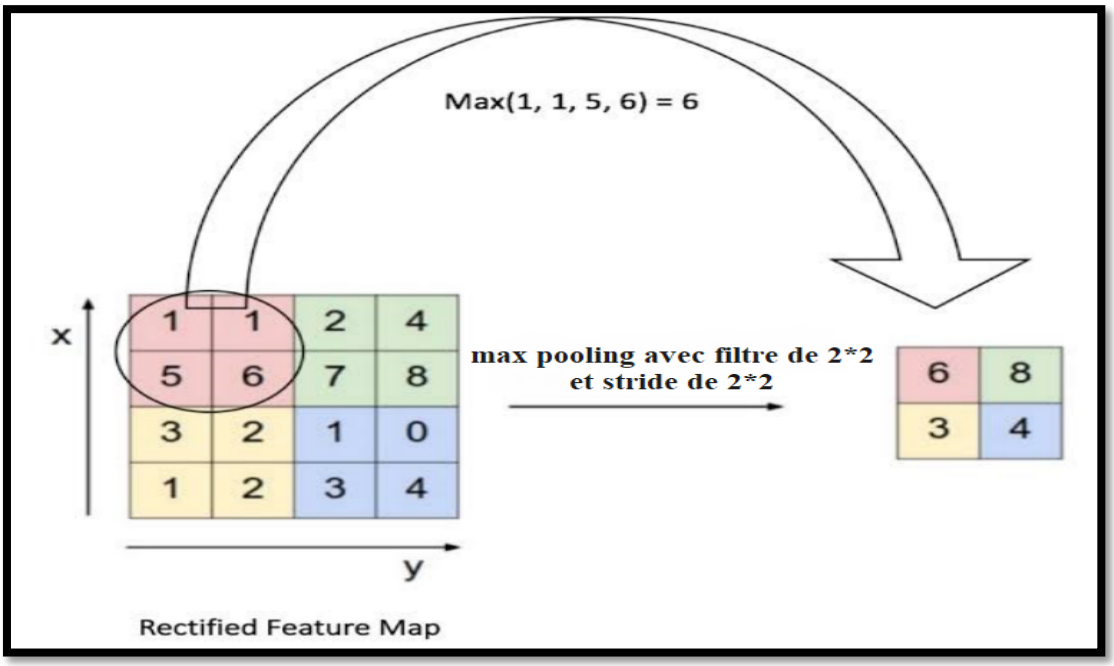

FIGURE 2.6-. La couche de *pooling* **[60]**

Nous obtenons le même nombre de features map que l'entrée, mais elles sont beaucoup plus petites. Cela permet d'accélérer non seulement les calculs, mais d'éviter également le problème du sur-ajustement.

Il existe deux principaux types de pooling : max et min. Comme son nom l'indique, pooling maximal est basé sur la détection de la valeur maximale dans la région sélectionnée et le pooling minimal sur la détection de la valeur minimale dans la région sélectionnée.

- **La couche de correction ReLU (*Rectified Linear Units*)**

Représente la fonction réelle non-linéaire permet de remplacer toutes les valeurs négatives reçues en entrées par des zéros. Elle joue le rôle de fonction d'activation.

$$\text{ReLU}(x) = \max(0, x).$$

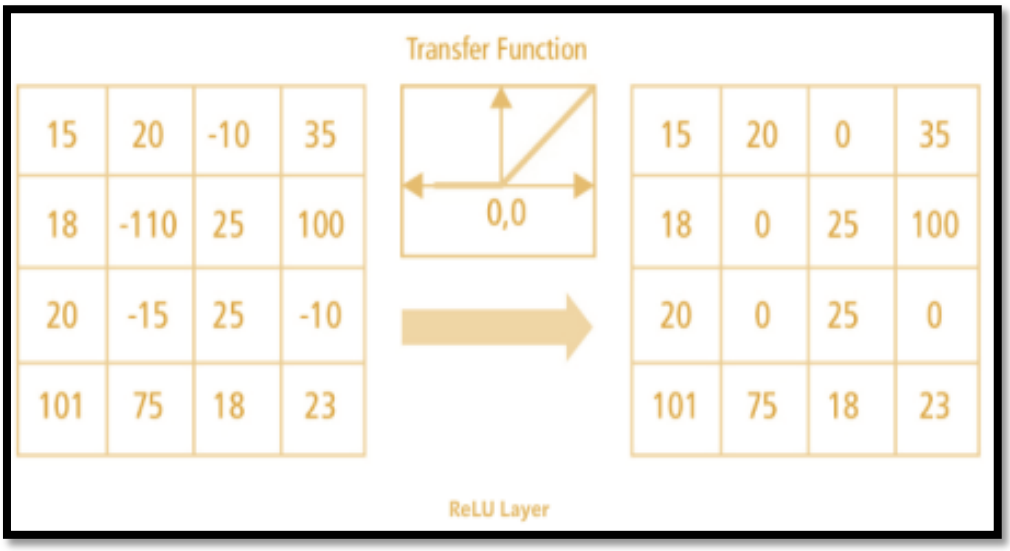

FIGURE 2.7-. La couche de RELU **[61]**

- **La couche entièrement connectée *(fully-connected)***

Elle permet de classifier l'image en entrée du réseau dans une étiquette. Cette couche reçoit un vecteur en entrée (représente le résultat d'aplatir de notre matrice) et produit un nouveau vecteur en sortie. Chaque élément du vecteur représentant une probabilité qu'une certaine caractéristique appartienne à une étiquette.

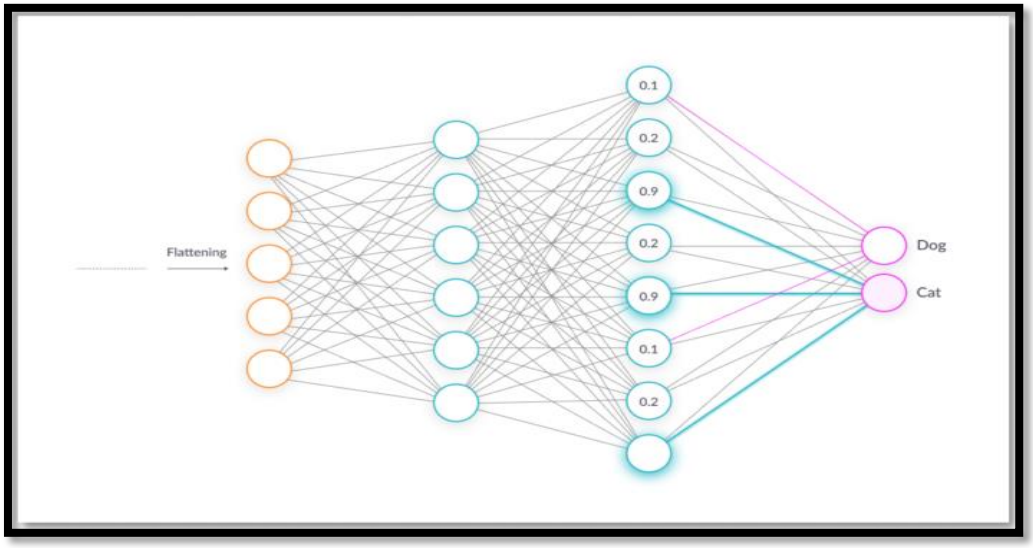

FIGURE 2.8-. La couche entièrement connectée. **[62]**

2. Les architectures de classification des vidéos

Comme nous avons vu dans les sections précédentes, les vidéos peuvent être considérer comme une série séquentielle d'images. Dans le processus de traitement, de nombreuses architectures d'apprentissage profond ont classifié rapidement une vidéo comme en effectuant une classification des images à plusieurs de fois, dans lequel ce nombre est le nombre total d'images dans une vidéo. Les méthodes de classification vidéo conventionnelle contiennent

généralement deux tâches : l'extraction et la classification des fonctionnalités. Ces méthodes sont largement utilisées dans de nombreux domaines d'applications de l'AA, comme la classification vidéo de football, médical, etc. Parmi ces méthodes nous pouvons mentionnées : les SVM, les arbres de décision, etc. Tandis que les méthodes de classification modernes sont des architectures d'apprentissage profond. Ces architectures se caractérisent par ses excellentes performances et contiennent généralement une seule tâche (ex, la classification). Parmi ces méthodes, on peut mentionnées :

2.1. Recurrent Convolutional Neural Network(RCNN)

L'apprentissage profond basée sur RCNN (Recurrent Convolutional Neural Network) été largement adoptés pour la classification vidéo. Cette architecture est fondamentale pour apprendre les relations à partir des Inputs de séquences vidéo [**21**].

Joe et al. [**22**], et Jeffery. [**23**], utilisent RNN avec des unités de mémoire à court terme (LSTM) [11] comme méthode de fusion pour extraire des caractéristiques temporelles sur des représentations spatiales CNN. Zhenqi et al**. [24]**, ont proposés une architecture d'apprentissage profond basée sur RCNN (Recurrent Convolutional Neural Network) pour la classification vidéo. En particulier, l'architecture proposée est basée sur l'extraction des caractéristiques locales et denses des trames d'images ainsi que l'apprentissage des caractéristiques temporelles entre les trames consécutives. De plus, l'opération de convolution et les liens récurrents sont également combinés pour les tâches de classification vidéo. Les caractéristiques locales et denses pouvant être apprendre grâce à une opération de convolution (CNN), la modélisation du mouvement entre les trames d'image par la liaison du pas de temps précédent (RNN), la réduction de la redondance et l'utilisation des stratégies d'échantillonnage aléatoire. L'opération de convolution dégage les avantages suivants :

- Peut gérer des entrées de longueur variable sans couper une vidéo en plusieurs clips
- Peut explorer la fonction de mouvement à la fois au niveau bas et au niveau haut
- Convient aux tâches de vision, car l'image est pertinente localement
- Visualiser les caractéristiques de mouvement tout comme les caractéristiques spatiales.

Xu et al [**24**], soutiennent que RCNN n'est pas non plus un bon moyen de modéliser le mouvement. Parce que le but de l'architecture CNN est d'apprendre les fonctionnalités de bout en bout invariables, en particulier dans la fonctionnalité de haut niveau. Ainsi il sera invariant au mouvement. Plus le modèle CNN est performant, moins il reste de mouvement dans la fonction CNN. Les résultats de l'expérience rapportés dans [**23**] montrent que la caractéristique inférieure (fc6) fonctionne mieux que la dernière caractéristique de couche (fc7), ce qui est cohérent avec

notre analyse. Un autre inconvénient est qu'il est difficile de visualiser la fonction RNN, ce qui est important pour les tâches de vision.

2.2. Les réseaux convolutifs (CNN)

Les réseaux de neurones convolutifs (CNN) **[26]**, ont montrés récemment une classe efficace de modèles profond pour l'extraction des objets des images **[27]**, atteindre un niveau souhaité de reconnaissance **[28]**, de classification **[29]**, de détection et la récupération des informations **[30]**. Ce type de réseau englobe les techniques permettent d'étendre les modèles en sortie de quelques paramètres à des millions paramètres. Ainsi, des jeux d'apprentissage étiquetés massifs permettent de construire le modèle d'apprentissage. Avec ces modalités, les réseaux de neurones convolutifs (CNN) ont montrés une grande puissance interprétable d'extraction des caractéristiques, encouragés également par des résultats positifs dans le domaine d'étude. Watcharin Maungmai et al. **[31]**, ont proposé une architecture CNN contient quatre couches importantes qui est détaillé précédemment. L'objectif de l'architecture est en double : contourner les problèmes qui sont la classification des types de véhicules et la classification des couleurs des véhicules, améliorer la précision du type de véhicule et la classification des couleurs de véhicule. Après avoir coupé la vidéo en plusieurs frames. Les opérations CNN proposées contiennent deux couches de convolution, ensuite une couche de Pooling appliqué avec la fonction d'activation après chaque couche de convolution, enfin ils appliquent trois couches de entièrement connecté **[31]**. Akçay et al. **[32]**, étudions les performances des CNN en classification vidéo à grande échelle, où les réseaux ont accès non seulement aux informations d'apparence présentes dans des images statiques uniques, mais aussi à leur évolution temporelle complexe. Il y a plusieurs défis à l'extension et à l'application des CNN dans ce contexte.

Le modèle proposé par Wei et al. **[33]**, est basée sur l'apprentissage profond CNN avec deux tâches : pour détecter et classer des personnes dans des données vidéo capturées à des distances de plusieurs kilomètres (vidéo en champ lointain) via une caméra vidéo à objectif haute puissance. L'objectif de ce travail est impliqué de surveiller les frontières de loin distance afin de détecter les personnes et d'identifier si elles porter un paquet ou un bras long. Pour un objective de détection, un ensemble d'étapes de traitement d'image efficaces en termes de calcul sont envisagées pour identifier les zones en mouvement dans une image ; suivie par un détecteur Adaboost Person pour voir s'il y a une personne dans les zones en mouvement. Ensuite, ces informations sont transmises à un classifier de réseau de neurones convolutif dont les couches convolutives consistent l'apprentissage par transfert sur GoogleNet. Ce qui rend la distinction entre une personne portant un paquet, une personne portant un bras long et une personne ne portant aucune charge.

Le modèle proposé par Ali et al. [**34]** repose sur deux tâches : la détection et la classification des véhicules dans la vidéo. L'objectif principal de ce travail est de comparer et d'évaluer six architectures CNN et celle qui fonctionne mieux entre autres. Dans l'architecture proposée, l'étape de la détection est repose sur trois parties, la première partie consiste à choisir les régions d'image d'entrée qui passeront à l'étape suivante (Cela s'appelle l'étape de proposition de région) par le SSD, ou faster R-CNN. La deuxième partie consiste à séparer l'image d'entrée en un nombre fixe de grilles par YOLO. La troisième étape consiste à extraire toutes les fonctionnalités nécessaires et obtenir le score de classification pour chacune des régions ou grilles par une certaine architecture CNN. Enfin, l'algorithme "suppression non maximale" est utilisé pour combiner les cases qui se chevauchent dans le résultat dans une seule case pour chaque objet.

Dans la phase de la classification des véhicules, ils mettent en œuvre un certain nombre d'architectures CNN, pour faire la comparaison entre elles. Les différentes techniques d'apprentissage profond contenu dans les Réseaux Neurone Convolutif (CNN) est : ResNet, Inception-ResnetV2, InceptionV3, NASNet, MobileNetV2 et PNASNet.

Une architecture basée sur les réseaux neuronaux convolution (AlexNet CNN) est proposée par *Minhas et al.* **[35]**, pour la classification des tirs pour les vidéos de sports de terrain. Le modèle est robuste aux variations de caméra, au changement de scène, aux vitesses d'action et aux conditions d'éclairage ; c.-à-d. (Lumière du jour, lumière artificielle, ombre). Ils améliorent les performances globales de formation et de validation évaluées sur un ensemble de données divers de vidéos de cricket et de football Grâce à réponse normalisation et les couches *dropout* dans les *Features Map*.

L'architecture proposée dispose d'un réseau à *huit couches* qui se compose de *cinq couches* convolutives et de *trois couches* entièrement connectées pour classer les prises de vue en prises de vue longues, moyennes, rapprochées et hors champ (hors du terrain).

