REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

الجمهورية الجزائرية الديمقراطية الشعبية

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

HIGHER SCHOOL IN APPLIED SCIENCES
--T L E M C E N--

المدرسة العليا في العلوم التطبيقية
École Supérieure en Sciences Appliquées

وزارة التعليم العالي والبحث العلمي

المدرسة العليا في العلوم التطبيقية
-تلمسان-

Mémoire de fin d'étude

Pour l'obtention du diplôme de master

Filière : automatique
Spécialité : automatique

Présenté par : Melle TAMAZIRT Melyssa
Melle NAGHRAOUI Djihane

Thème

Benchmark pour la reconnaissance automatique des émotions.

Soutenu publiquement le 27/10/ 2020 devant le jury composé de :

M Fouad BOUKLI	MCA	ESSA Tlemcen	Président
Mme Wahida HANDOUZI	MCB	Université Abou Bekr Belkaid	Directrice du mémoire
M Ali RIMOUCHE	MCB	ESSA Tlemcen	Co-directeur du mémoire
Mme Latéfa GHOMRI	MCA	Université Abou Bekr Belkaid	Examinatrice
M Ghouti ABDELLAOUI	MCB	ESSA Tlemcen	Examinateur

Année universitaire : 2019/2020

Dédicace

« A nos très chères mères »

Quoi que nous fassions ou que nous disions, nous ne saurons pas vous remercier comme il se doit. Votre affection nous couvre, votre bienveillance nous guide et votre présence à nos côtés a toujours été notre source de force pour affronter les différents obstacles de la vie**.**

« A nos très chers pères »

Nous dédions ce travail marquant de nos vies à la mémoire du père disparu trop tôt. Nous espérons que du monde qui est sien maintenant, qu'il apprécie cet humble geste comme preuve de reconnaissance de la part d'une fille qui a toujours prié pour le salut de son âme. Puisse Dieu, le tout puissant, l'avoir en sa sainte miséricorde.

A mon père qui était avec nous depuis le début de ce travail, pour le goût à l'effort qu'il a suscité en nous, de par sa rigueur. Ceci est nos profondes gratitudes pour ton éternel amour, que ce rapport ne soit qu'un début des cadeaux que je puisse t'offrir.

A nos frères, nos sœurs, nos grands-parents et tous ceux qui ont partagé avec nous tous les moments d'émotion lors de la réalisation de ce travail. Ils nous ont chaleureusement supportés et encouragés tout au long de notre parcours. A nos familles, nos proches et à ceux qui nous donnent de l'amour et de la vivacité.

Remerciements

On tient tout d'abord à remercier le dieu tout puissant pour nous avoir donné le courage, la patience, la force ainsi que la santé pour nous permettre de finir ce travail.

On remercie également notre encadrante Mme Wahida HANDOUZI et notre co-encadrant M Ali RIMOUCHE de nous avoir si bien accompagné et conseiller, leurs encouragements et leur patience nous ont été d'une aide psychologique précieuse et spécialement Mme W.HANDOUZI, qui a fait preuve d'énormément de professionnalisme, de bienveillance et de gentillesse à notre égard.

Nos remerciements s'adressent aussi aux membres du jury : M F.BOUKLI, Mme L.GHOMRI et M G.ABDELLAOUI qui nous ont fait l'honneur de lire et d'évaluer ce mémoire. En outre, nous tenons à remercier notre école, que ce soit le staff administratif ou encore nos enseignants, ils ont toujours été à notre écoute et nous ont toujours bien accompagnées et ce, durant tout notre cursus.

Faire un mémoire de fin d'étude est pour nous l'accomplissement de cinq années de travail et de sacrifices. Même si c'est un projet dont la durée ne s'étale que sur un semestre, on en a énormément appris, mais on s'est aussi heurté à des moments plus difficiles. Pour surmonter ces difficultés, le soutien de personnes qui vous entour est déterminant. On a eu le privilège de bénéficier de cet appui, et c'est avec gratitude que nous nous adressons à nos familles respectives qui ont toujours cru en nous et nous ont toujours tirés vers le haut. Sans eux, nous n'en serions pas là. Spécial merci à nos amis respectifs, Oussama BORSALI et Leila BENAZZA qui nous ont encouragées et ce, depuis le début.

Table des matières

Liste des abréviations

Acc : Précision.

CNN/ConvNet : Réseau de neurones convolutifs.

Conv : Convolution.

DB : Base de données.

DL : Apprentissage profond.

FC : Entièrement connectées.

ML : Apprentissage automatique.

RN : Réseau de neurones.

Val : Taux/valeur.

Liste des tableaux

Liste des figures

Introduction générale

L'émotion est présente dans les meilleurs comme dans les pires moments d'une personne, cela étant dit, la société moderne semble sous l'emprise de ces cotés les plus sombres. En effet, le nombre de troubles anxieux et de troubles de l'humeur ne cessent de croitre et cela semble se ressentir surtout dans le domaine de travail. Aujourd'hui, la première cause d'absentéisme pour raison médicale revient directement, non plus à des raisons médicales comme c'était le cas avant, mais bien à des problèmes liés à la régulation de l'émotion (SCHENE et Van DIJK, Vander KLINK, BLONK). D'un autre coté, l'essor que connait la robotique nécessite des technologies capables d'encoder et de decoder l'émotion, chose qui ne peut se faire si les caractéristiques visibles ou non de l'émotion ne sont pas étudiées. En outre, les différents troubles psychologiques sont intrinsèquement liés aux émotions, que se soit pour les sujets phobiques dont ce trouble résulte de la coloration subjective d'une expérience par l'émotion, ou tout simplement pour les personnes Lambda victimes d'anxiété. Que se soit pour les domaines techniques ou pour les domaines relevant de la psychologie, l'émotion tient aujourd'hui une place centrale.

La reconnaissance des émotions peut se faire de plusieurs manières: Signaux vocaux, expressions faciales, signaux physiologiques, et expressions corporelles. En 1968, Albert MEHRABIAN a souligné le fait que l'interaction entre les humains est constituée de 7% par des indices verbaux, 38% par les signaux vocaux et la majeure partie, soit 55% est apportée par des expressions faciales. Ainsi, l'expression faciale est l'une des composantes les plus importantes pour la reconnaissance des émotions. L'objectif de ce travail est de mettre en place un benchmark pour la reconnaissance automatique des émotions. Il permettra de reconnaitre les six émotions de bases décrites selon EKMAN, à savoir, la colère, le dégout, la joie, la tristesse, la peur ainsi que la surprise, à cela, nous ajouterons la neutralité.

L'apprentissage profond, plus précisément les réseaux de neurones convolutifs, ont permis de résoudre les problèmes rencontrés par le machine Learning et la reconnaissance d'images en générale, à savoir, le très grand nombre de paramètres engendrés par la taille des images. En outre, la disponibilité de très grandes bases de données étiquetées permettant l'entrainement de ces réseaux rendent leur utilisation beaucoup plus aisée. Dans ce travail, quatre architectures convolutives inspirées de LeNet et VGG16 seront testées avec trois bases de données (la Fer2013, la JAFFE et la RaFD), différents optimiseurs (Adam, SGD, Adagrad et Rmsprop), et avec un Batch size différent (60, 120 et 240).

Ce mémoire est divisé en quatre chapitres, à savoir :

- Un premier chapitre purement théorique consacré à l'expression de l'émotion via ses différentes composantes faciales, posturales, physiologiques et vocales.

- Un deuxième chapitre faisant état des différentes méthodes de reconnaissance automatique des émotions via les expressions faciales.

- Un troisième chapitre dédié aux réseaux de neurones convolutifs et dans lequel le fonctionnement de ce type d'architectures est décortiqué.

- Enfin, un quatrième chapitre dédié à la méthodologie, aux résultats et aux discussions.

1

Notions générales sur les émotions

1.1 Introduction

Dans ce chapitre, nous allons mettre en exergue les différentes notions concernant les émotions, en passant par leur définition, les différentes façons de les représenter ou encor, la divergence des théories traitant du sujet.

1.2 Définition

Etant un phénomène complexe, l'émotion ne possède pas de définition claire ni unique [1]. Les émotions sont des états motivationnels pouvant induire l'individu à changer ses relations intérieures soient-elles ou extérieures, kit à être contradictoire avec la réalité comme décider de maintenir une relation qui lui fait du mal. C'est donc une expérience psychophysiologique intense et complexe de l'esprit d'un individu en réponse aux influences externes (environnementales), ou internes (biochimiques) [2].

En mettant en perspective les quatre théories de l'émotion subsistantes (voir figure 1-1), on obtient la définition suivante : Les émotions sont des états affectifs (relatif à l'affectivité, qui est un ensemble de phénomènes psychiques qui influencent à la fois l'état propre de l'esprit, la vision du monde, l'attitude, le comportement, et la pensée), suivis de réactions physiologiques (transpiration, tremblements, l'accélération du rythme cardiaque ou encor le faite de rougir) (Janet 1926, James 1884), qui sont engendrés par des stimulis externes ou internes, et qui sont de courte durée en général. Elles sont repérables grâce à l'expression du visage (EKMAN 1994), le ton de la voix et les gestes (Scherer 1986) et par les signaux verbaux (RIME, CORSINI, & Herbette, 2002). Elles peuvent être de faible ou forte intensité et ont une valence soit positive (agréable) ou négative (désagréable) (Russel 1999, FELDMAN 1995). On peut notamment l'évaluer de par ses conséquences négatives ou positive sur notre paix intérieure (bien-être) (SCHERER, SCHOR et JOHNSTONE 2001, LAZARUS 1999et FRIDJA 1986) [3].

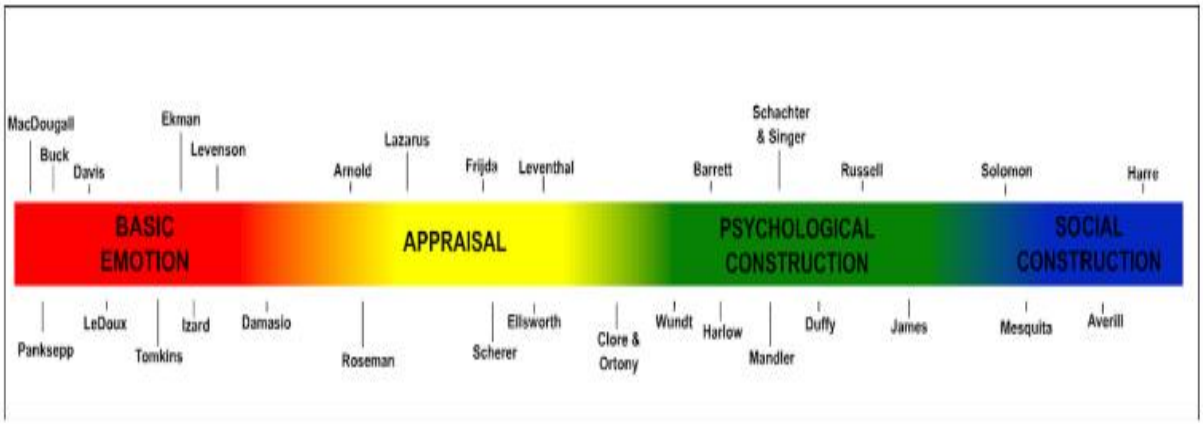

Figure 1. Perspectives on emotion can be loosely arranged along a continuum. We have populated this continuum with representative theorists/researchers drawn from the field of psychology. We distinguish four "zones": (1) basic emotion, in red, e.g., MacDougall (1908/1921), Panksepp (1998), Buck (1999), Davis (1992), LeDoux (2000), Tomkins (1962, 1963), Ekman (1972), Izard (1993), Levenson (1994), and Damasio (1999); (2) appraisal, in yellow, e.g., Arnold (1960a, 1960b), Roseman (1991), Lazarus (1991), Frijda (1986), Scherer (1984), Smith and Ellsworth (1985), Leventhal (1984), and Clore and Ortony (2008); (3) psychological construction, in green, e.g., Wundt (1897/1998), Barrett (2009), Harlow and Stagner (1933), Mandler (1975), Schachter and Singer (1962), Duffy (1941); Russell (2003), and James (1884); (4) social construction, in blue, e.g., Solomon (2003), Mesquita (2010), Averill (1980), and Harré (1986). Given space constraints, as well as the goals of this article, we have limited ourselves to a subset of the many theorists/researchers who might have been included on this continuum (e.g., those who only study one aspect of emotion were not included in this figure).

Figure 1-1 : Continuum des théories de l'émotion [4].

1.3 Composantes de l'émotion

L'émotion induit deux types de changements chez l'individu (voir figure 1-2) et peut être mesuré via trois composantes : la composante physiologique, la composante cognitive subjective et la composante expressive comportementale.