2.3. Les architectures hybrides :

2.3.1. Architecture basé sur les réseaux convolutifs (CNN) et les réseaux Long Short Memory(LSTM) :

Soentanto et al. [**37**], explorent les modèles d'apprentissage profond pour reconnaître les actions humaines dans une vidéo et en classant leurs objets. Ils regroupent les modèles en se basant sur deux tâches, qui sont la classification des objets et la reconnaissance de l'action humaine à partir d'une vidéo. Pour la reconnaissance de l'action humain, l'architecture extraire initialement des informations relatives au mouvement humain ensuite les informations sont représentées par le flux optique (le motif de déplacement d'objets d'une image à une autre image) entre deux trames (image) vidéo et introduites dans les réseaux de neurones convolutifs (CNN).

La classification des actions est effectuée par les architectures de mémoire à court terme (LSTM), et la classification des objets est effectuée uniquement par les réseaux de neurones convolutifs à travers les images vidéo en entrée. D'une part, la reconnaissance d'action humain est effectuée en utilisant quatre couches de convolution et trois couches Pooling, mais avant de classer la représentation d'entrée en utilisant le entièrement connecté. Le modèle de reconnaissance d'action utilise une couche de mémoire à long terme à court terme (LSTM) avec mille unités cachées. D'autre part, pour la classification des objets, le modèle englobe 12 couches de convolution et trois couches entièrement connecté.

2.3.2. Architecture basé sur les réseaux convolutifs (CNN) et les réseaux récurent (RNN)

NG et al. [**38**], proposent une architecture hybride de réseaux de neurones profonds (CNN et RNN) pour combiner des informations d'image à travers une vidéo sur des périodes plus longues (gérer la vidéo complète) que précédemment. D'abord, l'architecture explore diverses aspects temporelles convolutives et de Pooling Feature et examiner les différents choix de conception qui doivent être faits dans l'adaptation de CNN. Ensuite, les réseaux neurone convolutives utilise l'architecture : Conv pooling GoogleNet. Puis, la couche entièrement connecter à la sortie du CNN sous-jacent modélise explicitement la vidéo comme une séquence ordonnée d'images. Cette tâche modélisée par réseau neuronal récurrent qui utilise des cellules de mémoire à court terme à long terme (LSTM). Les avantages de cette architecture peuvent être résumés comme suit :

- Pour obtenir un niveau vidéo global et démontrent que l'utilisation d'un nombre croissant d'images améliore considérablement les performances de classification.
- Par le partage des paramètres dans le temps, le nombre de paramètres reste constant en fonction de la longueur de la vidéo à la fois dans Pooling Feature et les architectures LSTM
- Confirmer que les images de flux optique peuvent grandement bénéficier de la classification vidéo et fournir un avantage lorsqu'elles sont couplées avec des LSTM

2.3.3. Architecture basé sur les réseaux convolutifs (CNN) plus les réseaux récurent (RNN)

La classification est une tâche fondamentale et importante pour la compréhension et la classification générale des scènes vidéo. Dans le travail de *Russo et al.*[**39**], une approche d'apprentissage profond est proposée en combinant à la fois les réseaux de neurones convolutifs et récurrents. Les CNN est pour extraire des caractéristiques d'apprentissage ou des indices visuels. Les RNN est utilisée pour déterminer la relation entre les trames dans le domaine temporel. La normalisation par lots a été effectuée après chaque activation, RMSProp a été

utilisé comme optimiseur. L'avantage de cette architecture est de construire un système qui peut classer les actions de football, le cricket, le tennis, le basket-ball, le hockey sur glace en cinq classes différentes de sports sur la base d'informations visuelles uniquement.

Cette approche consiste également à fournir le system par une séquence des images colorées en RGB (frames), chaque image est envoyée à une couche de convolution séparée. La partie de convolution est en quatre couches : la première couche consiste en 32 *Feature Map* avec un *kernel* de (3*3), et est suivie par *Max Pooling* de *Stride* de (2*2). Les trois couches suivantes sont trois types différents de couches de convolutions dilatées. Ce dernier est efficace pour apprendre les relations entre les séquences d'action humaine et leur contexte environnemental. Dans la sortie, une unité linéaire rectifiée (*ReLU*) est utilisée comme fonction d'activation et placée à la sortie des couches CNN. Enfin, les résultats sont renforcés de la partie de convolution et la partie récurrente par une couche de entièrement connecté (*fully connected*), en utilisent 64 neurones.

2.3.4. Architectures basé sur les réseaux 3DCNN et les réseaux Long Short Memory (LSTM)

Noorkholis et al. **[40]**, proposent une architecture d'apprentissage profond pour résoudre le problème de reconnaissance et classification des gestes en situation d'application en temps réel (application vidéo). L'architecture proposée est une architecture combinée du réseau neurone de convolution tridimensionnel (3DCNN) suivi du modèle de mémoire à court terme à long terme (LSTM). De plus, ce modèle été utilisée pour extraire les caractéristiques spatio-temporelles surtout avec la reconnaissance dynamique des gestes, suivi par la suite par un contrôle *Finite State Machine* (FSM) qui communique le modèle pour contrôler les résultats de la décision de classe en fonction du contexte de l'application et restreindre certains flux de gestes dans la classification et pour limiter les classes de reconnaissance.

L'aspect multimodal de l'architecture proposée se compose de couches 3DCNN, d'une couche LSTM stack et d'une couche entièrement connectée suivie de la couche *Softmax*, et *Batch normalization* utiliser pour accélérer le processus d'apprentissage. C'est en faisant permettre au modèle d'utiliser des taux d'apprentissage beaucoup plus élevés et moins préoccupé par l'initialisation.

3. Étude comparative entre les architectures de classification des vidéos :

Comme nous avons discuté dans section précédentes, les réseaux de neurones convolutifs sont des modèles d'apprentissage profond inspirés du comportement humain qui remplacent les trois étapes par un réseau neuronal unique qui est apprend les valeurs de pixels brutes aux sorties du

classificateur. La structure spatiale des images est explicitement mise à profit pour la régularisation grâce à un couplage restreint entre les *Kernels*, les convolutions et des neurones locaux spéciaux de création d'invariance (mise en commun maximale).

La plupart des approches qui ont été proposés précédemment démontrer l'efficacité de l'utilisation de l'apprentissage profond pour la classification vidéo. Cependant, ces approches utilisent généralement des modèles profonds renommées dans le domaine de traitement d'image et de la reconnaissance vocale. La nature complexe des informations un vidéo inclut de nombreux indices spatiaux, temporels et acoustiques, rend les modèles profonds prêts à l'emploi insuffisants pour ***les tâches*** liées à la vidéo à savoir, classification, identification, détection, etc. Ce qui met en évidence la nécessité d'un modèle d'apprentissage profond personnalisé pour interpréter efficacement les informations spatiales et acoustiques, et surtout pour modéliser la dynamique temporelle.

Les travaux cités auparavant ont prouvés de manière décisive au fil du temps que les réseaux de neurones surpassent les autres algorithmes en termes de précision et de vitesse. Les réseaux de neurones, inspirés du cerveau humain, sont de plus en plus utilisés dans la classification des informations complexes Avec diverses variantes comme CNN, RNN, LSTM, etc. Dans ce point en va faire une étude comparative sur un tableau entre ces approches avec des ensembles critères pour comprendre mieux les fonctionnalités de chaque architecture.

Pour le fait de comparaison, nous proposons quelques critères pour évaluer les architectures d'apprentissage profond pour les tâches d'apprentissage vidéo. Dont, chaque ***architecture*** se caractérise par un nombre de ***tâches*** à savoir, la classification, la détection et l'identification. Par exemple, dans les travaux [**21,39**] propose des architectures d'apprentissage profond à une seule tâche de classification. Par contre dans le reste des approches [**37,40**] proposent des architectures pour deux tâches : la classification et la reconnaissance. Dans les propositions de [**31, 38, 33,34**] c'est classification et la détection. Le ***taux de calcule*** est un critère de grande importance dans le processus de classification. En particulier, ce critère montre la souplesse de cette approche par rapport les autres. Le ***nombre de classe*** est également un critère signifie que dans la sortie de système nous trouvons multi classe [**37, 39, 33, 34, 40,35**] ou classe binaire [**21,31**] ou score de classe [**38]**. Le cinquième critère est la ***précision*** du modèle, est la carte d'identification de la qualité de cette approche dans la classification ; c'est à dire l'architecture proposée est le bon de résoudre ce problème. Par exemple, dans [**38, 39, 33, 40,35**] la précision atteindre entre 89% et 94%, par contre dans le reste approche la précision atteindre entre 50% et 82%. Le dernier critère est le ***pas du temps*** qui utilise pour déterminer la relation entre les trames (frame) et la modélisation du mouvement entre les trames d'image dans une séquence de la vidéo ; ce qui est analogue à la persistance de la vision dans le système visuel humain. Ce critère distingue les

architectures qui contient les réseaux récurrent (RNN) comme [**37, 38, 39,40**].Le ***tableau 1*** représente une étude comparative entre les approches qui ont été proposées pour plusieurs tâches d'apprentissage vidéo telles que la classification, la détection et la reconnaissance. Nous constatons que l'apprentissage profond à l'aide de modèles hybrides (CNN / LSTM, CNN / RNN, etc.) nécessite des ensembles de données supervisés étiquetés manuellement. Cette tâche d'étiquetage est généralement fastidieuse et longues à acquérir, en particulier lorsque les ensembles de données deviennent volumineux. Le ***sous-titrage*** de vidéos en utilisant des techniques d'apprentissage profond [**41, 42, 43,44**] est une direction prometteuse implique d'utiliser pleinement des quantités substantielles de données vidéo non supervisées et de riches indices contextuels pour obtenir une meilleure tâche d'apprentissage.

TABLEAU 2.1-Etude comparatifs entre les architectures

	L'architecture proposée	**Le nombre de tâches**	**Le nombre de classes**	**La précision**	**Pas du temps**	**Le taux de calcul**
R-CNN [21]	GoogleLeNet-RCNN	Classification	Binaire Classe	81.0 %	oui	élevé
CNN+LSTM [37]	CNN + LSTM + Optical Flow	Reconnaissance + classification	Multi Classe	56.41% 76.92%	oui	élevé
CNN+RNN [38]	Conv pooling + GoogLeNet LSTM+Optical Flow	Détection + Classification	Scores de classe	90.4 88.6	oui	élevé
CNN [31]	CNN	Classification .C + Classification .T	Binaire Classe	81.0% 70.09%	non	moyen
CNN+RNN [39]	CONV+GRU	Compréhension + classification	Multi classe	96.66%	oui	élevé
CNN [34]	Faster-R-CNN +SSD. ResNet Inception-ResnetV2 InceptionV3 NASNet MobileNetV2 PNASNet.	Détection + Classification	Multi classe	76.78% 78.76% 75.17% 75.75% 77.45% 74.70%	non	élevé
CNN [33]	Adaboost person detector. GoogleNet transfer learning	Détection + Classification	Multi classe	90.0%	non	élevé
3DCNN + LSTM [40]	3dcnn + lstm + FSM	Reconnaissance + Classification	Multi classe	89% .91%	oui	Faible
CNN [35]	AlexNet	Classification	Multi classe	94.07	non	moyen

CONCLUSION

Dans ce chapitre, nous avons découvert l'apprentissage profond comme un sous domaine de l'intelligence artificielle, très large et qui permet d'appliquer plusieurs méthodes en collaboration avec des techniques ou des tâches d'apprentissages à savoir la classification (CNN, RNN, 3DCNN, LSTM, etc.) saper les données et extraire de nouvelles connaissances qui servent la recherche scientifique avec excellence, et a essayé de clarifier les concepts qui nous intéressent dans ce domaine et chaque méthode d'apprentissage profond à ces caractéristiques et son mode d'application qui diffèrent à des autres méthodes. Nous avons mentionné également une étude comparative entre ces méthodes avec des ensembles de critères. Ces méthodes représentent ensemble la force du domaine.

CHAPITRE 03

CONCEPTION ET IMPLÉMENTATION D'UN MODÈLE DE DÉTECTION ET DE CLASSIFICATION

1. Les techniques d'analyse dans le football.
2. La détection des joueurs dans les séquences des vidéos.
3. La génération des datasets.
4. Les modèles de classification.
5. Validation des modèles.

INTRODUCTION

La lecture du jeu et l'analyse tactique en football est d'une grande importance dans les décisions d'entraîneur et les résultats des matchs. Récemment, le traitement vidéo a occupé une place importante dans le travail de coaching soit pour préparer leurs équipes, soit pour déstabiliser l'équipe adverse, ce qui a attiré l'attention de nombreux chercheurs pour démarrer ce domaine de recherche notamment avec l'important développement technologique. L'apprentissage profond est l'une des techniques les plus utilisées au cours de ces dernières décennies et a vu beaucoup de travail dans la classification vidéo. La détection et la classification des vidéos de sport ont été largement explorées en raison des avantages commerciaux potentiels et de l'audience mondiale massive, mais en même temps, cela devient une tâche difficile.

Ce chapitre est contient un travail important basé sur l'apprentissage profond pour la détection des positions des joueurs dans un match de football à partir des séquences vidéos, cette détection se fait de plusieurs vues et de plusieurs caméras. Dans une autre étape notre travail sert à classifier la tactique de jeu d'une équipe à partir du placement des jours dans les frames extraire de vidéo

1. Les techniques d'analyse dans le football

Les enjeux sportifs et financiers du football professionnel ont conduit les entraîneurs sportifs à constamment rechercher les meilleurs moyens d'évaluer et d'améliorer les performances individuelles et collectives. Le besoin d'un retour rapide, objectif, précis et pertinent sur les performances des joueurs en compétition a conduit au développement de systèmes d'analyses de matches perfectionnés. Basés sur les technologies de pointe de l'informatique et de la vidéo, ces systèmes sont devenus un élément capital dans le processus d'entraînement et de préparation pour la performance au plus haut niveau **[65].** Quels sont les caractéristiques de ces nouvelles techniques ? En quoi leurs observations peuvent-elles faire évoluer la pratique ?

Il est donc intéressant de détailler les différents moyens d'analyse ainsi que d'évaluer la pertinence de leurs résultats à travers la performance sportif.

1.1. Les différents systèmes d'enregistrement

1.1.1. Description, utilité et limites des techniques d'enregistrement

Les techniques d'enregistrements sont variées et comportent chacune une utilité spécifique pour chaque séquence de jeu. L'enregistrement manuel est une technique peu onéreuse, rapide et accessible à tous puisqu'il ne requiert que des crayons et du papier. Cependant avec l'évolution technologique ce mode d'enregistrement est de moins en moins utilisé.

L'enregistrement audio, quant à lui, est très peu utilisé malgré qu'il représente une base de données assez élevées, le temps de conversion de ces données est beaucoup trop long pour qu'elles soient exploitables et donc utiles en temps réel.

L'enregistrement vidéo non informatisé est la technique la plus répandu dans le monde du football. Elle permet un accès illimité à tous ce qui se passe sur le terrain, L'observateur peut ainsi comme bon lui semble revenir en arrière ou s'arrêter sur un fait de jeu qu'il semble pertinent. Cette technique est souvent complémentaire avec une analyse manuelle puisqu'elle n'est pas informatisée.

Les systèmes d'analyses informatisés ont progressivement été adopté par les équipes professionnels qui cherchent à améliorer leur préparation et bien évidement leurs résultats. Par exemple, le système AMISCO localise les positions et suit les déplacements des joueurs, 25 fois par seconde, en utilisant simplement du matériel vidéo et informatique basique sans le besoin d'avoir à intervenir sur la rencontre. La vidéo informatisée est un moyen qui facilite l'enregistrement d'un très grand nombre de données, cependant elle demande du personnel qualifié. La taille du matériel est un frein à son utilisation, mais surtout son coût qui est très élevé **[64].**

Enfin le système de navigation par satellites, Global Positioning System (GPS), est une technique qui permet de déterminer avec précision, la position d'un joueur sur le terrain ainsi que sa vitesse de déplacement instantanée, sa direction, la distance parcourue et la durée du déplacement. Contrairement à l'enregistrement manuel, audio et vidéo, le GPS permet d'obtenir le nombre d'arrêts complets, mais ne permet pas de distinguer les différents types de courses (latérales, avant ou arrière). De plus, le port de tels appareils est interdit en compétition. Malgré ces limites, c'est un outil intéressant dans l'étude des actions du football lors de la simulation de matchs ou d'opposition lors des entrainements.

1.1.2. Statistiques et caractéristiques du jeu

Les informations fournies par les systèmes d'analyse aident à concevoir et à mener des programmes d'entraînements en fonction de la prestation des joueurs. Ces outils fournissent des informations statistiques détaillées sur les forces et faiblesses des adversaires et aident dans le choix de l'effectif et dans le recrutement de joueurs (figure 1). Grâce à cela, un entraîneur peut ainsi évaluer les performances de son équipe à travers toute une saison. A titre général, par exemple, de part, les analyses et statistiques faites au niveau professionnel et international, une équipe a plus de chance de marquer à partir d'un corner rentrant (71%), qu'un corner sortant (21%). Les équipes gagnantes ont un meilleur ratio du nombre de buts par rapport au nombre de tirs (5 à 1 contre 16 à 1 pour les équipes perdantes). Cela montre que le réalisme et l'efficacité à marquer des buts est l'un des principaux facteurs qui différencie une équipe de niveau international à celle d'une équipe de niveau inférieur. De plus, la majorité des buts sont marqués après des séquences de jeu de moins de 3 passes et de moins de 10 secondes. En ce qui concerne, la performance physique, un joueur de haut niveau doit en moyenne courir 10 kilomètres et effectuer une succession de sollicitations explosives de l'ordre d'une action de haute intensité toute les 60 secondes. Les sprints dépassent rarement une distance et une durée de 20 mètres et 4 secondes respectivement mais la performance physique dépend du poste du joueur et des tactiques de l'équipe **[65]**

1.2. L'importance et l'observation de l'analyse de match

1.2.1. Le besoin d'analyse approfondie

Certains entraîneurs se passent de l'enregistrement systémique (notes écrites, enregistrement électronique, etc...) se référant à leur seule mémoire pour se souvenir des évènements d'un match. Cependant, c'est loin d'être suffisant puisque pour un entraîneur de niveau international, seul 45% de ce qui se passe durant un match est retenu et la probabilité de se rappeler des moments critiques du match est de 42%. Lorsqu'il y a une phase arrêtée, l'entraîneur peut

aisément prendre en compte plusieurs informations, mais dès que le ballon est remis en jeu, l'entraîneur se retrouve face à une multitude d'actions prenant place simultanément. D'une certaine manière, une analyse plus aboutis permet de mieux connaître l'évolution des besoins physiologiques et ainsi mieux planifier et éviter les blessures des joueurs. Cependant, la quantité de résultat fournie par des systèmes plus élaborés (informatique, vidéo, etc...) exige un travail assez conséquent de la part du staff technique.

1.2.2. Les différents types d'analyse

L'analyse permet de stocker un ensemble important d'informations et de rétroactions (aussi bien qualitatif que quantitatif) qui vont permettre au footballeur de se corriger et donc mieux se connaître. L'information et la rétroaction sont perçus de deux façons : intrinsèque et extrinsèque. Intrinsèque puisqu'elle correspond à l'information obtenue par le sportif lui-même, les différentes sensations de son corps par rapport à l'exercice fourni, mais aussi extrinsèque qui se rapporte aux informations obtenus par observation vidéo et de l'analyse de l'entraîneur, c'est-à-dire un point vue extérieur à celui du sportif qui permet de déterminer les limites techniques et tactiques du joueur. **[65].** Pour cela deux types d'analyses sont utilisées, l'une qualitative et l'autre quantitative. L'une complète l'autre, mais l'analyse qualitative est plus difficile et compliquée à réaliser car elle demande une formation spécifique des observateurs qui peut s'avérer aussi bien avantageuse que handicapante. La combinaison information/rétroaction permet donc de dresser un bilan qualitatif et quantitatif de la performance du joueur et de mettre en place un travail plus spécifique lors de sa préparation physique.

1.3.Les systèmes de jeu

Un Système de Jeu peut être décrit comme une structure générale pour les plans tactiques de l'entraîneur. C'est la forme ou formation de base d'une équipe qui implique les rôles défensifs ou offensifs d'un joueur en affectant des zones de terrain distinctes à chaque joueur.

Dans le football moderne, les différents systèmes peuvent être décrits par une combinaison de chiffres. Un exemple courant étant le système 4-4-2 dans lequel les trois chiffres représentent les différentes zones du terrain (défense, milieu, attaque) et le nombre de joueurs affectés à ces zones: 4 défenseurs, 4 milieux de terrain et 2 attaquants. On caractérise souvent ces groupes de joueurs d'unités, ex. l'unité défensive.

Au moment de choisir un système de jeu particulier, l'entraîneur doit prendre en compte un certain nombre de facteurs : le type de match, le fait de savoir si le match est joué à domicile ou non mais aussi d'autres facteurs tels que le temps, la taille du terrain, etc. qui jouent tous un rôle important. Pour plus d'information sur les différents facteurs qui affectent la qualité de la

performance de jeu, consultez notre section sur les facteurs externes.

En fin de compte, c'est la qualité et le type de joueurs disponibles qui s'imposent pour le choix final de l'entraîneur. Par exemple, les joueurs de milieu de terrain sont souvent connus pour avoir à courir de plus grandes distances que les défenseurs. Par conséquent, un niveau de forme physique plus élevé est requis pour jouer à cette position. De même, il est indispensable que les défenseurs aient un bon jeu de tête. L'entraîneur pourra décider de donner un rôle défensif important aux joueurs qui ont cette compétence. Les tactiques choisies par l'entraîneur vont aussi affecter son choix de système de jeu. Ainsi, des équipes qui effectuent un marquage près du but adverse pour récupérer le ballon, pourront avoir besoin de plus de joueurs dans les zones de milieu de terrain.

Dans l'ensemble, on peut aussi faire la liste des caractéristiques requises dans les systèmes de jeu modernes :

- Les joueurs doivent être répartis sur l'ensemble du terrain afin de maintenir l'équilibre de l'équipe et d'empêcher que les joueurs soient surpassés en nombre dans des zones dangereuses.
- Le plus grand nombre possible de joueurs autour du ballon pour empêcher les équipes adverses de dominer et augmenter les chances de récupération.
- Une participation maximale des chaque joueur tant en attaque qu'en défense et chaque joueur doit avoir le même niveau de responsabilité au sein de l'équipe.
- Une flexibilité maximale du système : les joueurs doivent pouvoir échanger leur rôle sur la longueur et la largeur du terrain et aussi, chaque joueur doit pouvoir exprimer librement ses talents et compétences de la meilleure façon possible.

Globalement, l'entraîneur devra choisir les joueurs les plus adaptés lors de chaque match. Mieux le joueur est préparé, plus grandes seront les chances pour le système de jeu de fonctionner. Cependant, il ne faut pas oublier qu'en fin de compte, ce n'est pas le système de jeu qui gagne le match mais les joueurs eux-mêmes ! Aucun système de jeu ne pourra rattraper une mauvaise technique et un mauvais marquage !

1.4. Les tactiques de football

La victoire ou la défaite ne dépendent pas que des 11 joueurs présent sur le terrain, elle dépend aussi de la tactique football mise en place ! Tactique qui, bien souvent, fait partie du rituel d'avant match.

Le football étant un sport collectif, les questions d'intelligence collective sont primordiales. L'issue d'un match est aussi liée à l'aspect tactique du jeu mis en place par l'entraîneur. Il est donc important que chaque joueur assimile son positionnement et son rôle précis dans le schéma

tactique de jeux choisi.

Le résultat d'une rencontre dépendra donc de la capacité des onze joueurs à pratiquer un football homogène en adéquation avec les choix tactiques de l'entraîneur. On parle alors de lucidité, d'intelligence de jeux.

Les différents schémas tactiques au football

1.4.1. La tactique du 4-4-2

Le 4-4-2 (4 défenseurs, 4 milieux de terrain et 2 attaquants) est l'un des schémas tactiques les plus utilisés aujourd'hui dans le football. On peut considérer que c'est un schéma de jeu qui est déployé pour dominer l'adversaire. Les défenseurs jouent généralement alignés, usant beaucoup du hors-jeu. Dans ce schéma, il existe ensuite deux dispositifs tactiques :

- Le 4-4-2 avec milieu de terrain en diamant *:* Il y a deux milieux défensifs centraux postés devant la défense, et deux milieux de terrain latéraux plus offensifs assimilés à des ailiers, dont la mission est d'alimenter les 2 attaquants.
- Le 4-4-2 avec milieu de terrain en losange (4-1-2-1-2) : On retrouve les 4 défenseurs, un milieu défensif posté juste devant la défense, deux milieux de terrain sur les ailes, et un milieu offensif situé juste derrière les deux attaquants.

1.4.2. La tactique du 4-3-3

Le 4-3-3 (4 défenseurs, 3 milieux de terrain et 3 attaquants) est un schéma tactique qui favorise l'attaque. Ce schéma n'est pas très utilisé en France, car les entraîneurs optent pour une bonne défense avec un milieu de terrain solide composé de 4, voir 5 joueurs. Cependant ce schéma s'emploie souvent en cours de match lorsqu'une équipe est menée au score (en faisant entrer un attaquant supplémentaire).

Généralement, les latéraux ont une vocation offensive et doivent utiliser le couloir pour apporter leur soutien. Les milieux de terrain sont plus défensifs (deux récupérateurs et un relayeur ou vise et versa) et effectuent plus de travail au niveau de la récupération que dans le 4-4-2 classique. Trois attaquants, dont un 9 et deux ailier qui peuvent permutés. C'est le système de jeu mis en avant par le FC Barcelone.

1.4.3. La tactique du 3-5-2

Le 3-5-2 (3 défenseurs, 5 milieux et 2 attaquants) est un schéma qui a pour objectif de "gagner" la bataille des milieux en créant le surnombre. Ce schéma est censé produire plus de jeu grâce aux milieux offensifs.

Il existe plusieurs variantes de 3-5-2 qui peuvent être employées suivant le contexte du match et l'évolution du score :

4 milieux alignés + 1 milieu offensif

2 milieux défensifs et 3 milieux offensifs en soutien des attaquants

3 milieux récupérateur et 2 meneurs positionné en escalier

1.4.4. La tactique 5-3-2

Le 5-3-2 (5 défenseurs, 3 milieux et 2 attaquants) est un système défensif dans lequel un des défenseurs prend le rôle de libéro, c'est à dire qu'il décroche de l'alignement constitué par les 4 autres défenseurs. Les deux défenseurs latéraux jouent un peu plus haut afin de renforcer le milieu de terrain, et d'apporter un soutien offensif aux milieux/attaquants. Cette tactique est mise en place lorsque l'équipe se sent inférieur à l'adversaire (coupe de France) et qu'elle choisit de se replier plutôt que d'attaquer.

1.4.5. La tactique 4-5-1

Le 4-5-1 (4 défenseurs, 5 milieux et 1 attaquants) est un système de jeu très défensif qui vise à étouffer son adversaire au milieu de terrain. Il n'y a qu'un seul attaquant en pointe qui se doit de jouer en pivot/remises dos au but. Ce système est très souvent utilisé par les équipes jouant en contre. Ce fut la tactique principalement utilisée par Raymond Domenech lors de la coupe du monde 2006.

1.4.6. La tactique 5-4-1

Cette formation absolument défensive est généralement développée durant le cours du match par des équipes ayant déjà marqué suffisamment de buts, ou voulant à tout prix éviter la défaite et opérant en contre-attaque.

Dans le cas du 5-4-1, on retrouve souvent une défense à quatre avec un libéro. Le milieu est lui disposé comme celui d'un 4-4-2 carré.

1.4.7. La tactique 3-6-1

Le 3-6-1 est une formation rarement utilisée, ce qui pourrait apporter des opportunités intéressantes dans le football de haut niveau. Le but de cette analyse est de donner un petit exemple d'un système utilisant cette formation.

FIGURE 3.1- Exemple de tactique du jeu

2. La détection des joueurs dans les séquences des vidéos

La détection des objets parmi les problèmes fascinants avec la vision par ordinateur ; pour cela on a créé un algorithme basé sur la bibliothèque optimisée open cv. L'objectif de cette détection et de localiser l'emplacement des joueurs et elle se fait par la détection des couleurs des objets (les joueurs) en prenant en considération le conteur de l'objet qui doit être de couleur vert pour garantir que l'objet détecté se trouve dans le terrain et non dans les tribunes par exemple. Cette détection ce fait de n'importe quelle vue et n'importe quelle caméra que ce soit leur position (niveau de la ligne médiane, mi-hauteur de la ligne médiane, dans l'axe central du terrain de jeu, la longueur à l'arrière des buts, à hauteur de la ligne des 16 mètres).

La détection nous permet de générer un dataset en premier lieu puis elle nous permet de générer les frames à classifier en seconde lieu.

Les étapes de notre algorithme se résument dans les étapes suivantes:

- La lecture du vidéo frame par frame.
- La détection des couleurs en temps réel dans la vidéo.
- La détection des objets spécifiques avec une dimension donnée en temps réel.
- La suivre des objets (les joueurs) en temps réel après la détection
- Calculer un nombre définit pour les objets détecté (nombre des objets données)
- Capturer de la vidéo image par image avec un spécifique nombre framerate.
- La conversation de vidéo, d'images couleur à images noire et blanc.
- La sauvegarde de la vidéo (images successives) dans un dossier donnée (par le path).