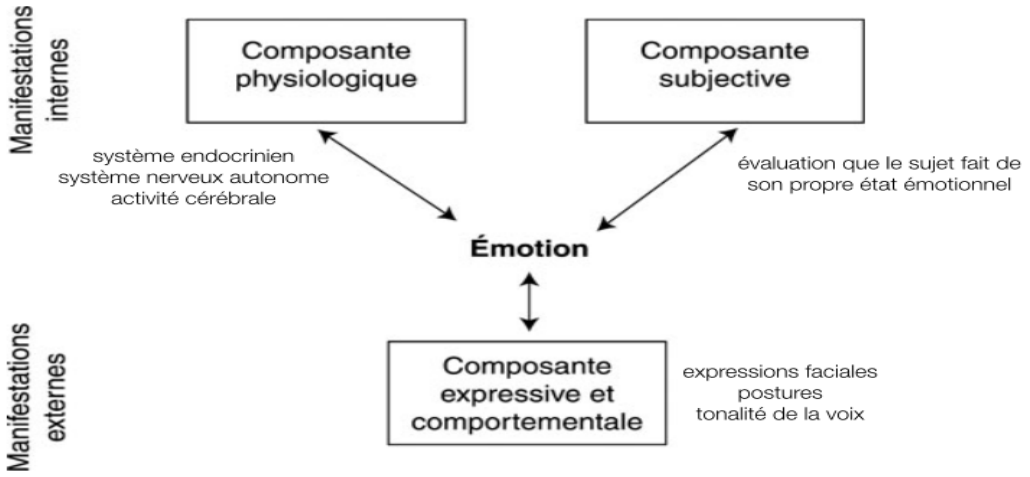

Figure 1-2 : Les trois composantes de l'émotion.

1.3.1 La composante expressive et comportementale

Dans son ouvrage « l'expression des émotions chez l'homme et l'animal » publié en 1872, Charles Darwin a mis en évidence la facette la plus figurative et représentative de l'émotion : L'expression. Qu'elle soit posturale (corporelle), vocale ou encor faciale, il a maintenu le fait que l'expression n'était pas non seulement une communication non verbale mais, serai aussi, une manière de s'adapter avec l'environnement et ses stimulis. Dans le prolongement de cette théorie, le psychologue Américain Sylvan TOMKINS, puis ses doctorants Caroll IZARD et Paul EKMAN considèrent que l'expression faciale 'est centrale dans l'émotion'.

a) L'expression faciale

Les expressions faciales représentent un moyen de communication plus rapide que la parole avec lesquelles les gens peuvent rapidement déduire l'état d'esprit de leurs compagnons, elles constituent ainsi, un puissant moyen de coordination sociale. Les expressions faciales de bases sont universelles, Friesen et Ekman ont reporté que six émotions, à savoir: La joie, la tristesse, le dégout, la colère, la surprise et la peur sont facilement reconnaissables et cela à travers plusieurs cultures [5]. La figure 1-3 est une illustration des muscles ainsi que du control nerveux responsables de l'expression faciale.

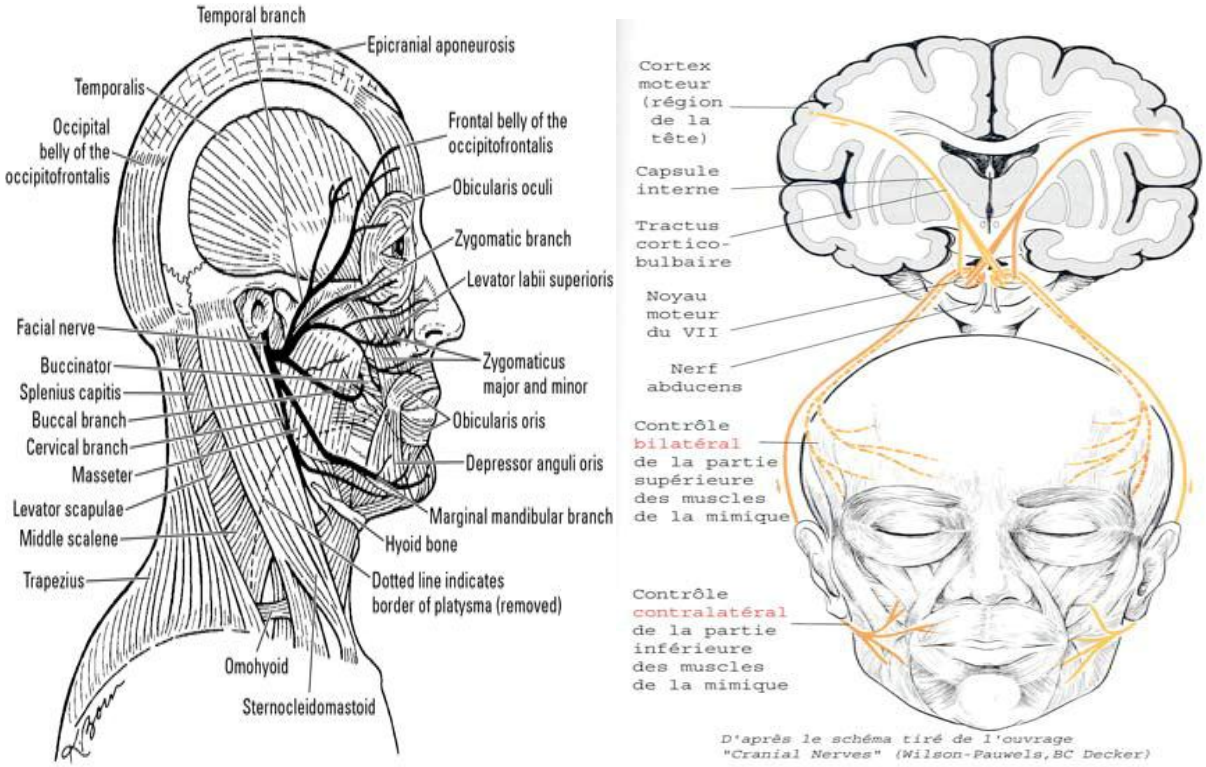

Figure 1-3 : Muscles faciaux et control nerveux [W1] [W2].

Le visage transmet deux types d'informations: Des indicateurs sur la personnalité de l'individu et des expressions de communication, d'émotions et d'intentions. La communication entre individus est très facilitée par la reconnaissance mutuelle des émotions qui sont dessinées sur leur visage via leurs expressions faciales: colère, gène, mécontentement. En outre, c'est grâce à la combinaison des mouvement des muscles faciaux et des traits du visage que les expressions spécifiques à chaque émotion sont représentées [2] [6] (voir figure 1-4).

Figure 1-4 : Les six émotions de base décrites selon Ekman [1].

b) L'expression vocale

Si les expressions faciales ont connu un véritable essor et un grand intérêt dés leur mention par Darwin en 1872, les phénomènes vocaux quant à eux ont connu un engouement que beaucoup plus tard dans les années 80 grâce aux travaux de Scherer (1986 et 1984). En effet, en grande partie à cause du manque de moyens, qui a été par la suite résolu par les méthodes d'enregistrements vocales et les différentes méthodes informatiques, le rythme et l'intonation de la voix n'ont pas été tout de suite reconnus comme étant évocateurs d'émotions [7].

La fréquence fondamentale F0, la durée du segment vocal ainsi que sont amplitude sont les paramètres les plus mesurés afin de déterminer l'étendu émotionnel d'une élocution. Chaque phrase contient deux types de prosodies : Une prosodie dite linguistique qui consiste à définir le type de phrase (affirmative, interrogative … etc), et une prosodie dite émotionnelle qui transcrit l'état émotionnel de l'individu [8]. En outre, d'autres types de mesures de types spectrales ont été mises au point afin de mieux étudier le timbre de voix et ainsi mieux évaluer la valence de l'émotion exprimée. Par exemple, une phonation grinçante est synonyme de la tristesse, la colère est exprimée par une voix roque, tandis que l'anxiété est représentée par une voix 'breathy', soit des parties vocales voilées à cause du rythme respiratoire [9].

c) L'expression corporelle

La posture (la position ainsi que l'orientation de parties du corps dans l'espace) ainsi que la gestuelle (mouvements visant à communiquer une information comme le faite de lever le pousse pour dire que quelque chose est bien, qu'un individu communique volontairement ou involontairement) ainsi que les différents mouvements du corps sont les trois aspects par lesquelles les expressions corporelles peuvent être exprimées. Même si cette composante connait un intérêt grandissant, elle présente des inconvénients majeurs qui sont:

- L'interférence avec la gestuelle utilisée dans le but de clarifier un discours.
- La similarité des mouvements et ceux pour plusieurs émotions, par exemple, le faite de mettre la tête en arrière est présente pour l'expression de la colère, de la joie, de la surprise et de la peur [10].

1.3.2 La composante physiologique

Une toute autre facette de l'expérience émotionnelle est la composante physiologique. C'est grâce à elle que le cerveau prépare l'action et cela grâce au système nerveux autonome (SNA) qui inhibe ou excite les parties du corps dont il est responsable (voir figure 1-5). Lors de la colère par exemple, les mains sont beaucoup plus irriguées en sang, préparant ainsi la personne à agir, comme le fait de prendre une arme par exemple, en outre une sécrétion plus accrue en adrénaline permet de fournir l'énergie nécessaire.

Plusieurs méthodes subsistent pour pouvoir détecter l'émotion avec les signaux physiologiques, parmi ces signaux: L'électro-encéphalographie (activité électrique du cerveau), l'activité électrodermale (variation des spécificités électriques de la peau), le volume respiratoire, la pression sanguine, l'activité musculaire, la température corporelle et la fréquence cardiaque.

Toutefois, elles sont heurtées à deux principaux obstacles : le premier est que l'outillage utilisé ainsi que le contact avec l'expérimentateur induisent des émotions parasites et du stress mentale pouvant fausser les résultats, le deuxième est que les mesures ne donnent d'informations que sur la valence ainsi que le degré d'activation de l'émotion , ce qui malheureusement ne permet pas de différencier les émotions entre elles [11]. Cela dit, la mesure de plusieurs signaux physiologiques permet la détection d'émotions [1].

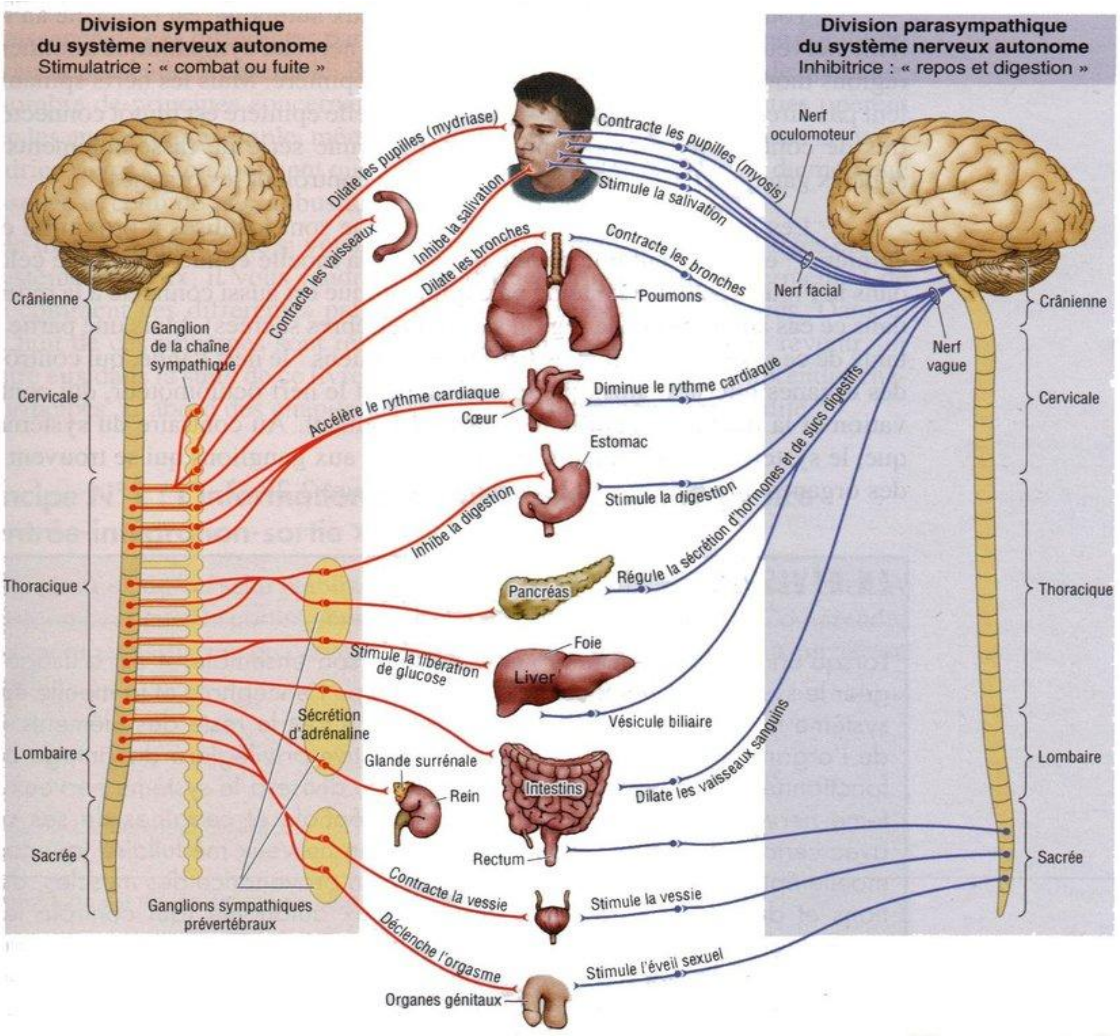

Figure 1-5 : Représentation du système nerveux autonome [W3].

1.3.3 La composante cognitive/subjective

Cette composante se base sur le ressenti de la personne ainsi que son évaluation pour la situation. Afin d'évaluer l'état émotionnel d'un individu en se basant sur cette composante, des questionnaires avec plusieurs sujets sont utilisés [1].

1.4 Les types d'émotions

1.4.1 Les émotions de base

Aussi appelées émotions basiques, primaires, ou encore fondamentales, elles constituent un petit ensemble d'émotions considérées comme innées. Elles ont un caractère universel, autrement dit, elles ont la spécificité de se manifester de façon identique quelque soit la culture [12]. A chaque émotion primaire son caractère somatique et une réaction du système nerveux [2]. Chaque auteur défini un nombre d'émotions différent, cela est en partie due aux critères d'inclusion qui diffèrent selon les études (voir tableau 1-1). Cela étant dit, parmi tous les ensembles d'émotions existants, celles d'Ekman restent les plus étudiées[13] .

Auteurs	Emotions basiques	Critères d'inclusion
EKMAN ET AL	Colère, dégout, joie, tristesse, peur, surprise, neutralité.	Expressions faciales universelles
PLUTCHIK	Acceptation, colère, anticipation, dégout, peur, joie, tristesse, surprise.	Relation avec les processus biologiques adaptatifs
FRIDJA	Intérêt, joie, désir, chagrin, émerveillement.	Formes de préparation à l'action
TOMKINS	Colère, intérêt, mépris, dégout, détresse, peur, joie, honte, surprise.	Densité du déclenchement neuronal
Arnold	Courage, colère, aversion, désir, désespoir, tristesse, amour, espoir, abattement, haine, peur.	Relation avec les tendances à l'action
McDougall	Peur, dégout, exaltation, colère, émotion tendre, soumission, émerveillement.	Relation à l'instinct

Tableau 1-1 : Emotions de base selon EKMAN et FREISEN, PLUTCHIK , Izard et TOMKINS [14].

1.4.2 Les émotions secondaires

Dites aussi émotions complexes, elle sont considérées par certains chercheurs comme issues de compositions d'émotions primaires dont elles gardent les propriétés fondamentales comme la valence [12] et qui se mettent en place à l'âge adulte[15] . Elles désignent aussi les émotions engendrées à l'évocation de souvenirs [2].

1.4.3 Les émotions sociales

Elles sont engendrées par les émotions primaires, l'éducation, et la culture. Elles sont inhérentes à la relation aux autres [2].

1.5 Représentation de l'émotion

1.5.1 Représentation catégorielle

Aussi appelée représentation discrète [16], elle consiste à définir un ensemble discret d'émotions dites universelles (les émotions de base), et considère que toutes les autres émotions découlent de ce dernier (à savoir les émotions secondaires). En outre, dans le domaine technique (l'informatique), cette représentation reste la plus utilisée [4].

1.5.2 Représentation dimensionnelle

Inversement à l'approche catégorielle, l'approche dimensionnelle classe de manière continue les émotions sur des axes représentant l'état émotionnel d'un individu suite à une expérience avec son environnement [16][2]. Afin d'évaluer la nature du vécu émotionnel, plusieurs dimensions ont été déterminées au fil des années et des études, cela étant dit, la valence et la dominance restent les plus mentionnés.

- **La valence :** Elle représente la qualité du vécu émotionnel en termes de positivité ou de négativité et varie selon an axe: un continuum continu variant du plaisant au déplaisant. En outre, en cas d'un nouveau stimuli, c'est la valence qui directement ressentie par la personne, avant de déterminé quel type d'émotion le stimuli a déclenché [17].

- **Le degré d'activation (arousal) :** Aussi dite 'éveil', elle représente une quantification de la valence. Elle varie selon un continuum continu allant d'une faible intensité à une forte intensité. En outre, elle représente le degré d'activation de la branche sympathique du système nerveux autonome (contrôlant par exemple le rythme cardiaque) [17].

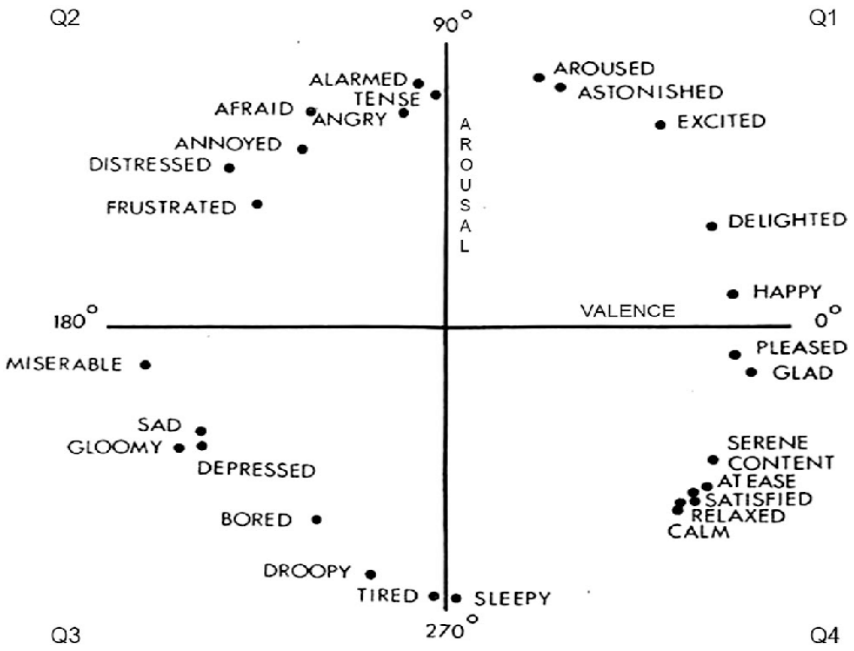

Figure 1-6 : Modèle de l'affect central (Roussel, 1980) [W4].

Au fil du temps, plusieurs modèles ont été conçus passant du tridimensionnel comme le modèle introduit par Mehrabian et Roussel en 1974, le PAD (**P**leasure **A**rousal **D**ominance), au bidimensionnel comme le circomplexe de Roussel en 1980 (aussi appelé modèle de l'affecte centrale) (voir figure 1-6). Toutefois, dans un article sorti en 2007, Lafontaine, Scherer, Roesch et Elsworth, affirment que pour décrire l'expérience émotionnelle au mieux, quatre dimensions sont nécessaires. Ainsi, au coté de la valence et le degré d'activation, deux autres dimensions ont été ajoutées : 'La dominance', qui correspond au sentiment que l'on a de pouvoir contrôler l'événement auquel on est opposé et 'l'imprévisibilité', qui correspond au sentiment que l'événement qui se produit n'est pas programmé [18] [19].

1.6 Conclusion

Dans ce chapitre, nous avons pu voir les différentes façons dont l'émotion apparait et vie à travers l'individu. Nous constatons toute fois des manières plus simples que d'autres pour l'élaboration de systèmes de détection automatique à savoir :

- Les expressions faciales ont la particularité d'être uniques pour chaque émotion fondamentale ressentie, ce qui simplifie la tache de détection.

- La composante posturale s'avère avoir des manifestations semblables pour plusieurs émotion, à savoir le mouvement de la tête ainsi que des bras qui est similaire pour plusieurs états émotionnels.

- La composante vocale nécessitera des traitements des signaux sonores (comme le traitement spectral) en plus du système de détection. En outre, plusieurs paramètres sont à prendre en compte (l'amplitude du signal, la durée, la différenciation entre prosodie linguistique et prosodie émotionnelle).

- La composante physiologique quant à elle, nécessite non seulement des capteurs mais aussi, nécessite de détecter non pas un mais plusieurs signaux physiologiques différents afin de pouvoir identifier les émotions ressenties.

Comme bilan de ce premier chapitre, nous dirons que le choix de la détection des expressions faciales pour la synthétisation de notre système de détection automatique est un choix judicieux qui nous permettra d'atteindre notre but avec le moins de difficultés possibles.

2 Etat de l'art des systèmes de détection automatique des émotions

2.1 Introduction

Dans ce chapitre nous allons introduire différentes méthodes de reconnaissance automatique des émotions via les expressions faciales. Nous aborderons en outre les avantages ainsi que les inconvénients de chaque technique. En outre, nous mettrons en exergue l'intérêt des réseaux de neurones et du DL dans le domaine de reconnaissance d'images en général.

2.2 Représentation des expressions faciales

Elle relie l'émotion aux mouvements des muscles responsables des expressions faciales. Il existe plusieurs types de représentation mais celles qui reviennent le plus sont : La représentation par le système FAC (**F**acial **A**ction **C**oding) et la représentation par le standard MPEG-4 (Une norme de codage d'objets audiovisuels spécifiée par le **M**oving **P**icture **E**xperts **G**roup).

2.2.1 Représentation par le standard MPEG-4

C'est un modèle de visage 3D construit sur la base de points caractéristiques (ensemble de points repères mis sur des zones d'intérêt du visage comme les yeux, la bouche et le nez) FFPs (Facial Features points). Les mesures faites sur ces FFPs forment ce qu'ont appelle les FAPUs (Facial animation parameter unit). Les FAPU représentent donc les distances entres différents FFPs. C'est grâce à la variation de ces distances que les mouvements élémentaires du visage FAPs (Facial Animation Parameters) sont estimés. Inversement à la représentation FAC, elle est utilisée pour les animations. Parmi ses applications les plus fantaisistes, mettre des émotions humaines sur des visages non humains [16] [20].

2.2.2 Représentation par le système FAC

Mise au point par Ekman et Friesen, c'est une méthode d'encodage qui consiste à décrire les expressions faciales selon les unités d'actions du visage (AU : unité d'action). Chaque AU représente le mouvement d'une partie du visage suite à l'activation du muscle responsable du mouvement (voir figure 2-1). Elle décrit donc l'expression faciale comme une combinaison de AUs [16] [21] [22] [23].

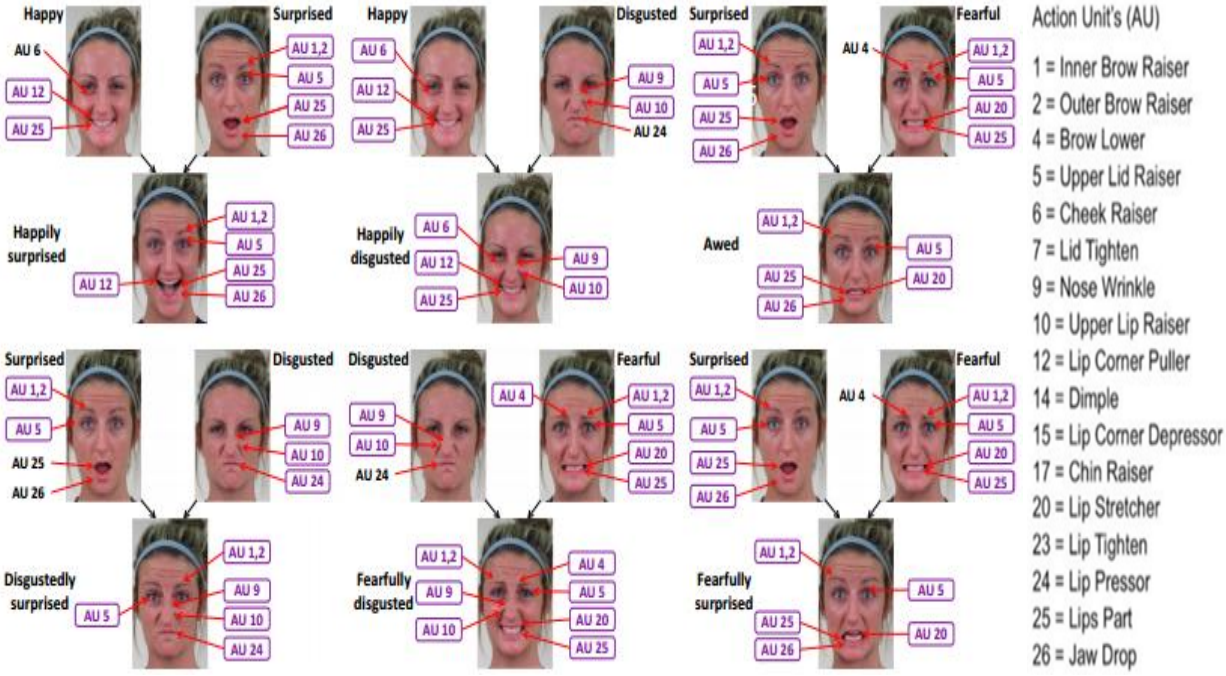

Figure 2-1 : Exemple de représentation de certaines des émotions de base et d'émotions secondaire avec le FAC [23].