Notre approche de détection est présentée dans le schéma illustratif suivant:

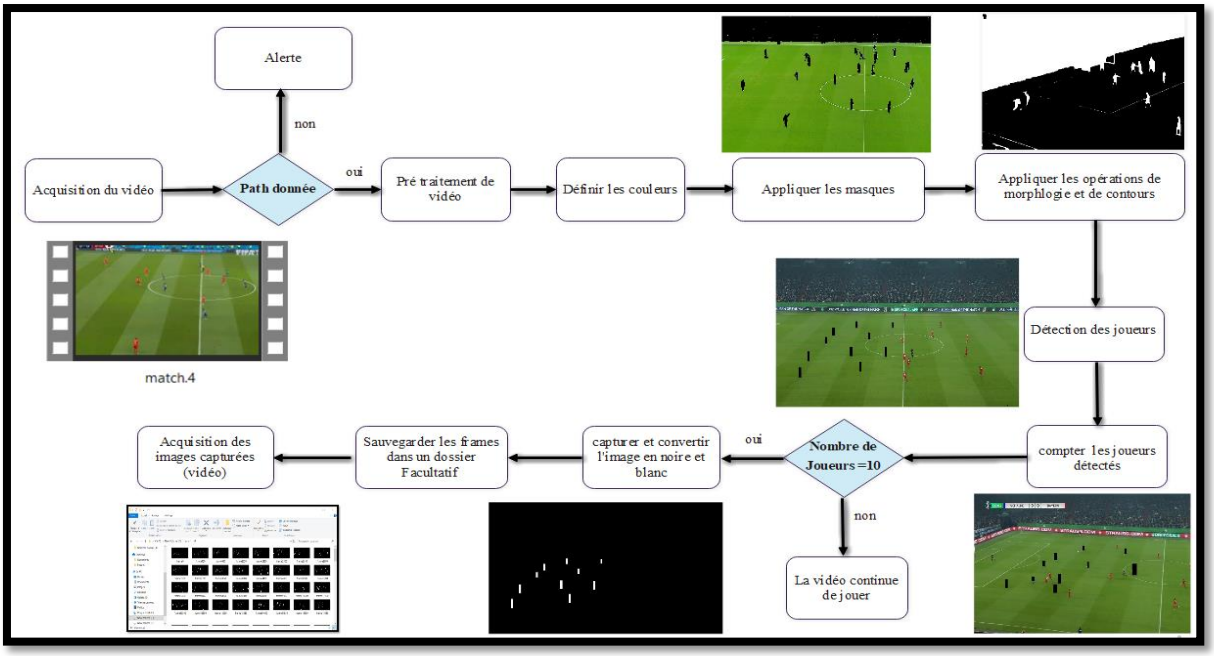

FIGURE 3.2- Système de détection des joueurs

Ce schéma ensembles des étapes où chaque étape exécute des instructions pour déterminer cette structure. En va discuter ces étapes dans la section suivante :

- **Importation des bibliothèques nécessaires et la lecture de la vidéo**

Ces package est destiné pour la détection et la segmentation des objets dans une vidéo (temps réel) ou image .il utilisé par l'apprentissage profond pour la reconnaissons et la détection des objets [25]; après, en va lire la vidéo (dans un base créé) par une méthode à partir un path donner.

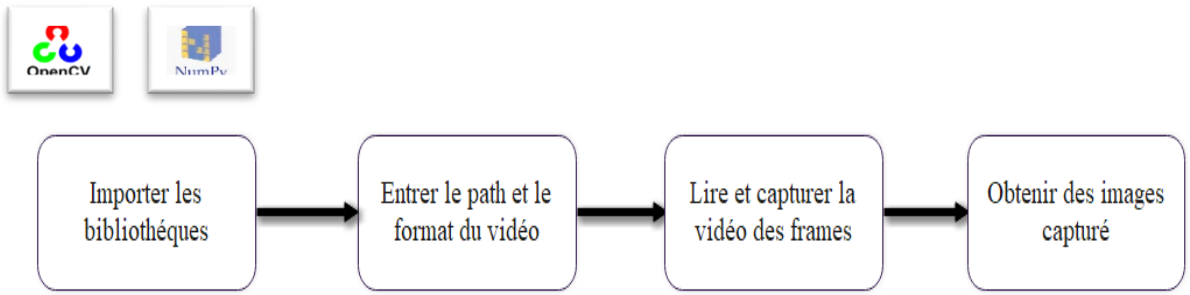

FIGURE 3.3-Lecture de la vidéo

- **Définir quelque variable nécessaire et définir le frame rate**

Le nombre d'images plein écran consécutives qui s'affichent chaque seconde. Il s'agit d'une spécification commune utilisée dans la capture et la lecture vidéo et est également utilisée pour mesurer les performances des jeux vidéo

- **Pré traitement de la vidéo** Lire la vidéo image par image et transformer les images au format HSV par une autre fonction, la nécessité de le convertir en HSV est pour détecter tout fond de couleur spécifié allant de codes de couleur spécifiques.

FIGURE 3.4-Pré traitement vidéo

- **Application des masques**

Nous définissons deux masques de notre gamme de couleur : vert pour détecter le sol vert, de sorte que l'autre partie deviendra noire, et une autre couleur choisissez par nous pour détecter les joueurs de même couleur de l'équipe adversaire dans le sol vert, il sera donc facile de détecter

nos joueurs. Cette étape est faire par l'opération de mask qui détecter les objets définissent par des couleurs

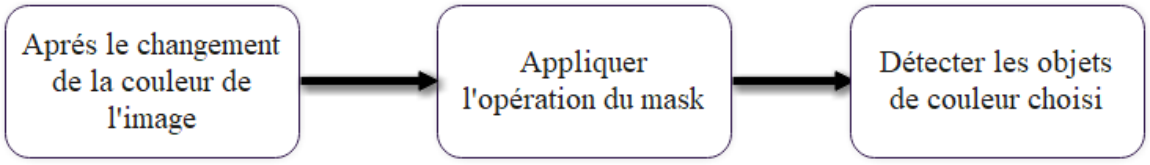

FIGURE 3.5-Détection des objets

- **Application des opérations de contour et de morphologie**

Dans cette étape en va définir un kernel pour effectuer une opération morphologique et une opération de contour dans threshold image pour obtenir une meilleure sortie. L'opération de contours joignant tous les points le long de la frontière d'une image qui ont la même intensité :

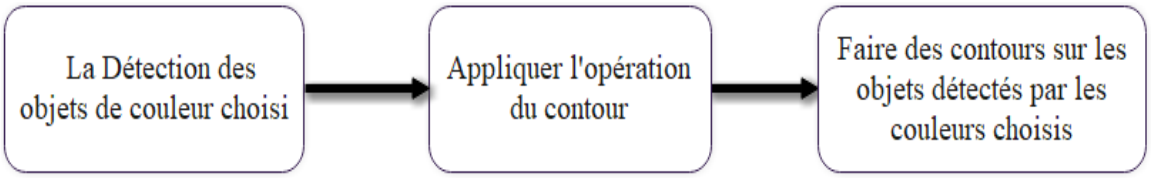

FIGURE 3.6-Application des opérations de contours

Et l'opération de morphologique remplira le bruit présent dans la foule, ce qui permet de réduire les fausses détections.

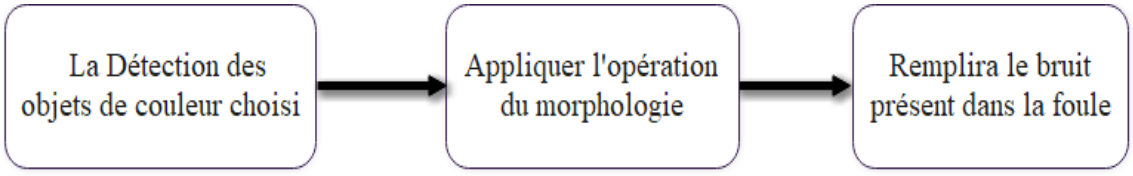

FIGURE 3.7- Application des opérations de morphologie

- **Détection des joueurs**

Maintenant en trouver des contours sur chaque image. Après avoir trouvé le contour, nous vérifierons uniquement les contours, où la hauteur est supérieure à la largeur, ceux-ci seront détectés par les joueurs. Après nous effectuerons une opération de masquage sur les joueurs détectés pour détecter la couleur de leur maillot avec la couleur donnée dans les thresholdes images (segmenter les joueurs avec la couleur spécifique) avec des rectangles noir :

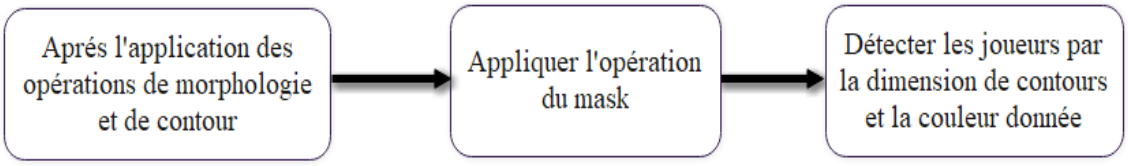

FIGURE 3.8-Détection des joueurs

Et une opération de rectangle pour faire des rectangles sur les joueurs détectées comme illustrer dans l'exemple de la figure au-dessous :

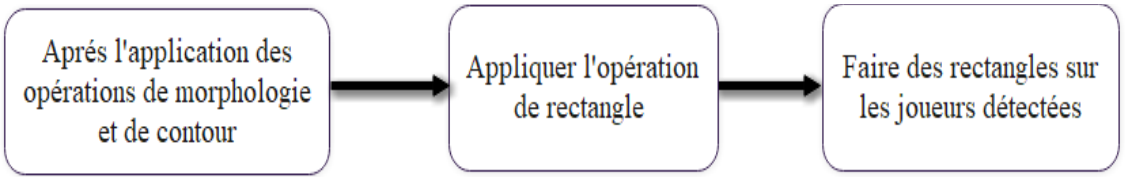

FIGURE 3.9-Détection des joueurs avec des rectangles

- **Capture et comptage les objets détectés**

Dans cette section notre système calculer 10 objets qui sont notre équipe ;et capturé les images avec un temps prédéfini (le temps de la capture entre image et autre image).après, en transformer l'image qui contient 10 joueurs en image noire et blanc.

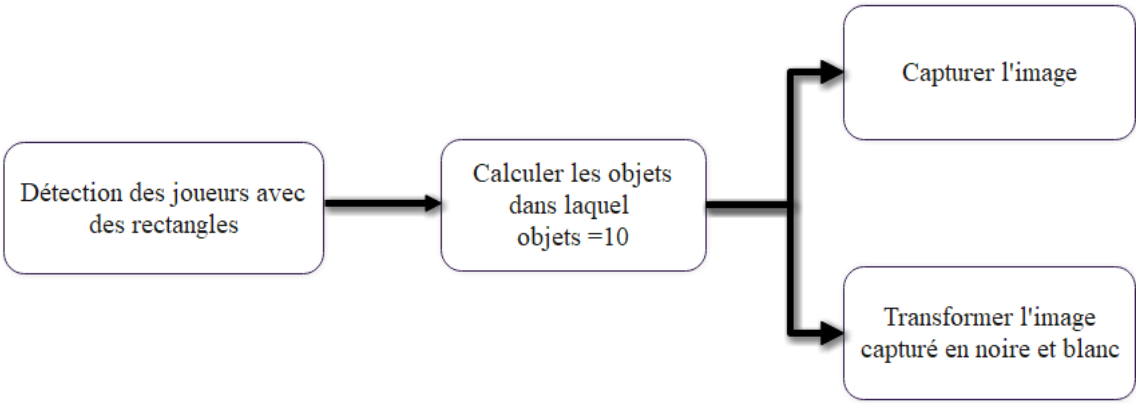

FIGURE 3.10-Capturer les images comptée

- **Sauvegarde les frames capturés**

Dans la fin de cet algorithme en enregistrer les images successives capturées (la vidéo) dans un dossier spécifique choisi par un path donnée par nous

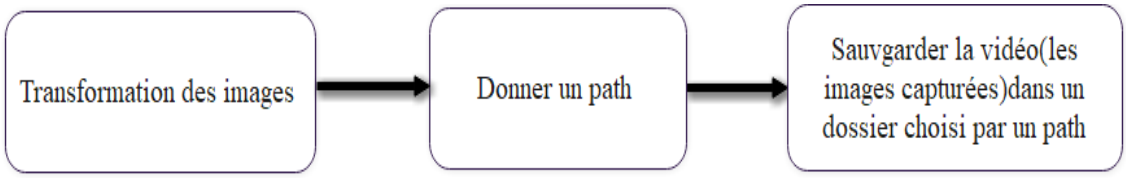

FIGURE 3.11-Sauvegarder les frames capturés

3. Génération des datasets

Pour l'évaluation des performances, nous avons sélectionné un ensemble de données diversifiés comprenant des vidéos de football (des matches) à partir de You Tube. L'ensemble de données comprend 19 vidéos de 45 min de différents équipes avec différents couleurs (rouge, vert, jaune, bleu, blanc…..etc.) .De plus, nous avons inclus les vidéos sportives de différents genres et tournois dans notre jeu de données. Les vidéos représentent différentes conditions d'éclairage (c.-à-d. Lumière du jour, lumières artificielles). Le datasets vidéos de l'ensemble de données sont composées de différents types de match, c'est-à-dire plusieurs positions de capture du terrain comme la montre la figure ci-dessous.

FIGURE 3.12-Vidéos de match du foot

Après l'application de notre algorithme de détection cité au-dessus et l'obtention des images noir et blanc, on a passé à l'aide de l'expert pour la classification des images générées. La résultat de cette étape est l'obtention de 8 catégorie de tactique de jeu comme illustré é dans la figure suivante :

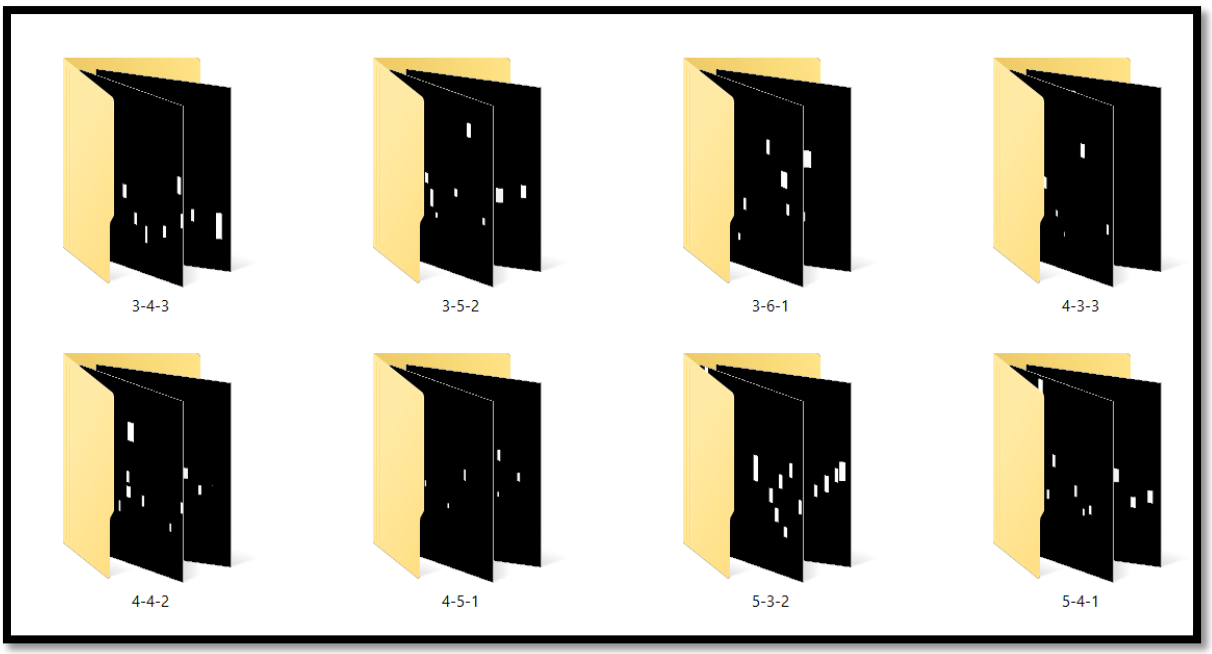

FIGURE 3.13-Basse des vidéos

4. Les modèles de classification

CNN est un modèle de réseau neuronal spécial proposé par LeCun, qui est utilisé pour la reconnaissance d'images de documents, et a fait une grande percée dans la classification et la récupération d'images **[45]**, la détection de cibles **[46]** et ainsi de suite. Le CNN profond réduit les dimensions de l'image en augmentant le nombre de couches cachées (couche convolutionnelle) et extrait les éléments d'image clairsemés dans un espace de faible dimension. En raison de leur partage de poids, CNN a beaucoup moins de neurones et de paramètres, il est donc plus facile de s'entraîner. Le CNN décomposer en plusieurs architectures performantes ; parmi les architectures

4.1. Modèle Alexnet proposé pour la classification

Le modèle Alex-Net se base sur les convolutions pour apprendre automatiquement et de manière supervisée les caractéristiques spatiales pour la classification des frames capturés à travers les vidéos. Nous allons décrire dans un premier temps l'architecture, puis nous avons intéressé à l'apprentissage en jouant avec les paramètres.