2.3 L'analyse des expressions Faciales

Afin de pouvoir synthétiser un système de reconnaissance automatique des émotions en se basant sur les expressions faciales, trois étapes cruciales sont nécessaires, à savoir : La détection du visage, qui va permettre de délimiter la zone du visage, l'extraction des caractéristiques faciales et enfin, la classification qui vas déterminer à quelle émotion la combinaison d'expressions faciales extraites correspond.

Les figures suivantes (2-2 et 2-3) représentent la différence d'étapes entre un système de reconnaissance automatique des émotions classique et pour un CNN.

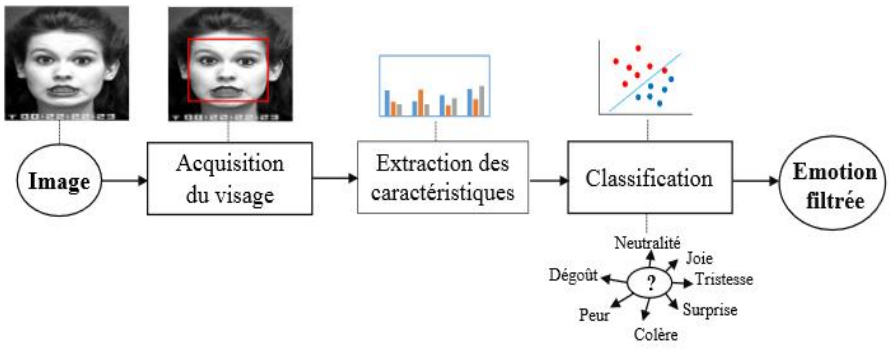

Figure 2-2 : Système de reconnaissance automatique des émotions [21].

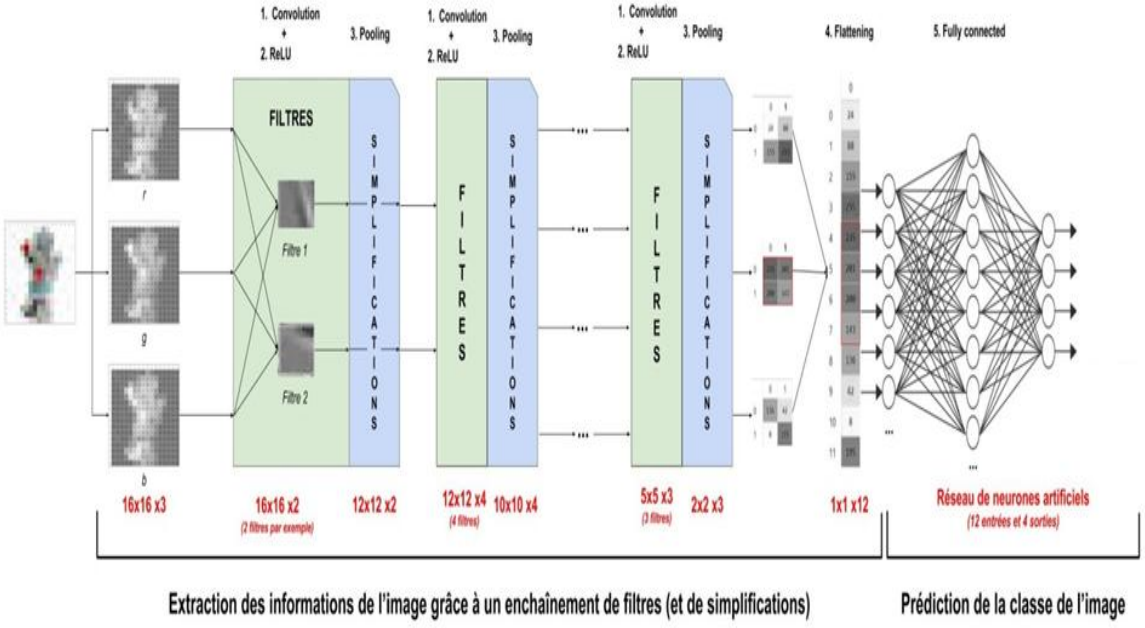

Figure 2-3 : Reconnaissance automatique des émotions avec les CNNs.

.

2.3.1 Détection du visage

Quatre classifications existent, à savoir, les méthodes se basant sur [2][36] :

- **Les connaissances acquises** : Elles se basent sur les caractéristiques faciales ainsi que sur les relations qui subsistent entre elles, parmi elles : Le calcul des positions relatives des différents éléments qui constituent le visage à savoir : La bouche, le nez, les yeux, la luminosité uniforme qui caractérise le centre du visage. Parmi les désavantages de ce genre de méthode : La difficulté de

pouvoir détecter les visages dans différentes orientations à cause de la difficulté à pouvoir recenser toutes les relations et caractéristiques faciales et cela pour toutes les positions du visage. En outre, il s'avère ardu de définir un visage de manière exacte, chaque visage et différent or, si la description est trop générale, ou à l'inverse trop précise, les erreurs de détection augmentent.

- **Les caractéristiques invariantes :** Similaires à la méthode citée précédemment, ces méthodes se basent sur des caractéristiques faciales mais invariantes à l'expression ou à la position et à l'orientation du visage comme la couleur de la peau. Cela étant dit, elles restent sensibles à la variation de luminosité.

- **La mise en correspondance :** Elles se basent sur la corrélation entre le candidat (le visage à détecter) et un modèle de visage préalablement créé.

- **L'apparence :** Elles se basent sur des méthodes d'apprentissage automatique à l'aide d'une base de données étiquetées. Parmi ces méthodes, celles basées sur les réseaux de neurones artificielles de T.Kanade, S.Baluja et H.Rowley ou encor, l'algorithme de Viola et Jones.

2.3.2 Extraction des caractéristiques faciales

Plusieurs méthodes ont été développées dans le domaine et peuvent être séparées de plusieurs manières, soit en méthodes globales et locales, soit en analyse bas, moyen et haut niveau [2], soit comme nous allons le voir : En caractéristiques d'apparence, en caractéristiques géométriques, en caractéristiques hybrides et en caractéristiques extraites à base de DL.

a. Caractéristiques d'apparence

Elles décrivent la texture de la peau comme les sillons et les rides et peuvent être extraites soit pour des parties du visage soit pour tout le visage. Parmi les méthodes les plus connues se basant sur cette dernière citons : les caractéristiques pseudo-Haar, les ondelettes de Gabor, l'analyse discriminante linéaire (LDA), l'analyse en composantes principales (ACP) ou encor les descripteurs basés sur le gradient [21].

Elles ont la caractéristique d'être robustes au désalignement de l'image, mais cela prend beaucoup de temps de calcul [21]. En outre, les traits du visage et notamment les rides ne dépendent pas que de l'expression faciale mais dépendent aussi de l'âge, chose qui peut fausser les résultats si la méthode n'intègre pas de techniques différenciant les rides naturelles au traits causés par l'expression faciale [20].

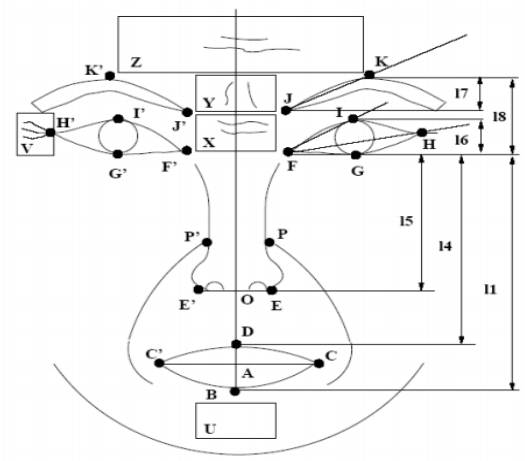

Figure 2-4 : Représentation des régions des sillons et des rides par les rectangles[2].

b. Caractéristiques géométriques

Elles décrivent la forme des composantes faciales du visage : les yeux, les sourcils, le nez, la bouche, et leur emplacement. Etant donné qu'une expression faciale affecte la position des composantes faciales ainsi que la taille des traits faciaux, c'est la mesure de ces changements et de ces mouvements qui permet de déterminer l'expression faciale correspondante. Parmi les travaux faits dans ce sens, citons ceux de Zhang et al (34 points), Tian el al (20 points) ainsi que Valstar et Pantic qui ont pu décrire le développement temporel des unités d'action en calculant les angles et les distances entre ces points. La figure 2-5 représente un exemple de points faciaux (fiduciaires) et traits faciaux [21].

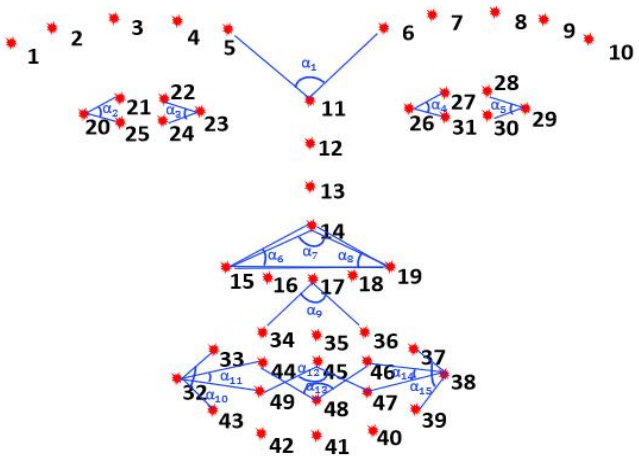

Figure 2-5 : Exemple de points caractéristiques [21].

Elles ont l'avantage d'être simples et de faibles dimension (moins d'espace mémoire), néanmoins, toutes s'avèrent être sensibles aux changements de luminosité. En outre, il est très difficile de concevoir un modèle décrivant les changements de traits faciaux pour toutes les expressions faciales. [21]

c. Caractéristiques hybrides

Ce sont des méthodes consistant à combiner les avantages des deux méthodes se basant sur les caractéristiques géométriques et d'apparence tout en essayons d'éviter leurs désavantages respectifs. Parmi les travaux faits dans ce sens, citons celui de Chen et al qui ont utilisé l'opérateur HOG-TOP pour pouvoir extraire les caractéristiques d'apparence puis l'ont combiné avec les caractéristiques géométriques ou encor, ceux de Tian et al qui ont utilisé les traits permanents (les yeux, les sourcils et la bouche) et les traits transitoires du visage (l'approfondissement des sillons faciaux) comme illustré dans la figure 2-6 [21] .

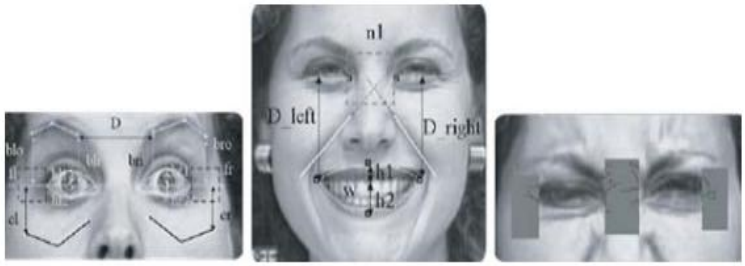

Figure 2-6 : Exemple des caractéristiques utilisées par Tian et al [2].

d. Caractéristiques extraites à base de deep learning

Khorrami et al ont montré empiriquement que les caractéristiques extraites par les CNN correspondent complètement aux FACs développés par Ekman et Friesen, en effet le DL en général permet l'extraction de caractéristiques faciales. Parmi les travaux faits dans ce sens, citons ceux de Kim et al qui ont exploré plusieurs architectures convolutives, ceux de Jung et al qui ont utilisé un combo de deux types d'architectures à savoir un CNN et un réseau entièrement connecté basé sur le DL ou encor, Liu et al qui ont construit une architecture profonde en utilisant des noyaux convolutifs pour apprendre les variations d'apparence locales provoquées par les expressions faciales et extraire des caractéristiques avec un Deep Belief Network (DBN) [21].

L'avantage de cette méthode est la robustesse des réseaux de neurones aux bruits de mesure (changements de luminosité), le désavantage réside dans la complexité à trouver une architecture performante ainsi que le choix de bons exemples de données à soumettre pour l'apprentissage. Cela dit, ils

ont montré de meilleures résultats que les autres méthodes citées, notamment pour les réseaux de neurones avec 'back propagation' qui peuvent atteindre 100% de bonnes estimations [20].

2.3.3 La classification

La dernière étape d'un système de reconnaissance automatique des émotions est la classification des émotions selon les caractéristiques faciales extraites. Elle se divise en deux types : La reconnaissance d'expressions basée sur des images statiques et la reconnaissance d'expressions basée sur des séquences vidéo. De nombreux classifieurs ont été adoptés pour la reconnaissance des émotions, parmi eux [21] :

- machines à vecteurs de support (Support Vector Machine, SVM).
- réseaux de neurone (Neural Networks, NN).
- réseaux bayésiens (Bayesian Network, BN).

2.3.4 Récapitulatif sur les méthodes de reconnaissance automatique des émotions

Il est difficile de concevoir un modèle déterministe capable de décrire les propriétés faciales pour toutes les rotations, toutes les expressions faciales et pour toutes les conditions de luminosité. Cela dit, les CNN confèrent plusieurs avantages et pas des moindres :

- L'extraction des caractéristiques est faite automatiquement.
- Ils sont robustes aux conditions de luminosité.
- Les images n'ont pas besoin d'un traitement particulier ou compliqué au préalable.

Le tableau (2-1) est un résumé des différentes méthodes utilisées pour la reconnaissance automatique des émotions.

Auteur	Caractéristique	Classifieur	Emotion	Type	Sujet/DB	Taux	Référence
Sebe et al	Mouvement de 12 unités	*KNN*	4	Image	CK:53 SD:28	93 % (CK) 95 % (SD)	[24]
Tong et al	Ondelettes de Gabor	*Adaboost* & RBD	14 AUs	Vidéo	CK:100 OD:10	93.2 % (OD) 93.3 % (CK)	[24]
Valstar et al	20 points faciaux	*Gentle SVM-sigmoide*	2	Vidéo	MMI:52	94 %	[24]
Whitehill et al	Descripteurs de HAAR	Adaboost	11 AUs	Image	/	92.35 %	[24]
Yeasin et al	L'intensité des pixels du visage	*KNN+HMM*	6	Image	CK:97 OD:21	90.7 % (CK) 72.82%(OD)	[24]
El Kaliouby et al	24 points faciaux	*RBD*	6	Vidéo	30	77.4 %	[24]
/	/	RN Hopfield	Tristesse, surprise, colère et joie	Image	/	92,2%	[20]
/	/	RN avec Back-propagation	6	Image	9 à 15 sujets	85% jusqu'à 100%	[20]

Tableau 2-1: Etat de l'art des méthodes utilisées pour la reconnaissance automatique des émotions.

2.4 Conclusion

A travers ses différentes composantes, Il est aujourd'hui possible de déterminer et d'identifier les émotions. Même si aucun modèle n'est parfait du faite que chaque individu soit différent, du faite aussi que le processus émotionnel reste un phénomène complexe, à cause aussi des limites matérielles, les performances des techniques utilisées ne cessent cependant de s'améliorer.

Une des méthodes que nous constatons s'avérer simple et efficace à la fois, est celle se basant sur le DL, plus précisément les CNNs. Cette méthode ne requière ni un modèle de visage prédéfini ni une implémentation de méthodes de calcules particulières, c'est le réseau qui fait tout et cela de bout en bout.

De la conclusion du chapitre 1, dans lequel nous avons constaté que la détection des émotions faciales était la plus simple et la plus efficace, nous allons introduire dans le chapitre qui suit (le chapitre 3), les caractéristiques ainsi que les spécificités des CNNs nécessaires à notre système de détection automatique d'émotions.

3 Les réseaux de neurones convolutifs

3.1 Introduction

Au cours des dernières années, et grâce aux progrès du DL, plus précisément en raison des CNNs [25], le domaine de détection d'objets ainsi que de reconnaissance d'images a connu un véritable essor. En effet, le concours ILSVRC (Imagenet Large Scale Visual Recognition Competition qui teste la performance des réseaux dans le domaine de la classification d'images ainsi que la détection d'objets) a montré que les architectures les plus efficaces pour le domaine de reconnaissance sont principalement basées sur l'apprentissage profond et les CNNs, chose qui a propulsé le concept à l'avant de la scène [26].

3.2 Définition de l'apprentissage profond

En anglais deep learning, deep structural learning ou encor hierarchical learning, l'apprentissage profond est une technique d'apprentissage automatique qui se base sur des architectures comprenant des transformations non linéaires et linéaires. Le mot deep qui veut dire profond, lui a été inculqué du faite de la superposition de plusieurs couches de traitement. Il a la capacité d'apprendre petit à petit les caractéristiques à travers chaque couche et ce, avec une intervention humaine minime [27].

Le DL est un domaine à croissance rapide et de nouvelles architectures, variantes ou algorithmes apparaissent toutes les semaines. Cela dit, on peut les classer en trois grand types de réseaux : Les CNNs, les réseaux récurrents (RNNs) dont les sorties ne sont pas indépendantes l'une de l'autre comme c'est le cas pour les CNNs, mais dépendent des sorties précédentes (réseaux avec mémoire)[28] et les réseaux génératifs. Contrairement aux CNNs et aux RNNs qui sont des réseaux discriminatifs et dont le but est de classer des données, les réseaux génératifs sont des réseaux qui prédisent et génèrent des données [29] .

Un des avantages du deep learning est la relation performance-quantité de données, contrairement aux autres algorithmes de ML, les performances des réseaux augmentent en fonction du nombre de données d'entrainement. En outre l'extraction des caractéristiques est faite automatiquement (voir figure 3-1).

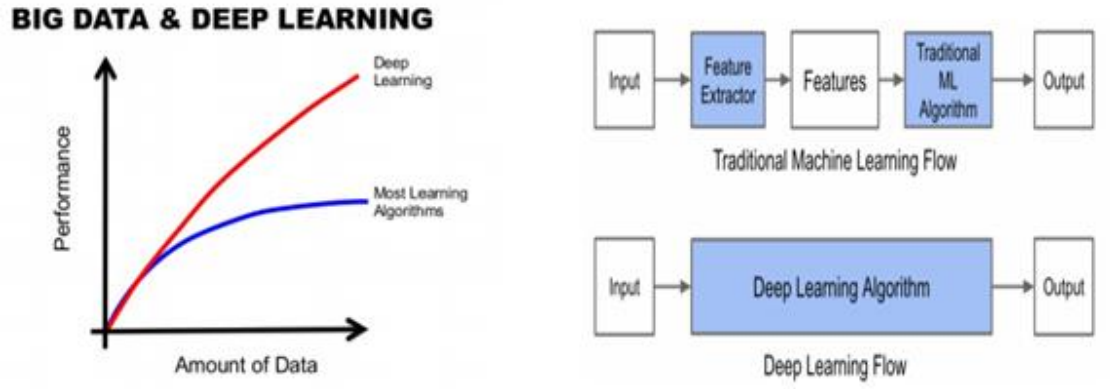

Figure 3-1 : Différence entre les algorithmes d'apprentissage profond et ceux de ML.

3.3 Définition des CNN

Ils représentent une catégorie de réseaux de neurones. Ils sont constitués de deux bloc principaux : Le premier bloc, vise à recevoir des images en entrées puis à effecteur du « template matching » en utilisant des opérations de filtrage par convolution. Le second bloc n'est pas spécifique aux CNNs et se présente à la fin de tous les réseaux destinés à la classification. Il est constitué d'une couche de sortie comportant autant de neurones qu'il y a de classes, les valeurs de cette dernière sont ensuite traduites en probabilités, le plus souvent via une fonction 'softmax'. Ils permettent l'extraction automatique des caractéristiques (dans notre cas les caractéristiques faciales) grâce aux couches de traitement composées de multiples transformations linéaires et non linéaires [30].

Hormis le faite qu'ils soient entrainables de bout en bout, les autres avantages des CNNs résident dans le faite que le nombre de paramètres soit drastiquement réduit grâce au 'parameter sharing' (voir figure 3.2) ainsi qu'au' pooling'. Un autre avantage est l'invariance par translation, en effet, grâce au simplifications faites lors du 'pooling', les 'features map' (résultat du filtrage d'une image par des noyaux de convolution) perdent les informations les moins importantes ce qui engendre que même si l'image est translatée ou n'est pas totalement de face, elle sera reconnue de nouveau [31] [32]. En outre, étant donné que les premières couches de ce type de réseau extraits seulement les informations générales de l'image (comme la luminosité et les bords) [33], l'apprentissage par transfert est très utilisé pour entrainer ce type de réseaux engendrant ainsi une baisse du temps de calcul.

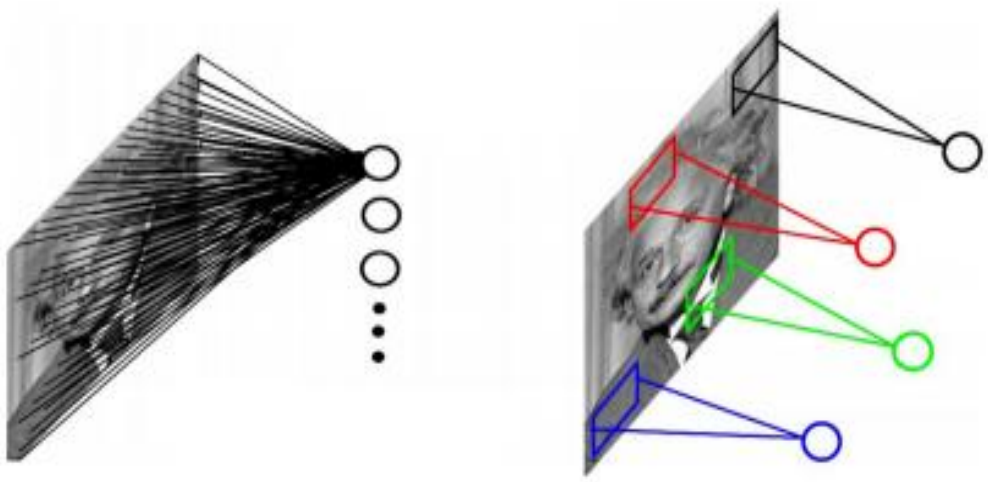

Figure 3-2 : Différence entre un réseau utilisant le 'parameter sharing' (à droite) et un réseau ne l'utilisant pas (à gauche).

3.4 Composantes d'un CNN

3.4.1 La couche de convolution

Son but est de repérer les informations pertinentes de l'image. Elle est conditionnée par trois hyper paramètres :'La profondeur de la couche' qui correspond au nombre de noyaux de convolution, 'le pas' (stride) qui contrôle le chevauchement des noyaux, 'une marge' (padding) qui permet de contrôler la dimension spatiale de la sortie par l'ajout de zéros et enfin, la taille des noyaux de convolution.

Après entrainement, les filtres ou noyaux et qui représentent une matrice de poids, sont mis à jour (appris) pour la reconnaissance. C'est grâce à ces filtres que le réseau après entrainement réussi à reconnaitre les images. 'Le filtre' est glissé sur l'image en entrée selon 'un pas' qui définit le nombre de pixels avec lequel le filtre se déplace sur l'image (voir figure 3-3). On peut parler de filtre si la matrice de poids contient 3 dimensions, si moins, on dit noyau (kernel).On peut distinguer, selon la manière dont se déplace le filtre ou le noyau, deux types de convolutions : La convolution en 2D si le noyau ne se déplace que verticalement ou une convolution en 3D si le filtre se déplace même en profondeur [W5] [34].

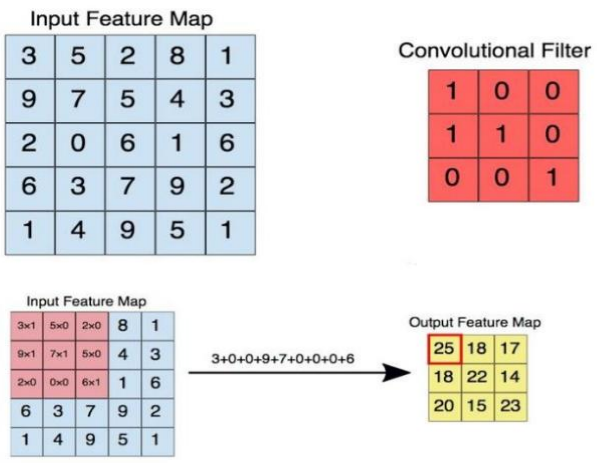

Figure 3-3 : Illustration d'une opération de convolution.