4.1.1. Architecture de réseau

L'architecture sur la Figure ci-dessus sur la dataset générés par le modèle de détection décrit dans la section précédente. Cet exemple comporte 13 couches : Une couche d'entrée, une couche de sortie et 11 couches cachées.

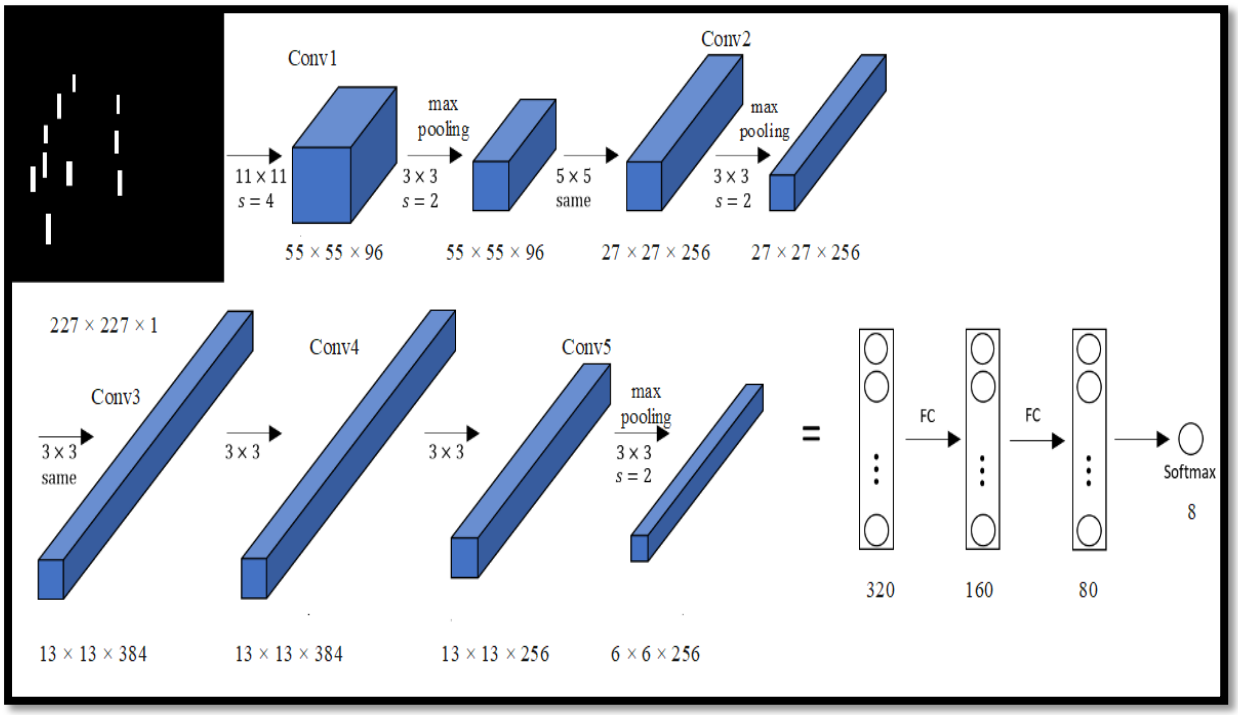

FIGURE 3.14-Architecture proposé Alexnet

Le réseau prend en entrée un volume spatial de taille M × N × T, c'est à dire une suite de T images successives de tailles M × N pixels chacune. Le choix de T est alors important, étant donné que :

- Une valeur trop faible de T conduit à des segments qui ne contiennent pas assez d'informations spatiales discriminantes.
- Une valeur trop importante de T augmente considérablement la qualité d'apprentissage (besoin du « data augmentation »).

Nous proposons donc d'entraîner le modèle Alex-Net à extraire des caractéristiques sur les séquences obtenues en découpant les séquences vidéo entières en frames en noir et blanc contenant des objets qui représente les joueurs. La nature de ces frames générés en noir et blanc nécessite donc la mise en place de stratégies de classification basée sur le modèle Alex-Net destiné à ce type d'image.

La couche de sortie contient 8 neurones qui représentent les différentes tactiques de la joue, et les couches cachées se répartissent en quatre catégories :

4.1.1.1.Les couches de convolution

Notre architecture comprend 5 couches de convolution laquelle :

- **La première couche** convolutionnelle **(conv 1)** est destiné à l'extraction d'identité (feature extraction).il effectue une convolution avec 96 filtres de taille 11x11 pour filtrer une image d'entrée de taille 227 × 227×1. Cette couche est effectuée avec une stride de (4,4)

avec un padding valide, la fonction d'activation utilisée est RELU. Le résultat de cette couche est une Channel de taille 55x55x96.

- **La deuxième couche** occupe la troisième position dans le modèle.il présente de nombreuses similitudes avec la première couche. Cette couche prend en entrée la sortie de la première couche convolutionnelle après une couche pooling avec 256 filtres de taille 5× 5 ; et une foulée (stride) de (1,1), le résultat de cette couche et une Channel de 27x27x256.

- **La troisième couche** occupe la cinquième position du modèle avec 384 filtres de taille 3x3 avec un stride (1,1) et sans utiliser le padding avec l'application d'une fonction d'activation RELU. Le résultat de cette couche est une Channel de 13x13x384.

- **La quatrième couche** occupe la sixième position est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 13×13×384.

- **La cinquième couche** vient directement après la couche de convolution précédente. Elle se caractérise par 256 filtres de taille 13x13 avec un stride (1,1) et sans padding. Le résultat de cette couche est un Channel de 13x13x256.

4.1.1.2.Les couches de pooling

Notre architecture se compose de 3 couches de pooling (maxPooling) pour réduire la dimension spatiale (feature map) des couches convoluionnelles

- **La première couche** de maxpooling (**Pool 1**) occupe la deuxième position dans le modèle, elle se caractérise par un pooling de taille 3x3 et un stride (2,2) et elle donne comme résultat une Channel de taille 55x55x96.

- **La deuxième couche** de pooling **(pool 2**) vient après la deuxième couche de convolution et elle dispose de même paramètres que la couche de pooling précédente. Le résultat de cette couche est un channel de 27 x 27 x256.

- **La troisième couche** de maxpooling (**pool 3**) précède la couche fully connected et il dispose de même caractéristique que les deux premières couches. Le résultat de cette couche est une Channel de taille 5 x 5 x 256.

4.1.1.3.Les couches de fully connected

Notre modèle se termine par trois couches entièrement connectées (FC) avec respectivement 320, 160 et 80 unités. En choisissant ce nombre des unités pour minimise le nombre des

paramètres et accélérer le calcule (le taux de calcules). Chaque couche utilisée une fonction d'activation RELU suivi par une couche dropout à la troisième couche pour réduire le problème de overfiting.

Dans la couche de sorti, en utilisent dans notre modèle une couche de 8 unités qui ressemble à nos classes de sortie avec une fonction **Softmax**.

4.1.2. Apprentissage :

Le modèle de classification basé sur le modèle (Alex-net), Incarné avec 5000000 paramètres entraînables ; et un nombre des échantillons de 1441 avant l'augmentation de notre dataset. L'augmentation des données par la technique de data-augmentation nous a permet d'améliorer la précision de notre modèle et de réduire le overfiting**.** Les fonctions d'augmentation utilisées sont :

- Augmentation horizontale inversée (Horizontal Flip).
- Augmentation de retournement verticale (Vertical Flip).

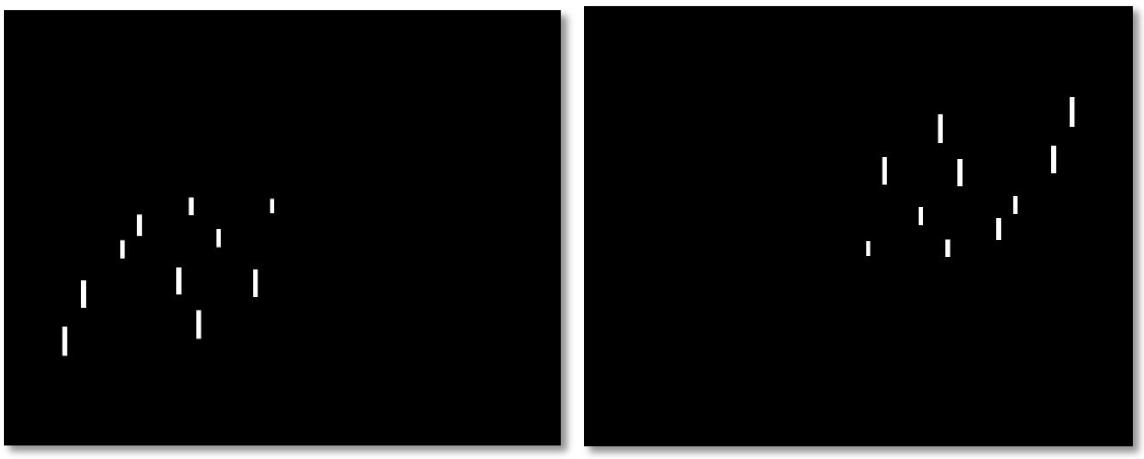

FIGURE 3.15-.Images augmentées avec horizontale inversée (Horizontal Flip)

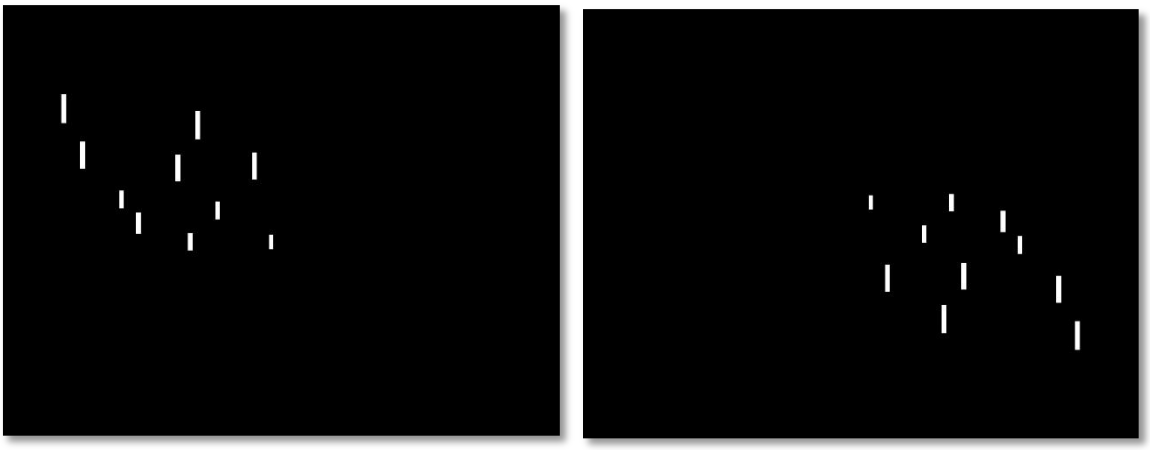

FIGURE 3.16-.Images augmentées avec verticale inversée (vertical Flip)

Après l'augmentation nous avons obtenu 4000 échantillons dans notre jeu d'entraînement (train data), la chose qui nous a permet d'arriver à un taux de précision de 99%.
Les caractéristiques utilisées pour l'entraînement de notre modèle sont :

4.1.2.1.La compilation de modèle :

Nous avons utilisé les paramètres suivants pour la phase de compilation:

- **Compile ()** : Compile définit la fonction de perte, l'optimiseur et les métriques
- **Optimiseur Adam ()** : nous allons utiliser l'algorithme d'optimisation Adam pour mettre à jour les poids de réseau itératifs qui basés sur les données d'entraînement.
- **loss='categorical_crossentropy'() :** en utilisant cette fonction de perte pour la catégorisation d'une étiquette.
- **Metrics = ['accuracy']** : nous utilisons cette fonction pour évaluer les performances de notre modèle.

4.1.2.2. L'entraînement de modèle

Dans l'entraînement de modèle nous utilisons les paramètres suivant :

- **fit_generator ()** : utilisé pour former nos modèles d'apprentissage en profondeur (Alexnet). Il nécessite un générateur pour les données de l'entraînement (training data)
- **Batch size: (32):** la taille de lot (batch size) qui utilisé dans notre modèle Alex-net est 32 parce qu'est un bon point de départ.
- **Step_per_epoch** () : le nombre total d'étapes (lots d'échantillons) qui utilisé dans notre approche est de 4000 échantillons.
- **Epoch ():** en a utilisé les epoch pour séparer la formation en phases distinctes, ce qui est utile pour la journalisation et l'évaluation périodique. Nous utilisons respectivement 50,100,150 nombre d'epoch pour obtenir accuracy parfait pour notre modèle ;mais, en a choisir 50 epoch par ce qu'il après l'epoch 50 la précision de l'apprentissage stabilisé.

4.1.3. Le test

Pour évaluer l'approche, nous avons réalisé les expériences sur un ensemble de données dans une variété de catégories des tactiques de jeu dans notre dataset. Les jeux de test (test data) composés de 321 images respectivement. Ce scénario nous a donné un accuracy de 50 % dans

l'évaluation du modèle. Ce taux réduit nous a ramené à appliquer les mêmes techniques d'augmentation des données utilisé dans la phase d'apprentissage. Cette augmentation nous a permis d'obtenir un jeu de 642 d'image pour arriver à un accuracy de 96 %.

4.1.4. Interprétation des résultats de modèle proposé Alexnet

La précision de l'apprentissage et de test augmente à chaque fois que nous augmente le nombre d'époque, la chose qui explique l'amélioration de la qualité d'entraînement à chaque fois.

La figure 3.17 représente les changements dans La précision de l'entraînement :

La précision de l'entraînement (training accuracy):

- De 0 époque à 50 époques, la précision de l'entraînement augmente jusqu'à la valeur 96% à l'époque 50.
- La précision d'entraînement maximale (training accuracy) est de 99,10%.