3.4.2 Les fonctions d'activation

Une fonction d'activation sert comme une fonction de décision au réseau et aide à la détection des caractéristiques. Le choix d'une telle ou telle fonction d'activation peut accélérer l'apprentissage et améliorer les performances du réseau. Plusieurs fonctions d'activation existent : Sigmoid, tnh, maxout, SWISH, ReLU, cela dit, c'est la fonction ReLU ainsi que ses variantes (ELU, LReLU) qui restent les plus utilisées étant donné qu'elles aident à surmonter le problème d'évanescence du gradient.

ReLU (Rectified Linear Unit) remplace toutes les valeurs négatives reçues par zéro (voir figure 3-4), opération décrite par l'équation (3-1). Ella a pour rôle d'éliminer la linéarité qui peut encore subsister après l'opération de convolution sans pour autant toucher au valeurs mises en évidence et considérées donc comme importantes (les valeurs positives). Elle engendre ainsi une baisse du temps de calcul [30]. Pour la couche de sortie, la fonction d'activation qui est la plus utilisée reste la 'softmax' (voir figure 3-5).

$$\text{ReLU}(x) = \max(0, x) \qquad (3\text{-}1)$$

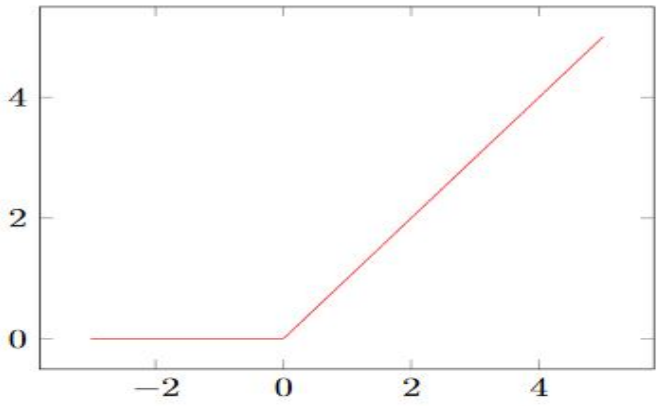

Figure 3-4 : Déroulement de la correction faite par la fonction ReLU [35] .

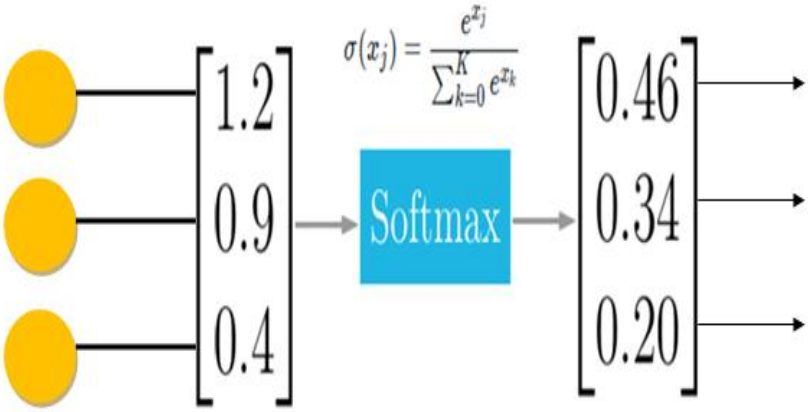

Figure 3-5 : Calcul effectué par la fonction 'softmax'.

3.4.3 La couche de pooling

Elle consiste en la réduction de la dimension des matrices de convolution obtenues après traitement des couches convolutives. Différents types de pooling existent :Le max Pooling (aussi le plus utilisé) et où le maximum des valeurs de la zone de pooling est pris et l'average Pooling, où la moyenne des valeurs est prise [36] (voir figure 3-6) .

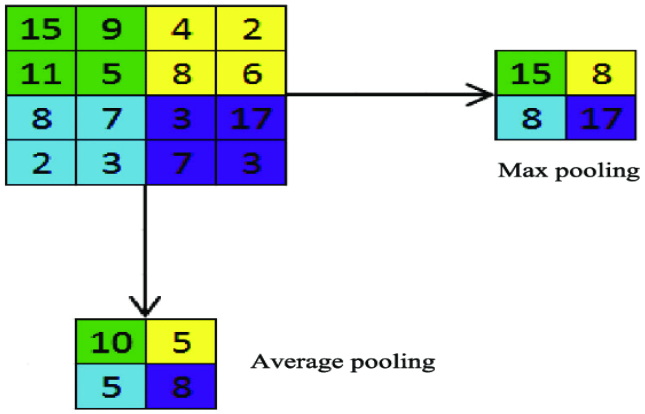

Figure 3-6 : Déroulement du 'max pooling' et de 'l'average pooling' pour un pas de deux et un pooling 2x2.

3.4.4 Les couches entièrement connectées (Perceptron multi couches)

Leur nombre diffère d'un réseau à un autre mais le nombre de neurones de la dernière couche correspond aux nombres de classes, et chaque valeur de sortie correspond à la probabilité d'appartenance à la classe correspondante. Le terme 'entièrement connectées' viens du faite que toutes les valeurs en entrées soient connectées en sortie (chaque neurone de sortie reçoit en entrées toutes les sorties des neurones de la couche précédente). En outre, un aplatissement est utilisé pour réduire la dimension des résultats des couches de traitement en un seul vecteur pour qu'il puisse être traité par les couches entièrement connectées, c'est ce qu'on appelle la mise à plat (flattering) (voir figure 3-7).

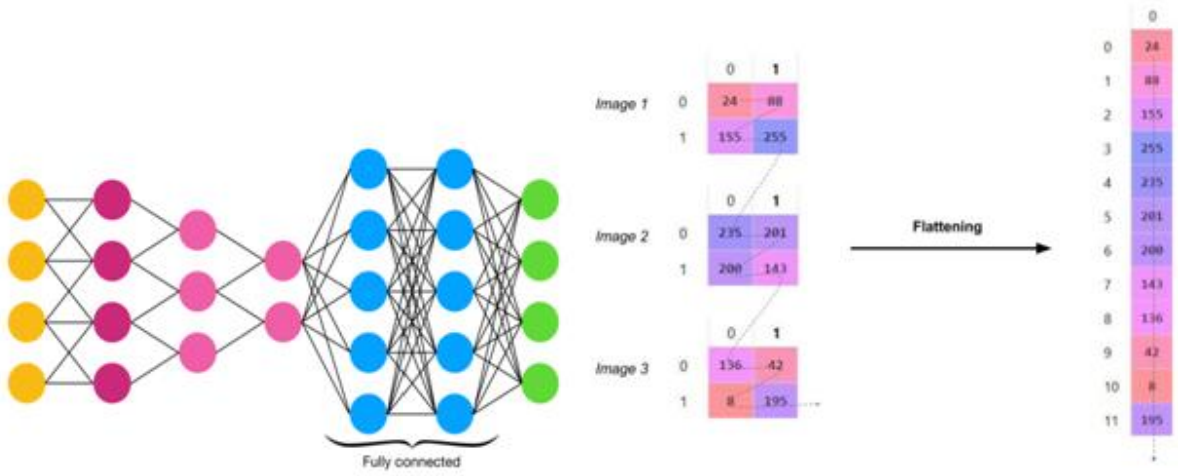

Figure 3-7 : De gauche à droite, couches entièrement connectées (MLP), opération de flattering.

3.4.5 La couche dropout

Elle correspond à la phase durant laquelle certains neurones sont aléatoirement désactivés (voir figure 3-8). Elle est utilisée surtout pour forcer le CNN à s'adapter au déficit de données mais peut aussi s'avérer utile pour éviter le sur-apprentissage du réseau [37].

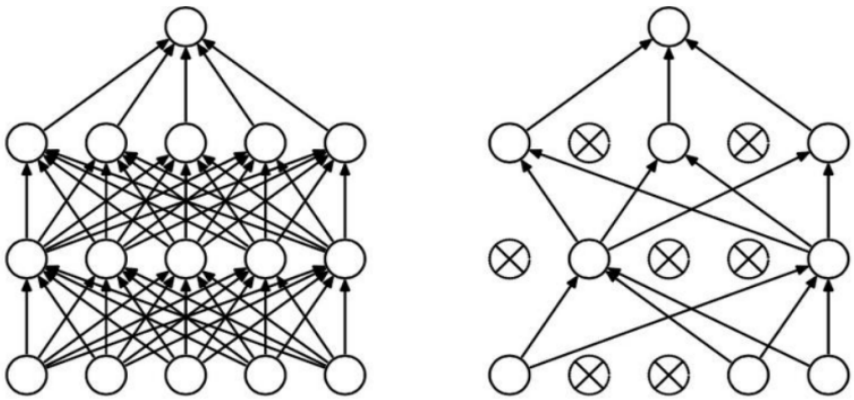

Figure 3-8 : Application de la couche dropout [37].

3.4.6 Une fonction de perte

L'étape la plus importante dans la conception d'un réseau est sans doute l'entrainement. Outre le choix de la base de données qui doit être fastidieux, le choix de la fonction de perte est aussi important. La fonction de perte évalue l'erreur qui peut subsister entre la sortie du réseau $s_{\Theta,n}$ avec la sortie désirée g_n. L'objectif de l'entrainement est donc d'ajuster les paramètres Θ du réseau (les poids) de sorte à avoir la valeur la plus petite possible pour la fonction de perte et ceux, pour chaque entrée (avec N le nombre de classes). Parmi ces fonctions citons, la 'Mean Squared Loss' (MSL), et la plus utilisée la 'cross entropy' (CE) [38].

$$\text{MSL}(\Theta) = \sum_{n=1}^{N}(s_{\theta,n} - g_n)^2 \qquad (3\text{-}2)$$

$$\text{CE}(\Theta) = -\sum_{n=1}^{N} g_n \log\left(s_{\theta,n}\right) \qquad (3\text{-}3)$$

Pour la classification multi-classes, la 'Categorical Cross Entropy'(CCE) reste la plus employée. Elle est une combinaison de la fonction d'activation 'softmax' et de la CE.

$$\text{CCE}(\Theta) = -\sum_{n=1}^{N} g_n \log\left(F(\theta)_n\right) \quad (3\text{-}4)$$

$$\text{Avec : } F(\theta)_i = \frac{e^{s_{\theta,i}}}{\sum_{j=1}^{N} e^{s_{\theta,j}}} \quad (3\text{-}5)$$

3.4.7 Les optimiseurs

Les optimiseurs sont des fonctions mathématiques visant à optimiser une fonction objective. Dans le cas des réseaux de neurones, leur travail consiste à mettre à jour les paramètres (poids) afin de minimiser l'erreur estimée par les fonctions de perte. Parmi ces optimiseurs :

a. Adam

De son nom, estimateur de moment adaptatif, l'optimisateur Adam a un taux d'apprentissage adaptatif pour chaque poids (ou paramètres en général) et garde le carré des gradients précédents. Son fonctionnement est comme suit :

Soit : $\beta 1$, $\beta 2$ des constantes <1, f(Θ) la fonction objective (la fonction de perte)

$g(\theta_t) \leftarrow \nabla_\theta f_t(\theta_{t-1})$ Calcul du gradient de la fonction de perte.

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g(\theta_t)$

$u_t \leftarrow \beta_2 \cdot u_{t-1} + (1-\beta_2) \cdot g(\theta_t)^2$

$\hat{m}_t \leftarrow m_t/(1-\beta_1^t)$

$\hat{u}_t \leftarrow u_t/(1-\beta_2^t)$

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t/(\sqrt{\hat{u}_t} + \epsilon)$ Mise à jour des paramètres.

Les valeurs des paramètres par défaut préconisés sont de : α=0.001, β_1=0.9, β_2=0.999, $\epsilon=10^{-8}$ [39].

b. Rmsprop

Mise au point par Geoff Hinton, Rmsprop est une méthode d'optimisation non publiée. Elle consiste à diviser le taux d'apprentissage d'un poids par une moyenne mobile 'E' des amplitudes des gradients récents pour ce poids.

$$E[g^2]_t = \gamma\, E[g^2]_{t-1} + (1-\gamma)\, g(\theta_t)^2 \quad (3\text{-}6)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]t+\epsilon}}\, g(\theta_t) \quad (3\text{-}7)$$

Pour la valeur des paramètres, G.Hinton propose que le taux d'apprentissage η soit de 0.001 et que γ soit fixé à 0,9 [40] [W6].

c. SGD

De son nom Descente de gradient stochastique, pour chaque exemple de l'ensemble de données, il fait une mise à jour des paramètres de la fonction objective J qui est en fonction de la donnée x et du label y, η est le taux d'apprentissage [41].

$$\Theta_{t+1} = \Theta_t - \eta\, g(\theta_t) \quad (3\text{-}8)$$

d. Adagrad

De son nom Adaptative Gradiant Algorithme, il est basé sur la descente de gradient qui adapte le taux d'apprentissage pour chaque paramètre. La mise à jour des paramètres se fait ainsi avec α le stepsize et g_t le gradient de la fonction objective J(qui est la fonction de perte) [39] :

$$\theta_{t+1} = \theta_{t,} - \frac{\alpha}{\sqrt{\sum_i^t g_t^2}}\, g(\theta_t) \quad (3\text{-}9)$$

3.5 Évolution architecturale des CNN

De nos jours, les CNNs sont considérés comme les algorithmes les plus largement utilisés parmi les techniques d'intelligence artificielle (IA) d'inspiration biologique. Au fil des décennies, différents efforts ont été menés pour améliorer leurs performances et plusieurs architectures sont apparues : LeNet, AlexNet (gagnant ILSVRC 2012), ZFNet (gagnant ILSVRC 2013), GoogleNet (gagnant ILSVRC 2014), VGG, ResNet (gagnant ILSVRC 2015) [30].

3.5.1 LeNet

LeNet a été proposé par Lecun en 1998. Il est célèbre en raison de son importance historique car il a été le premier CNN, qui a montré des performances de pointe sur les tâches de reconnaissance de chiffres à la main. Il a la capacité de classer les chiffres sans être affecté par de petites distorsions**,** rotation et changements de position et d'échelle. LeNet est un réseau neuronal à rétroaction composé de cinq couches de convolution (figure 3-9)**,** suivies de deux couches entièrement connectées. La principale limitation du CNN multicouche traditionnel entièrement connecté est qu'il traite chaque pixel comme une entrée distincte et le transforme, ce qui représente beaucoup de calcul, surtout à l'époque (Gardner et Dorling 1998). LeNet est la première architecture CNN, qui non seulement réduit le nombre de paramètres, mais peut également apprendre automatiquement les caractéristiques des pixels d'origine. [30]

3.5.2 VGG

L'utilisation réussie des CNNs dans les tâches de reconnaissance d'image a accéléré la recherche en conception architecturale. À cet égard, Simonyan et al ont proposé un principe de conception simple et efficace pour les architectures CNN : Leur architecture nommée VGG. ZFNet, qui était un réseau de première ligne de la compétition 2013-ILSVRC, a suggéré que les filtres de petites tailles peuvent améliorer les performances des CNNs. Sur la base de ces résultats, VGG a remplacé les filtres 11x11 et 5x5 par une pile de couches de filtres 3x3 et a démontré expérimentalement que le placement simultané de filtres de petite taille (3x3) pouvait induire l'effet du filtre de grande taille (5x5 et 7x7). L'utilisation de filtres de petite taille offre un avantage supplémentaire de faible complexité de calcul en réduisant le nombre de paramètres. Ces résultats ont établi une nouvelle tendance dans la recherche à travailler avec des filtres de plus petite taille. VGG montre de bons résultats dans la classification des images et les problèmes de localisation. Il s'est classé deuxième dans le concours 2014-ILSVRC, mais il a été reconnu pour sa simplicité. Sa principale limitation est liée à est l'utilisation de 138 millions de paramètres, ce qui rend le coût de calcul élevé et difficile à déployer sur des systèmes à faibles ressources [30]. La figure 3-10 représente l'une de ces variantes les plus connues : VGG16.

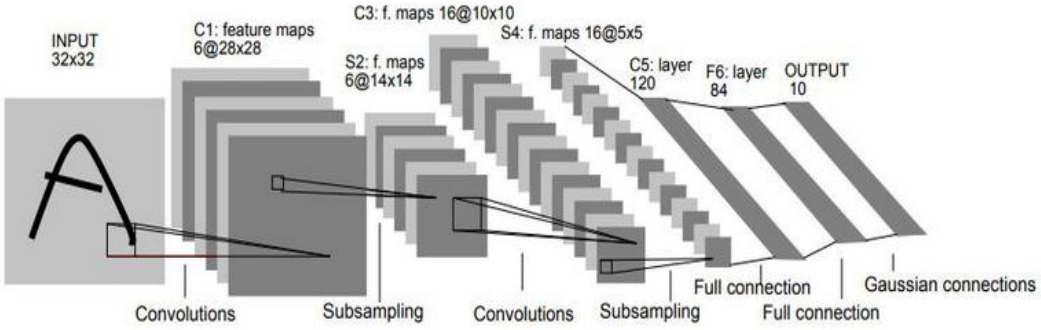

Figure 3-9 : Architecture LeNet [W7].

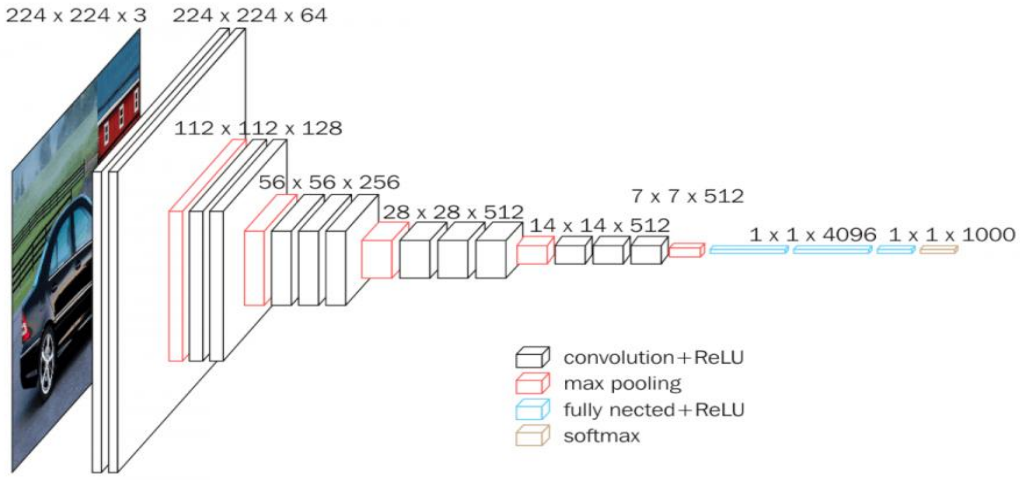

Figure 3-10 : Architecture VGG16 [35] .

3.6 Les bases de données

3.6.1 RAFD

La base de données RADBOUD Faces Dataset (RAFD) est un ensemble d'images de 67 modèles entrainés par deux spécialistes du FACs (des hommes et des femmes de race blanche, des enfants de race blanche, garçons et filles, et des hommes hollandais et marocains) affichant 8 expressions émotionnelles (colère, mépris, joie, tristesse, surprise, peur, dégout, et le neutre). Elle constitue une base de données de visages de haute qualité, qui contient 1608 distribuées ainsi : Chaque émotion est représentée avec une

résolution of 640 × 1024 pixels [42].Les images ont été prises avec trois directions de regard différentes et avec cinq angles de caméra simultanément (voir figure 3-11) [43].

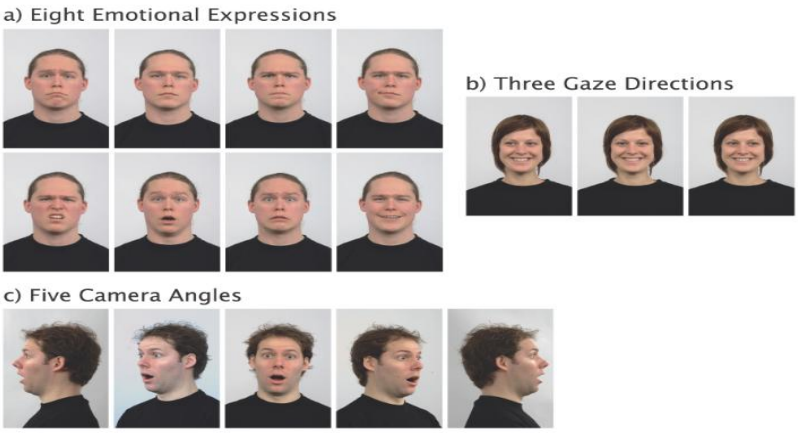

Figure 3-11 : Echantillons d'images de la DB RaFD [43].

3.6.2 JAFFE (Japanese Female Facial Expression)

Créée et assemblée par Michael Lyons, Miyuki Kamachi et Jiro Gyoba, elle contient 213 images d'expressions faciales simulées par dix sources Femmes japonaises. Elles ont simulé 3 à 4 exemples pour chacune des sept émotions basiques (colère, dégout, joie, tristesse, peur, surprise et neutres). La résolution de l'image est de 256 × 256 pixels.

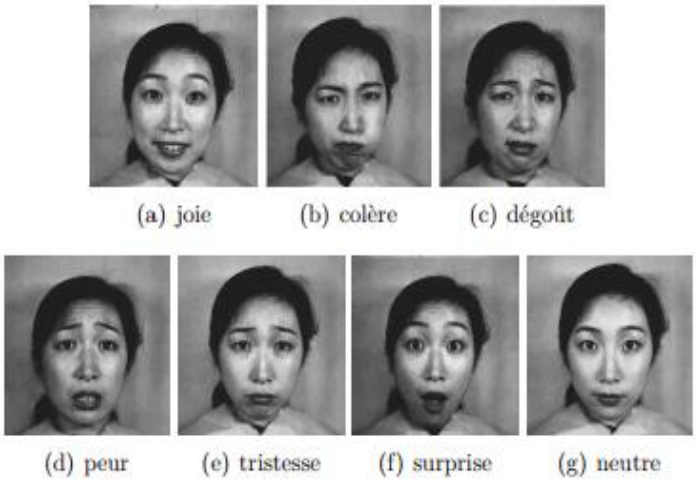

Figure 3-12 : Echantillons d'images de la JAFFE [16].

3.6.3 Fer2013

Fer2013 est un ensemble de données open source crée par Pierre-Luc Carrier et Aaron Courville pour un projet, puis partagé publiquement pour un concours 'Kaggle'. Elle se compose de 35 887 images en niveau de gris de 48x48 et contient 7 émotions toutes libellées, à savoir : 4 593 d'images pour la colère, 547 images pour le dégout, 5 121 images pour la peur, 8 989 images pour la joie, 6 077 images pour la tristesse, 4 002 images pour la surprise et 6 198 images pour le neutre. Contrairement aux autres DB qui sont des images, la fer2013 est représentée sous forme de tableau Excel. [W8]

Figure 3-13 : Echantillons d'images de la fer2013 [W8].

3.7 Conclusion

La conclusion qu'on peut tirer de ce chapitre est que les CNNs sont synonymes d'un air nouveau pour les domaines de vision par ordinateurs et de reconnaissance d'images en général. Les avantages qu'ils confèrent justifient leur succès et les rendent plus faciles à implémenter. Parmi ces avantages citons :

- L'entrainement de bout en bout.
- L'invariance par translation.
- La puissance des ordinateurs d'aujourd'hui (GPU).
- Le grand nombre de bases de données étiquetées.
- Le grand nombre de bibliothèques facilitant le traitement des images (Opencv).
- La facilité d'implémentation
- Le grand nombre de codes déjà existant (Github) favorisant ainsi, l'utilisation de l'apprentissage par transfert qui diminue le temps d'entrainement du réseau.

Les CNNs offrent un vaste choix de modifications et de développement. Leurs performances dépendent non seulement de leur architectures en elle-même (nombre de couche de convolution, le type de pooling, le nombre de filtres etc), mais pas seulement. Dans le chapitre qui suit nous allons étudier l'influence de la base de données ainsi que de la modification de certain des hyper-paramètres sur l'apprentissage de quatre architectures inspirées de VGG et LeNet.

4 Méthodologie, résultats et discussion

4.1 Introduction

Ce chapitre est consacré à la manière avec laquelle le benchmark a été construit. Les performances des réseaux seront comparées en utilisant les taux de validation (val_acc et val_loss) qui vont nous donner une première idée sur les performances de nos architectures ainsi que sur le fait qu'ils soient en sur-apprentissage ou non (figure 4-1). Nous allons aussi les évaluer selon les taux du test (test_acc et test_loss).