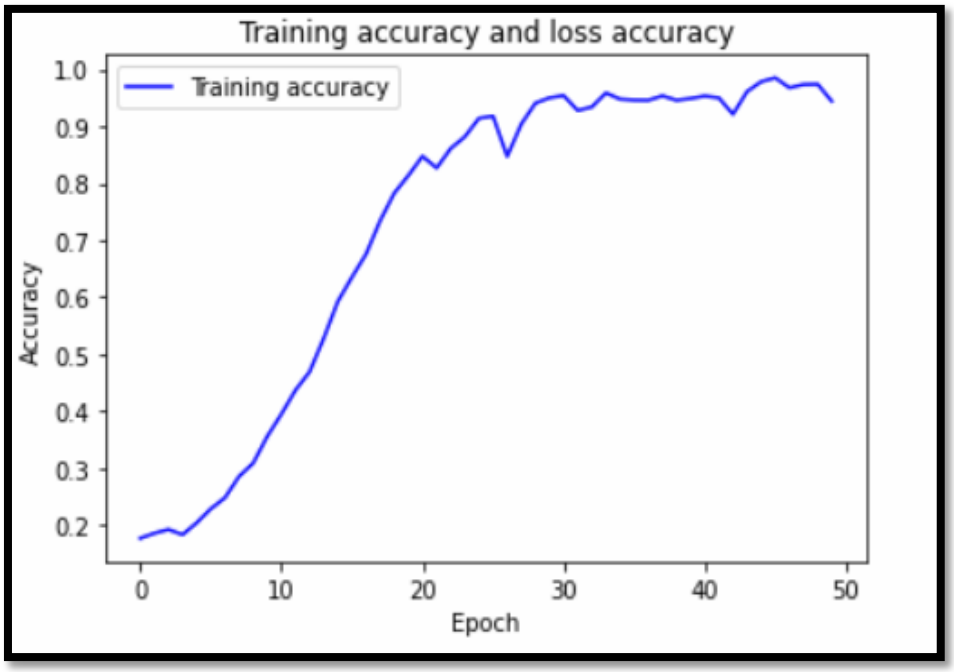

FIGURE 3.17-La précision de l'entraînement « Alex net »

La figure 3.18 représente des changements dans L'erreur de l'entraînement :

La perte de l'entraînement (Loss accuracy):

- De 0 époque à 50 époques, la perte d'entraînement (Loss accuracy) diminue de la valeur 2.0 jusqu'à la valeur proche de 0,110 à l'époque 50.
- La perte d'entraînement minimale est de 0,0410.

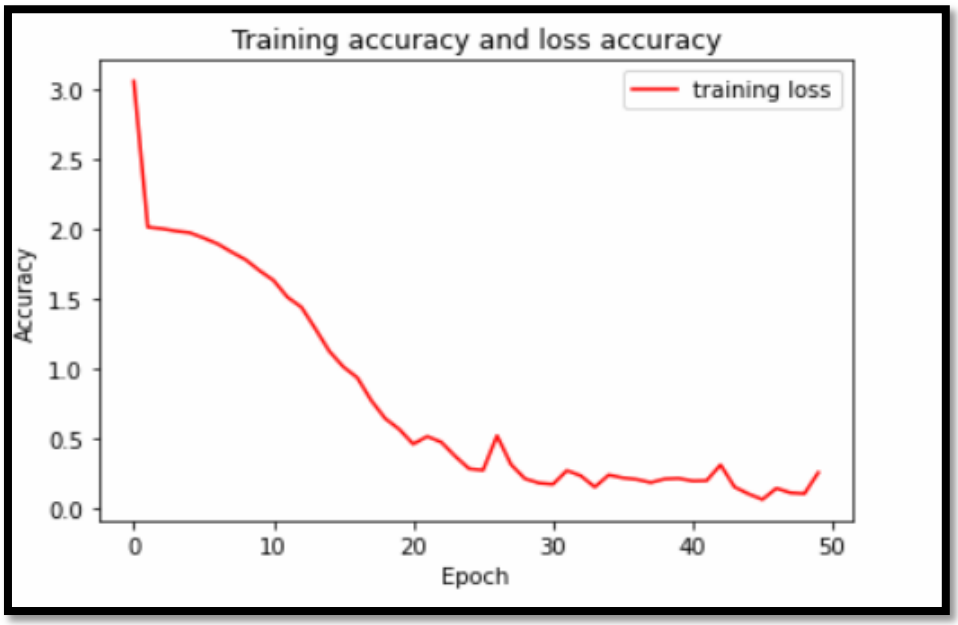

FIGURE 3.18-La perte de l'entraînement « Alex net »

Et en comparant la précision et la perte d'apprentissage dans la même figure nous remarquons de la stabilisation de la précision et la perte sont similaire.

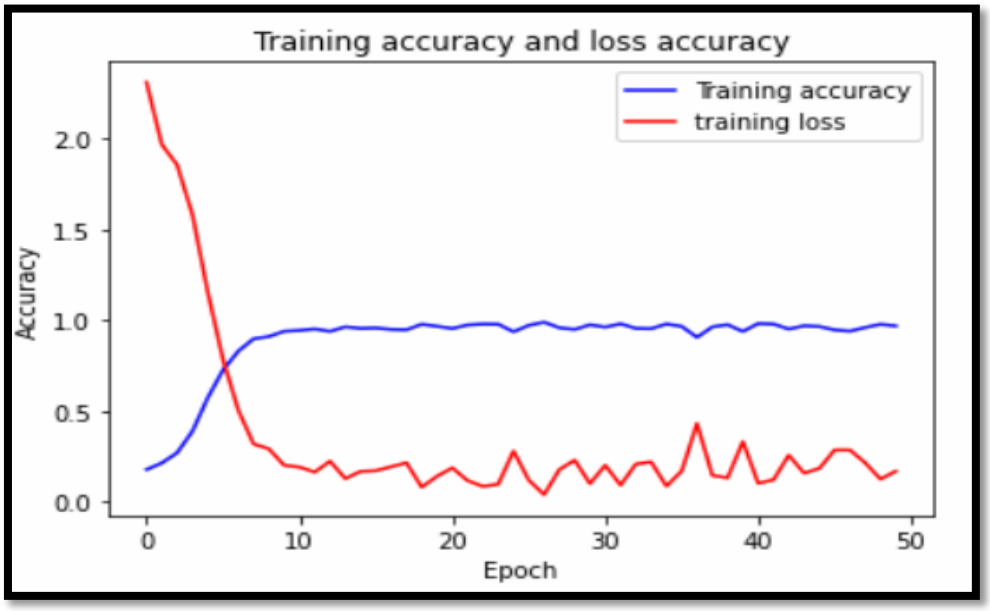

FIGURE 3.19-La précision et la perte de l'entraînement « Alex net »

4.2.Modèle VGGNET16 proposé pour la classification

Le deuxième modèle VGGNET16 destiné à la classification des frames capturés à travers les vidéos est extrait du modèle VGGNET16 modifié qui est basé sur les réseaux de neurones convolutionelles.

4.2.1. Architecture de réseau

L'architecture de modèle VGGNET16 est montré dans la figure ci-dessus et se compose de 20 couches : Une couche d'entrée, une couche de sortie et 18 couches cachées. La dataset utilisé est la même utiliser dans l'architecture alexnet qui a été générés par le modèle de détection décrit dans la première section de ce chapitre.

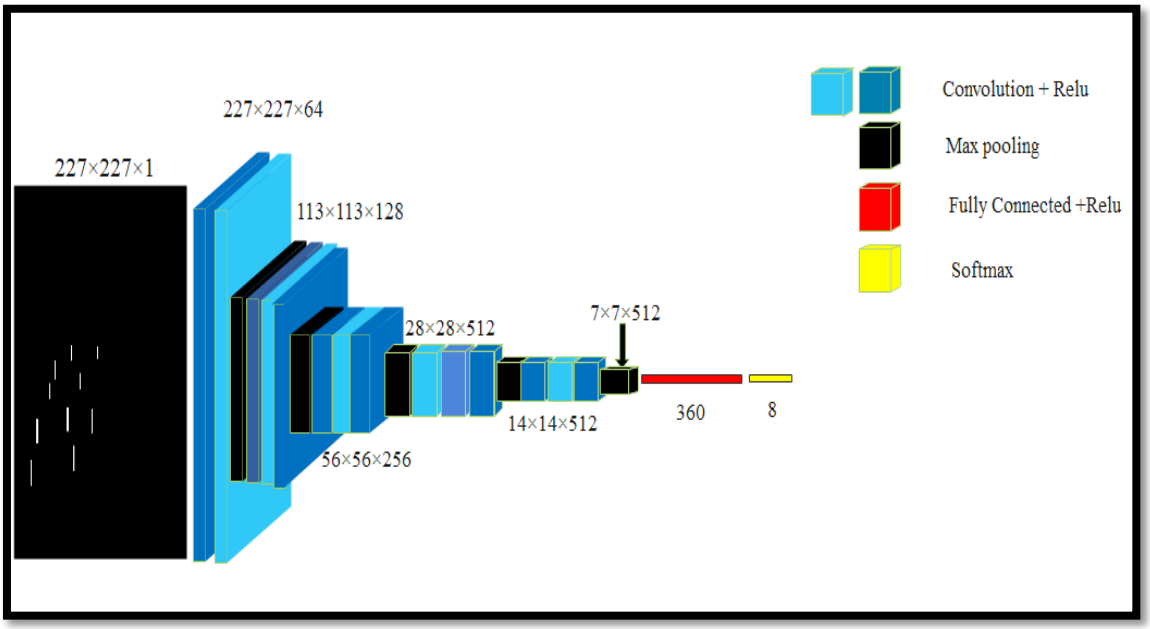

FIGURE 3.20-Architecture proposé VGGNET16

Le réseau prend en entrée un volume spatial de taille M × N × T, c'est à dire une suite de T images successives de tailles M × N pixels chacune. Le choix de T est alors important, étant donné que :

- Une valeur trop faible de T conduit à des segments qui ne contiennent pas assez d'informations spatiales discriminantes.
- Une valeur trop importante de T augmente considérablement la qualité d'apprentissage (besoin du « data augmentation »).

Nous proposons donc d'entraîner le modèle VGG16 et extraire des caractéristiques sur les séquences obtenues en découpant les séquences vidéo entières en frames en noir et blanc contenant des objets qui représente les joueurs. La nature de ces frames générés en noir et blanc nécessite donc la mise en place de stratégies de classification basée sur le modèle VGG16 destiné à ce type d'image.

La couche de sortie contient 8 neurones qui représentent les différentes tactiques de la joue, et les couches cachées se répartissent en quatre catégories :

4.2.1.1.Les couches de Convolution

L'architecture proposée comprend 13 couches de convolution laquelle :

- **La première couche** convolutionnelle **(conv1)** est destiné à l'extraction d'identité (feature extraction).il effectue une convolution avec 64 filtres de taille 3x3 pour filtrer une image d'entrée de taille 227×227×1. Cette couche est effectuée avec un padding same, la fonction d'activation utilisée est RELU. Le résultat de cette couche est une Channel de taille 227x227x64.

- **La deuxième couche** est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 227×227×64

- **La troisième couche** occupe la quatrième position du modèle avec 128 filtres de taille 3x3, et sans utiliser le padding avec l'application d'une fonction d'activation RELU. Le résultat de cette couche est une Channel de 113x113x128.

- **La quatrième couche** occupe la cinquième position est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 113×113×128.

- **La cinquième couche** occupe la septième position de modèle avec 256 filtres de taille 3x3, un padding same, la fonction d'activation utilisé est RELU. Le résultat de cette couche est une Channel de taille 56x56x256.
- **La sixième couche** occupe la huitième position de modèle est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 56×56×256.
- **La septième couche** est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 56×56×256.
- **La huitième couche** occupe la onzième position du modèle avec 512 filtres de taille 3x3, et sans utiliser le padding avec l'application d'une fonction d'activation RELU. Le résultat de cette couche est une Channel de 28x28x512.

- **La neuvième couche** vient directement après la couche de convolution précédente, est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 28×28×512.

- **La dixième couche** est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 28×28×512.

- **La onzième couche** occupe la quinzième position du modèle avec 512 filtres de taille 3x3, et sans utiliser le padding avec l'application d'une fonction d'activation RELU. Le résultat de cette couche est une Channel de 14x14x512.

- **La douzième couche** occupe la quinzième position du modèle avec 512 filtres de taille 3x3, et sans utiliser le padding avec l'application d'une fonction d'activation RELU. Le résultat de cette couche est une Channel de 14x14x512.
- **La treizième couche** est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 14×14×512.

- **La quatorzième couche** vient directement après la couche de convolution précédente, est contient les mêmes paramètres que la couche de convolution précédente et donne comme résultat un Channel de 14×14×512.

4.2.1.2.Les couches de pooling

L'architecture composée par 5 couches de pooling (maxPooling) pour réduire la dimension spatiale (feature map) qui est dérivée de la couche précédente (couche convolution)

- **La première couche** de maxpooling (**Pool 1**) occupe la troisième position dans le modèle, elle se caractérise par un pooling de taille 2x2 et un stride (2,2) et elle donne comme résultat une Channel de taille 227x227x64.

- **La deuxième couche** de pooling **(pool 2**) vient après la quatrième couche de convolution et elle dispose de même paramètres que la couche de pooling précédente. Le résultat de cette couche est une Channel de 113 x 113 x128.

- **La troisième couche** de pooling **(pool 2**) occupe la dixième position dans le modèle. elle dispose de même paramètre que la couche de pooling précédente. Le résultat de cette couche est une Channel de 56 x 56 x256.

- **La quatrième couche** de pooling **(pool 2**) vient après la dixième couche de convolution et elle dispose de même paramètres que la couche de pooling précédente. Le résultat de cette couche est une Channel de 14 x 14 x512.

- **La cinquième couche** de maxpooling (**pool 3**) précède la couche fully connected et il dispose de même caractéristique que les autres couches. Le résultat de cette couche est une Channel de taille 7 x 7 x 512

4.2.1.3.Les couches de fully connected

A la fin, notre modèle contient une seul couche entièrement connectées (FC) avec 360 unités. En choisissant ce nombre des unités pour minimisé le nombre des paramètres et accélérer le calcule (le taux de calcules). Chaque couche utilisée une fonction d'activation RELU.

Dans la couche de sorti, en utilisent dans notre modèle Alex-net une couche **Softmax** à la fin avec 8 unités (output class) qui présente le nombre de nos classes.

Le tableau suivant résume notre approche de classification :

Tableau 3.1-La structure du réseau neuronal convolutif VGG16

Layers	**Filtres**	**Kernel size**	**Stride**	**Padding**	**Fonction d'activation**
Image	227×227×1				
Conv2D	227×227×64	3×3	1×1	Same	Relu
Conv2D	227×227×64	3×3	1×1	Same	Relu
MaxPool	113×113×128	2×2	2×2	Same	Relu
Conv2D	113×113×128	3×3	1×1	Same	Relu
Conv2D	113×113×128	3×3	1×1	Same	Relu
MaxPool	56×56×256	2×2	2×2	Same	Relu
Conv2D	56×56×256	3×3	1×1	Same	Relu
Conv2D	56×56×256	3×3	1×1	Same	Relu
Conv2D	56×56×256	3×3	1×1	Same	Relu
MaxPool	28×28×512	2×2	2×2	Same	Relu
Conv2D	28×28×512	3×3	1×1	Same	Relu
Conv2D	28×28×512	3×3	1×1	Same	Relu
Conv2D	28×28×512	3×3	1×1	Same	Relu
MaxPool	14×14×512	2×2	2×2	Same	Relu
Conv2D	14×14×512	3×3	1×1	Same	Relu
Conv2D	14×14×512	3×3	1×1	Same	Relu
Conv2D	14×14×512	3×3	1×1	Same	Relu
MaxPool	7×7×512	2×2	2×2	Same	Relu
Fully Connected	360				Relu
Fully Connected	8				Softmax

4.2.2. Apprentissage

Ce modèle de classification basé sur le modèle (VGG16), Incarné avec 9000000 paramètres entraînables ; et un nombre des échantillons de 1441 avant l'augmentation de notre dataset. L'augmentation des données par la technique de data-augmentation nous a permet d'améliorer la précision de notre modèle et de réduire le overfiting**.** Les fonctions d'augmentation utilisées sont les même que le modèle précédent.

Après l'augmentation nous avons obtenu 4000 échantillons dans notre jeu d'entrainement (train data), la chose qui nous a permet d'arriver à un taux de précision de 99.87%.

Les caractéristiques utilisées pour l'entraînement de notre modèle sont :

4.2.2.1.La compilation de modèle

Nous avons utilisé les paramètres suivants pour la phase de compilation:

- **Compile ()** : Compile définit la fonction de perte, l'optimiseur et les métriques
- **Optimiseur Adam ()** : nous allons utiliser l'algorithme d'optimisation Adam pour mettre à jour les poids de réseau itératifs qui basés sur les données d'entraînement.
- **loss='categorical_crossentropy'() :** en utilisant cette fonction de perte pour la catégorisation d'une étiquette.
- **Metrics = [‘accuracy’]** : nous utilisons cette fonction pour évaluer les performances de ce modèle.

4.2.2.2.L'entraînement de modèle

Dans l'entraînement de modèle nous utilisons les paramètres suivant :

- **fit_generator ()** : utilisé pour former nos modèles d'apprentissage en profondeur (Alexnet). Il nécessite un générateur pour les données de l'entraînement (training data)

- **Batch size: (32):** la taille de lot (batch size) qui utilisé dans notre modèle Alex-net est 32 parce qu'est un bon point de départ.

- **Step_per_epoch** () : le nombre total d'étapes (lots d'échantillons) qui utilisé dans notre approche est de 4000 échantillons.

- **Epoch ():** en a utilisé les époques pour séparer la formation en phases distinctes, ce qui est utile pour la journalisation et l'évaluation périodique. Nous utilisons respectivement 50,100,150 nombre d'époque pour obtenir accuracy parfait pour notre modèle ;mais, en a choisir 50 époche par ce qu'il après l'epoche 50 la précision de l'apprentissage stabilisé.

4.2.3. Le test :

Pour évaluer l'approche, nous avons réalisé les expériences sur un ensemble de données dans une variété de catégories des tactiques de jeu dans notre dataset. Les jeux de test (test data) composés de 321 images respectivement. Ce scenario nous a donné un accuracy de 61 % dans l'évaluation du modèle. Ce taux réduit nous a ramené à appliquer les mêmes techniques d'augmentation des données utilisé dans la phase d'apprentissage. Cette augmentation nous a permis d'obtenir un jeu de 642 d'image pour arriver à un accuracy de 99,53 %.

4.2.4. Interprétation des Résultats du modèle proposé VGG16

La précision de l'apprentissage et de test augmente avec le nombre d'époque, ceci reflète qu'à chaque époque le modèle apprenne plus d'informations. Comme en a vue dans la figue précédemment.

La FIGURE 3.21 représente des changements dans La précision de l'apprentissage où:

La précision de la formation (training accuracy):

- De 0 époque à 50 époques, la précision de l'entraînement augmente jusqu'à la valeur 99 % à l'époque 50.
- La précision d'entraînement maximale (training accuracy) est de 99,87%.

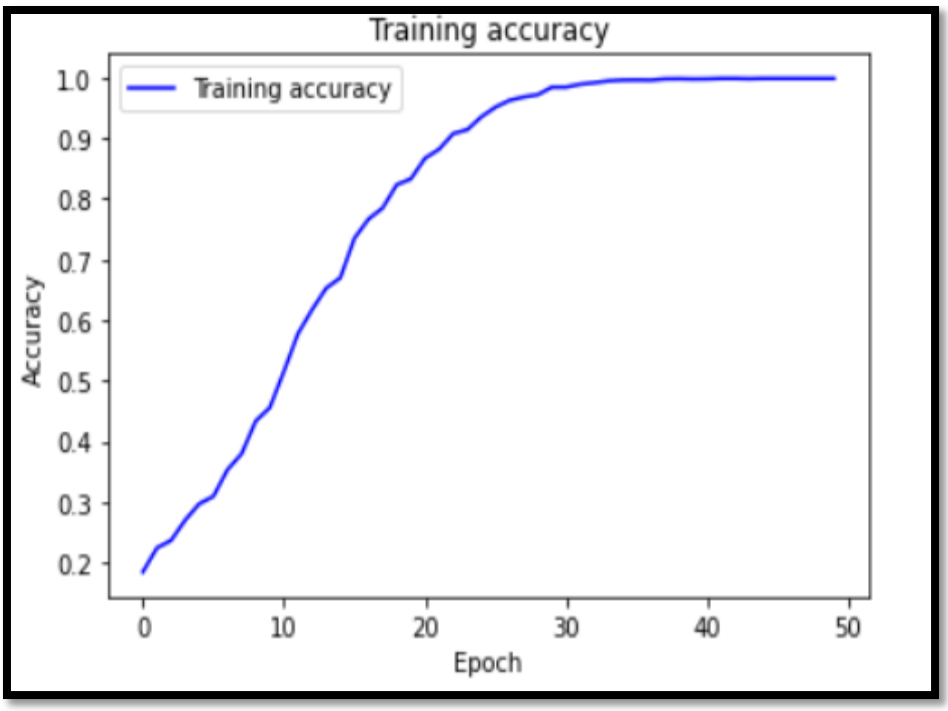

FIGURE 3.21-La précision de l'entraînement « VGG16 »

La FIGURE 3.22 représente des changements dans L'erreur de l'apprentissage où:

La perte de l'apprentissage (Loss accuracy):

- De 0 époque à 50 époques, la perte d'entraînement (loss accuracy) diminue de la valeur 2.0 jusqu'à la valeur proche de 0,0219 à l'époque 50.
- La perte d'entraînement minimale est de 0,0219.

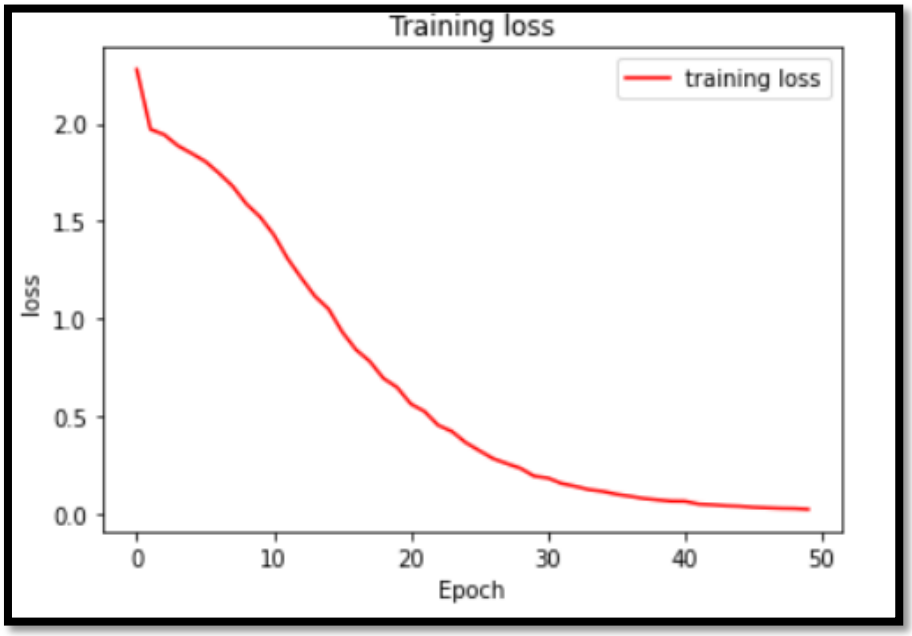

FIGURE 3.22-La perte de l'entraînement « VGG16 »

Et en comparant la précision et la perte d'apprentissage dans la même figure nous remarquons de la stabilisation de la précision et la perte sont similaire.

FIGURE 3.23- La précision et la perte de l'entraînement « VGG16 »

5. Validation des modèles

Pour donner plus de concret à notre modèle, on a opté pour une phase de validation, pour ce faire on a choisi une vidéo d'une mi-temps d'un match de football de la première League anglaise entre Manchester City et Manchester United.

La première chose à faire est de déclarer la couleur de l'équipe cible (couleur bleu de Manchester city) afin d'extraire les frames qui représentent l'enlacement des joueurs puis on passe à la deuxième phase de classifications que ce soit Alexnet ou VGGNET16 dans le but de donner la tactique appliquée par l'équipe de « Citizens ».

Les résultats obtenus par nos deux modèles sont similaire. La tactique de jeu générer pour l'équipe de Manchester City est le **4-5-1.** Cette résultat est confirmé par les experts, ces derniers ont été d'accord sur le tactique **4-2-3-1** qui est la forme la plus explicatifs du tactique **4-5-1.**

Le même travail est refait avec l'équipe de Manchester City, aussi les deux modèles ont donné les mêmes résultats 4-4-2. Toujours selon les experts la tactique appliquée par cette équipe 4-4-2 en losange (4-1-2-1-2).

La figure ci-dessous montre les différentes étapes de validation :

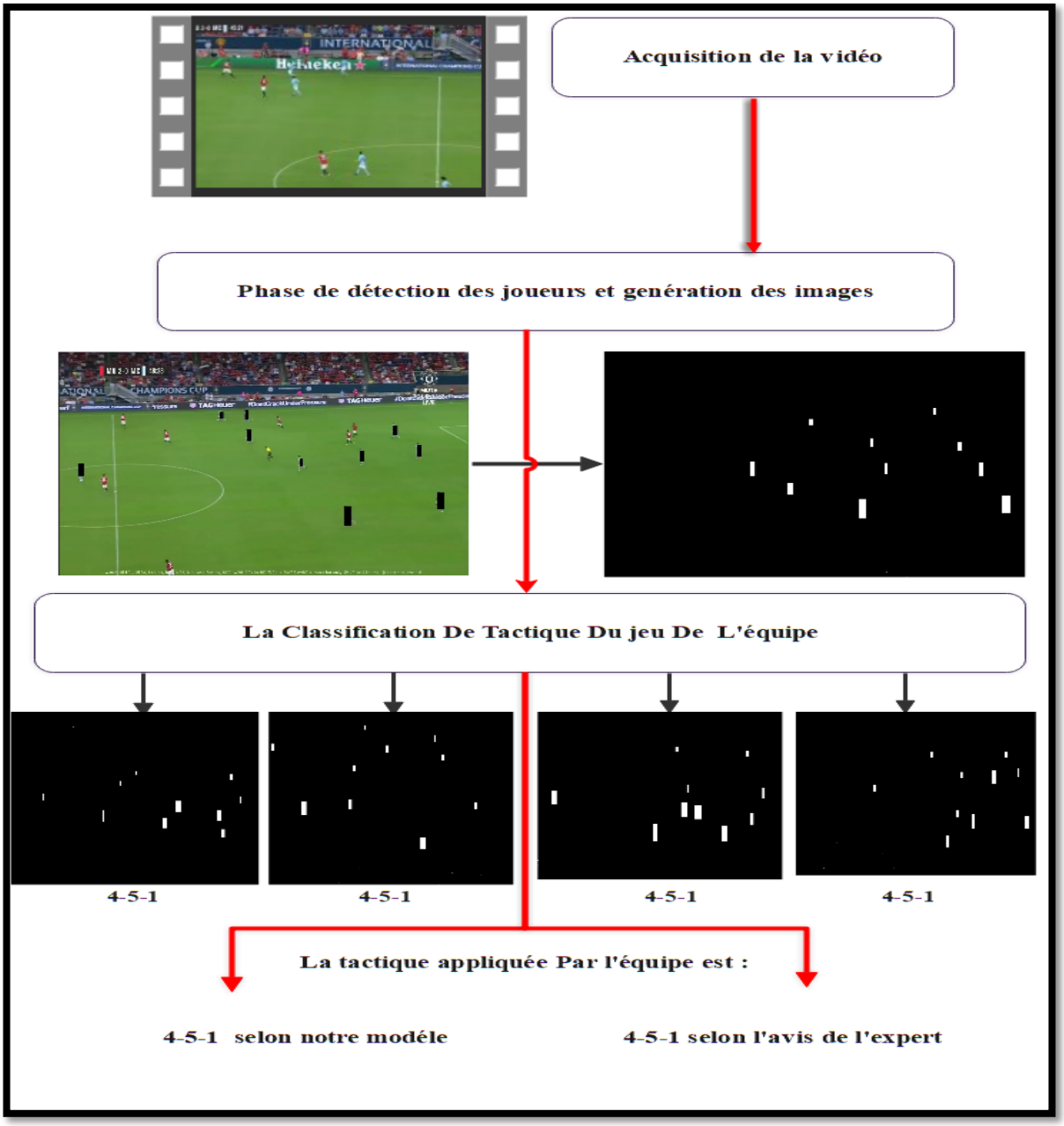

FIGURE 3.24- Validation des modèles de L'équipe choisie « Manchester City

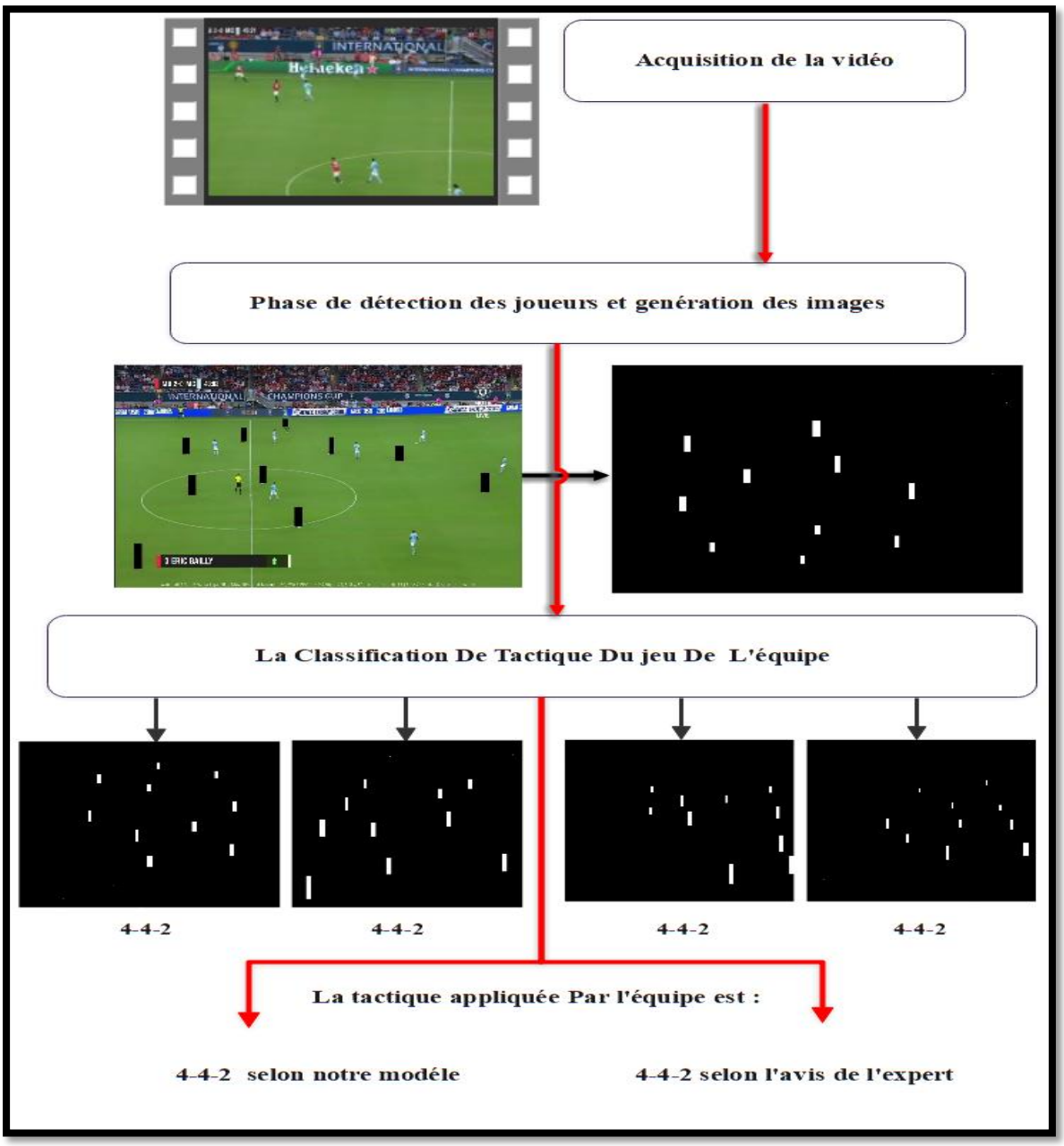

FIGURE 3.25- Validation des modèles de L'équipe choisi « Manchester United »

Conclusion

Dans ce chapitre, nous avons présenté la réalisation de nos approches dédiées à la classification des vidéos qui s'appuient sur des réseaux de convolution avec les résultats obtenus. Dans ce cadre, nous avons étudié les performances de nos modèles (Alexnet et VGG16) en classification vidéo à grande échelle. Nous avons constaté que ces architectures CNN sont capables d'apprendre des fonctionnalités puissantes à partir de données faiblement étiquetées qui dépassent de loin les méthodes basées sur les performances et que ces avantages sont étonnamment robustes aux détails de la connectivité des architectures par temps.

Le but de ces approches est de détecter les joueurs dans un équipe choisissez et classifier les différents tactique de jeu de cette équipe au cours de la vidéo du match de foot en temps réel

Conclusion Générale

Ces dernières années, avec l'explosion technologique, l'art du traitement vidéo s'est considérablement développé, principalement avec l'apparition de l'apprentissage profond, notamment dans les tâches de détection et de classification.

Dans notre travail, nous avons utilisé des approches de détection et de classification afin d'aider les entraîneurs de football dans leur jeu, que ce soit pour leur équipe ou pour les équipes adverses. L'objectif principal de nos travaux est de parvenir à une approche de classification multi-classes basée sur des réseaux de neurones convolutifs par apprentissage supervisé après une étape de détection de la localisation des acteurs sur le terrain.

Les tâches effectuées au cours de ce travail sont divisées en grandes lignes, en premier lieu nous avons généré le jeu de données à partir d'une approche de détection, ou nous avons extraire la vidéo en image et traité par détection l'objet (joueurs) selon une couleur choisie au départ, pour éviter de détecter les objets de couleur choisis hors champ, opter pour tester le conteur pour chaque objet détecté. Le conteur doit être vert, qui est la couleur du terrain. La condition de sélection du cadre est le nombre d'objets qui doit être égal à 10, ce qui représente le nombre de joueurs dans le terrain. Les images sélectionnées seront enregistrées avant d'être classées avec l'aide d'un expert dans le domaine.

La deuxième étape de notre travail est de proposer un modèle de classification basé sur l'apprentissage profond. Pour ce faire, nous avons opté pour un modèle basé sur AlexNet et un autre modèle basé sur VGG où nous sommes arrivés à un taux de précision de 50% et 56% respectivement un ensemble de données de 1441 images. Ces taux sont améliorés de manière très significative après l'augmentation de l'ensemble de données par une inversion verticale et horizontale. Les taux obtenus après cette augmentation sont de 96 % et 99.50 %.

Notre modèle final prendra une vidéo d'un match de football et donnera les tactiques de jeu à appliquer par l'équipe sélectionnée par sa couleur. Le point important de notre modèle et la vue multiple lors de la phase de détection où le modèle peut fonctionner avec plusieurs caméras depuis plusieurs positions.

Les perspectives de notre travail sont de réaliser un modèle qui analyse les vidéos des matchs de football afin de donner un aperçu de l'animation défensive - offensive et inversement. Ce travail donnera une grande flexibilité aux entraîneurs de football pour mieux lire et préparer leur équipe ou pour analyser également le jeu d'une équipe adverse.

Bibliographie

[1]*Christophe Savariaux, Esposito, A., Bratanić, M., & Keller, E. (Eds.). (2007). Fundamentals of verbal and nonverbal communication and the biometric issue (Vol. 18). IOS press.*

[2]*Vinod Kumari Sharma, RK Bhatnagar & Dipti Arora, Video Digitization and Editing: An Overview of the Process DESIDOC Bulletin of Information Technology, Vol. 22, No. 4 & 5, July & September 2002, pp. 3-8.*

[3]*Djindjian, F. (2016). Archéologie, de l'analogique au numérique: évolution technique ou révolution méthodologique?. Les nouvelles de l'archéologie, (146), 6-11.*

[4]*C. V. networking Index. Forecast and methodology, 2016- 2021, white paper. San Jose, CA, USA, 1, 2016.*

[5] *Lu, G., Ouyang, W., Xu, D., Zhang, X., Cai, C., & Gao, Z. (2019). Dvc: An end-to-end deep video compression framework. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 11006-11015).*

[6]*C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Kr¨ahenb¨uhl. Compressed video action recognition. In CVPR, pages 6026–6035, 2018.*

[7]*BARBOSA Philippe, GIDEL Marie-aude, DALL'AGNESE ,(2017), la vidéo analogique et numérique : Thomas http://patrick.dallagnese.free.fr/tpe/.*

[8]*Introduction à la vidéo numérique* .URL: *https://www.commentcamarche.net/contents/1493-introduction-a-la-video-numerique*

[9]*Van Der Schaar, M., Turaga, D. S., & Wong, R. (2006). Classification-based system for cross-layer optimized wireless video transmission. IEEE Transactions on Multimedia, 8(5), 1082-1095*.

[10] *Mhalla12, A., Chateau, T., Delsol, A., Gazzah, S., & Amara, N. E. B. Détection spatio-temporelle d'objets par apprentissage profond: l'entrelacement vidéo pour améliorer le suivi multi-objets.*

[11] *Kamble, P. R., Keskar, A. G., & Bhurchandi, K. M. (2019). A deep learning ball tracking system in soccer videos. Opto-Electronics Review, 27(1), 58-69.*

[12] *Guide,Javier Rey ,Object Detection with Deep Learning (2017) for the detection of acute intracranial haemorrhage from small datasets. Nature Biomedical Engineering, 3(3), 173*

[13] *Kislay Keshari Mar.12.19 ,Object Detection Tutorial in TensorFlow : Real-Time Object Detection,.*

[14] *Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4694-4702).*

[15] *Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4489- 4497)*

[16] *Taibaoui, M., & Debbar, D. La découverte des concepts sémantiques cachés avec plusieurs niveaux d'abstraction pour la recherche d'images (Doctoral dissertation).*

[17] *Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machine By Lean Yu, Shouyang Wang, Kin Keung Lai, Ligang Zhou.*

[18] *Belalia, M. (2016). Classification des images selon la sémantique, mémoire master , université de Mostaganem*

[19] *Deep Learning Long Short-Term Memory (LSTM) Networks: What You Should Remember : ttps://missinglink.ai/guides/neural-network-concepts/deep-learning-long-short-term-memory-lstm-networks-remember/*

[20] *Youcef Messaoud, (Nov 22,2018), LSTM, Intelligence artificielle sur des données chronologiques.*

[21] *Kawakami, K. (2008). Supervised sequence labelling with recurrent neural networks. Ph. D. dissertation, PhD thesis.*

[22] *Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici, "Beyond short snippets: Deep networks for video classification," arXiv preprint arXiv:1503.08909, 2015.*

[23] *Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625– 2634.*

[24] *McLaughlin, N., Martinez del Rincon, J., & Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1325-1334).*

[25] *Xu, Z., Hu, J., & Deng, W. (2016, July). Recurrent convolutional neural network for video classification. In 2016 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.*

[26] *Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.*

[27] *Li, X., & Wang, S. (2017). Object detection using convolutional neural networks in a coarse-to-fine manner. IEEE Geoscience and Remote Sensing Letters, 14(11), 2037-2041.*

[28] *de Menezes, R. S. T., Magalhaes, R. M., & Maia, H. (2019). Object Recognition Using Convolutional Neural Networks. In Artificial Neural Networks. IntechOpen.*

[29] *Akçay, S., Kundegorski, M. E., Devereux, M., & Breckon, T. P. (2016, September). Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 1057-1061). IEEE.*

[30] *Tompson, J., Stein, M., Lecun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics (ToG), 33(5), 1-10.*

[31] *Maungmai, W., & Nuthong, C. (2019, February). Vehicle Classification with Deep Learning. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 294-298). IEEE.*

[32] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

[33] *Wei, H., Laszewski, M., & Kehtarnavaz, N. (2018, November). Deep learning-based person detection and classification for far field video surveillance. In 2018 IEEE 13th Dallas Circuits and Systems Conference (DCAS) (pp. 1-4). IEEE.*

[34] *A Deep Learning Approach for Vehicle Detection, 2018 13th International Conference on Computer Engineering and Systems (ICCES), [Mohamed Ashraf Ali[1], Hossam E. Abd El Munim[2], Ahmed Hassan Yousef [2,3], and Sherif Hammad [1]].*

[35] *Minhas, R. A., Javed, A., Irtaza, A., Mahmood, M. T., & Joo, Y. B. (2019). Shot classification of field sports videos using alexnet convolutional neural network. Applied Sciences, 9(3), 483..*

[36] *Buduma, N., & Locascio, N. (2017). Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. " O'Reilly Media, Inc.".*

[37] *Soentanto, P. N., Hendryli, J., & Herwindiati, D. E. (2019, July). Object and Human Action Recognition From Video Using Deep Learning Models. In 2019 IEEE International Conference on Signals and Systems (ICSigSys) (pp. 45-49). IEEE.*

[38] *Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4694-4702).*

[39] *Russo, M. A., Filonenko, A., & Jo, K. H. (2018, September). Sports Classification in Sequential Frames Using CNN and RNN. In 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT) (pp. 1-3). IEEE.*

[40] *Hakim, N. L., Shih, T. K., Arachchi, K., Priyanwada, S., Aditya, W., Chen, Y. C., & Lin, C. Y. (2019). Dynamic Hand Gesture Recognition Using 3DCNN and LSTM with FSM Context-Aware Model. Sensors, 19(24), 5429.*

[41] *Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).*

[42] *Gao, L., Guo, Z., Zhang, H., Xu, X., & Shen, H. T. (2017). Video captioning with attention-based LSTM and semantic consistency. IEEE Transactions on Multimedia, 19(9), 2045-2055.*

[43] *Pan, Y., Yao, T., Li, H., & Mei, T. (2017). Video captioning with transferred semantic attributes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6504-6512).*

[44] *Baraldi, L., Grana, C., & Cucchiara, R. (2017). Hierarchical boundary-aware neural encoder for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1657-1666).*

[45] *Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).*

[46] *Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).*

[47] *Buduma, N., & Locascio, N. (2017). Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. " O'Reilly Media, Inc.".*

[48] *Digital communication Sampling –digital communication.URL : https://www.wisdomjobs.com/e-university/digital-communication-tutorial-1983/digital-communication-sampling-25953.html*

[49] *Frich, A. (2017). Color Management Guide. URL: http://www. Color-management guide. Com/how-to-calibrate-monitor. html.(9.9. 2015.).*

[50] *David E. Fisher, A. Michael Noll,January 31, 2020,Encyclopædia Britannica,: https://www.britannica.com/technology/television-technology*

[51] *Minsky, M. (2007). The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind. Simon and Schuster.*

[52] *Srishti Salwa, (2015),K-Nearest Neighbors is one of the most basic algorithm used for Classification.*

[53] *Write short note on decision tree based classification approach. URL: https://techblogmu.blogspot.com/2018/05/decision-tree-based-classification-approach.html*

[54] *Kshitiz Rima.,(2014,),difference-machine-learning-deep-learning*

[55] *Keller, F. Rapport d'information fait au nom de la délégation sénatoriale à la prospective sur les nouvelles menaces des maladies infectieuses émergentes par Mme. Fabienne Keller, Sénatrice.*

[56] *Stuner, B. (2018). Cohorte de réseaux de neurones récurrents pour la reconnaissance de l'écriture (Doctoral dissertation, Normandie).*

[57] *Angel, A. (2015). Towards Distortion-Predictable Embedding of Neural Networks. arXiv preprint arXiv:1508.00102.*

[58] *Ye, Y., Zhang, C., He, C., Wang, X., Huang, J., & Deng, J. (2020). A Review on Applications of Capacitive Displacement Sensing for Capacitive Proximity Sensor. IEEE Access, 8, 45325-45342.*

[59] *Understanding lstm networks. http://colah.github.io/posts/2015-08- Understanding-LSTMs/. Accessed April 6, 2019.*

[60] *Morice-Atkinson, X. (2019). Machine learning approaches for astrophysics and cosmology (Doctoral dissertation, University of Portsmouth).*

[61] *Prabhu, R. (2018). Understanding of Convolutional Neural Network (CNN)–Deep Learning. A Medium Corporation, US.*

[62] *Konjuh, N. (2019). Automatsko raspoznavanje akcija u rukometu (Doctoral dissertation, University of Rijeka. Department of Informatics.).*

[63] *Panadda konslip , (2020), CNN step4 connection. URL : https ://medium.com/ @ PK_KwanG/,cnn-step-1-convolution-d53a0803b255*

[64] Bekraoui, N., Cazorla, G., & Léger, L. (2010). Les systèmes d'enregistrement et d'analyse quantitatifs dans le football. *Science & Sports*, *25*(4), 177-187.

[65] *Bekraoui, N., Cazorla, G., Léger, L. (2008). Validité et limite de la technique du gps dans l'analyse de la tâche en football. In: B. Zoudji, Editor,* Science et football : recherches et connaissances actuelles. Valenciennes*, Presses Universitaires de Valenciennes. p.363-377.*