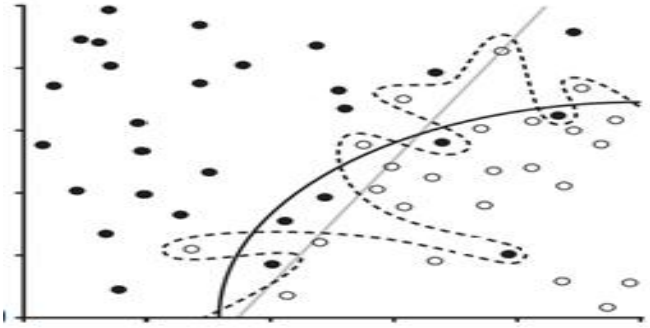

Figure 4-1 : Modèle correct (trait en continu), modèle sur - entrainé (trait discontinu) et un modèle sous entrainé (trait en gris) [44].

La correction des erreurs des différents réseaux se fera selon le concept de l'apprentissage supervisé comme illustré dans la figure suivante (4-2).

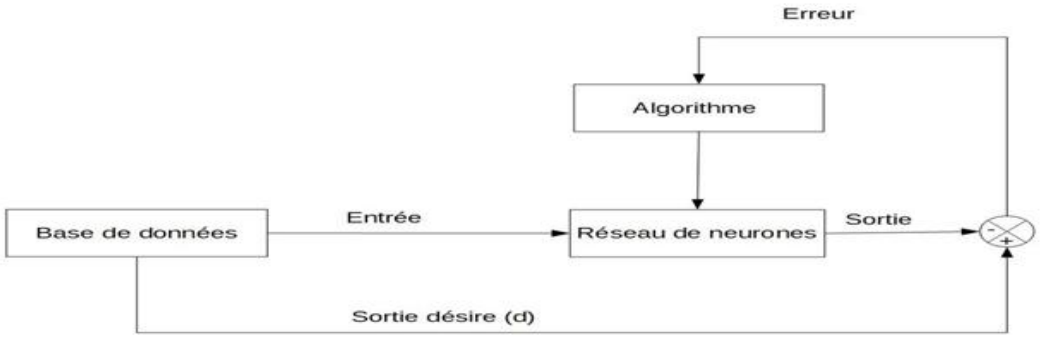

Figure 4-2 : Structure générale du système de détection mis au point.

Pour l'implémentation des architectures et les tests, nous diviserons cette partie réalisation en quatre parties. La première partie sera consacrée au choix de l'architecture durant laquelle nous testerons plusieurs architectures inspirées de LeNet et VGG16 et choisirons les plus performantes. La deuxième partie sera consacrée à l'influence du choix des DBs sur les résultats de l'implémentation. Dans la troisième partie nous testerons nos différents réseaux avec plusieurs optimiseurs : SGD, Adam, Rmsprop, et Adagrad. Enfin, nous allons varier le batch size afin d'avoir les meilleurs résultats possibles. Pour tous les modèles, nous utiliserons la fonction de perte 'categorical cross entropy', la fonction d'activation ReLU, la fonction 'softmax' pour le calcul des probabilités des couches de sortie et un nombre d'époques de 200.

4.2 Environnement de travail, logiciels et matériels utilisés

Caractéristiques du PC :

- **Marque du pc** Acer A715
- **Processeur** Intel (R) Core (TM) i5-8300H CPU
- **GPU** GTX 1050 4 GB
- **RAM** 20GB

Logiciels utilisés :

- **Anaconda** pour les environnements.
- **Spider** pour la programmation.

4.3 Bibliothèques utilisées

- **Opencv (Open Source Computer Vision)** : C'est une bibliothèque graphique libre distribuée sous une licence BSD spécialisée dans le traitement d'images. Elle propose un ensemble de plus de 2500 algorithmes de vision par ordinateur et peut être utilisée pour la programmation en C, C++, Python et ceux pour les plateformes Windows, GNU/Linux, Android et MacOS [W9].
- **Numpy (numerical python):** C'est un package pour le calcul scientifique sous forme de tableau. Elle permet de rendre les opérations beaucoup plus efficaces surtout sur les tableaux de grande taille. Les tableaux numpy ne contiennent des nombre que d'un seul type, si les nombre à stocker sont de type différent, numpy convertira tous les nombre au type le plus général (comme convertir le type 'int' en 'float') [W10] [W11].

- **Matplotlib:** C'est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python [W12].

- **Scikit learn:** C'est la principale bibliothèque d'outils dédiée au ML et à la data-science dans l'univers Python. Elle s'utilise conjointement avec d'autres bibliothèques python comme Matplotlib, Pandas ou Numpy. Parmi ses utilisations citons le découpage d'un ensemble de données en données de test et d'entrainement [W13].

4.4 Prétraitement des images

Pour le Benchmark, trois bases de données seront utilisées : La RaFD, la Fer2013 et la JAFFE. Afin d'accélérer l'exécution d'un programme et diminuer le nombre de paramètres que le réseau aura à apprendre, deux types de prétraitements peuvent être effectués sur la base de données à savoir : Un redimensionnement des images et une réduction du nombre de canaux.

Pour les trois bases de données, nous n'avons fait qu'un simple redimensionnement en images 48x48 pixels. Nous n'avons pas fait une réduction des canaux RVB car les images des dataset Fer2013 ainsi que la JAFFE sont déjà en gris.

En ce qui concerne la RaFD, nous avons décidé de la laisser en couleurs afin de reproduire au mieux les conditions réelles d'acquisition d'images. On y a omis les images représentant 'le mépris' étant donné que cette émotion ne figure pas parmi celles décrites par Ekman. En outre, les images non prises de profil ont été retirées puisque c'est l'algorithme de 'Viola et Jones' qui sera utilisé pour pouvoir détecter les visages dans les images afin de faire un redimensionnement correct.

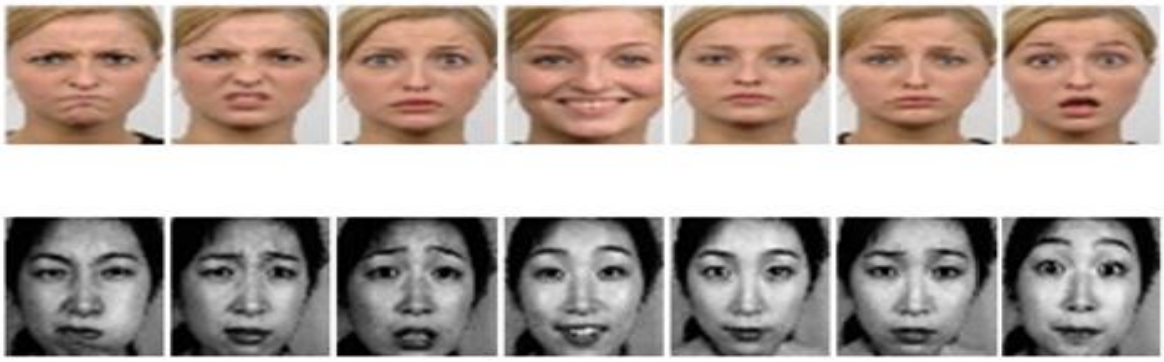

Figure 4-3 : Echantillons d'images prétraitées de la base de données RaFD et JAFFE pour la 'colère', le 'dégout', la 'peur','la 'joie', le 'neutre', la 'tristesse' et la surprise' respectivement.

4.5 Choix des CNN

Pour notre benchmark, nous allons tester quatre architectures différentes : Deux inspirées de LeNet, et deux inspirées de VGG16. Etant donné que nous n'avons pas une très grande quantité de données à traiter, nous n'avons pas à recourir à des architectures très profondes telles que ResNet ou Inception. En outre, l'entrainement de réseaux très profonds est compliqué et ceux pour deux raisons :Le temps nécessaire à l'entrainement qui croît plus le nombre de couches est grand ainsi que le problème de 'fuite du gradient' qui croit aussi avec le nombre de couches de traitement [30].

Les architectures sont modifiées de sorte à avoir un meilleur résultat pour un nombre d'époques moyen (200 époques). Pour ce fait, nous allons modifier les paramètres suivants

- Le type de pooling (average pooling ou max pooling).
- L'ajout ou non de couches ReLU, de couches de convolution et de couches dropout.
- Le nombre de filtre de convolution ainsi que leur taille.

4.5.1 CNN implémentés

- LeNet_v1 et LeNet_v11 qui sont des architectures LeNet modifiées se rapprochant le plus de l'architecture originale. Comme l'architecture décrite par Y.Lecun, elles contiennent deux couches de convolution, deux couches entièrement connectées et un nombre relativement petit de filtre de convolution avec une taille de filtres de 5x5 (voir figure 4-2).

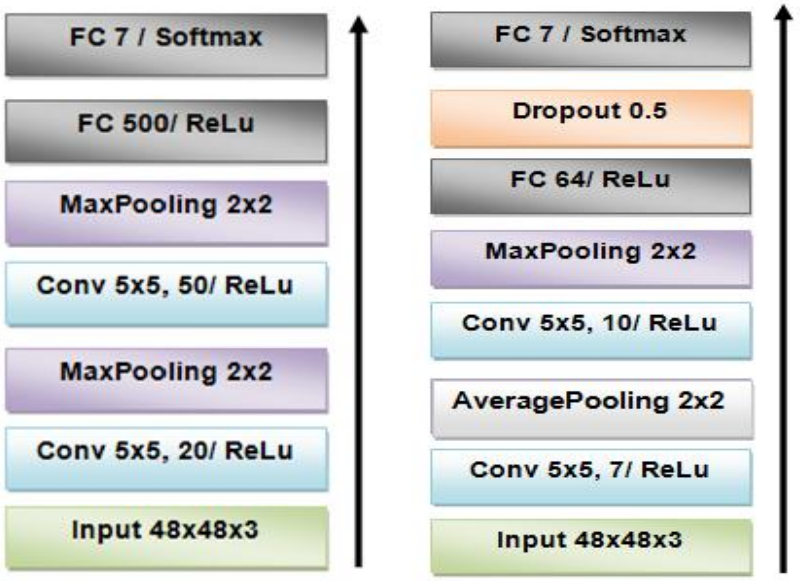

Figure 4-4 : De gauche à droite, implémentation LeNet_v1 et LeNet_v11.

- LeNet_v2 et LeNet_v22 qui sont aussi inspirées de l'architecture LeNet, la différence est l'utilisation d'un plus grand nombre de filtres avec une plus petite taille (voir figure 4-3).

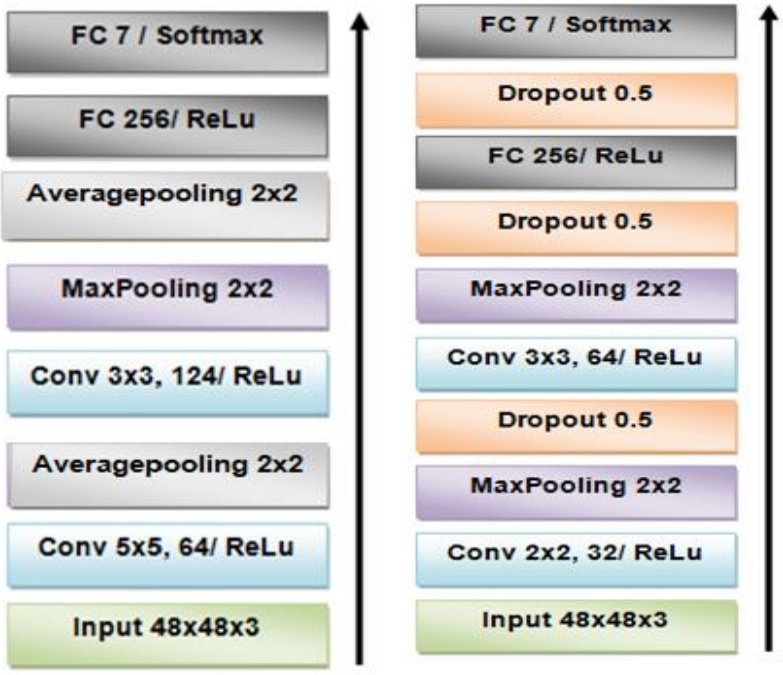

Figure 4-5 : De gauche à droite, Implémentation de LeNet_v2 et LeNet_v22.

- VGG_v1, VGG_v2 et VGG_v3 qui sont des architectures inspirées de l'architecture VGG16 originale. Pour VGG_v1, les grandes lignes de l'architecture originale ont été respectées, à savoir : la succession de couches de convolution sans couche de pooling entre elles ainsi que la petite taille des filtres de convolution, cela dit, les couches de correction ReLU n'ont pas été implémentées. Pour VGG_v2, des couches ReLU et des couches dropout ont été ajoutées et pour VGG_v3, le 'max pooling' a été remplacé par un 'average pooling' (voir figure 4-4).

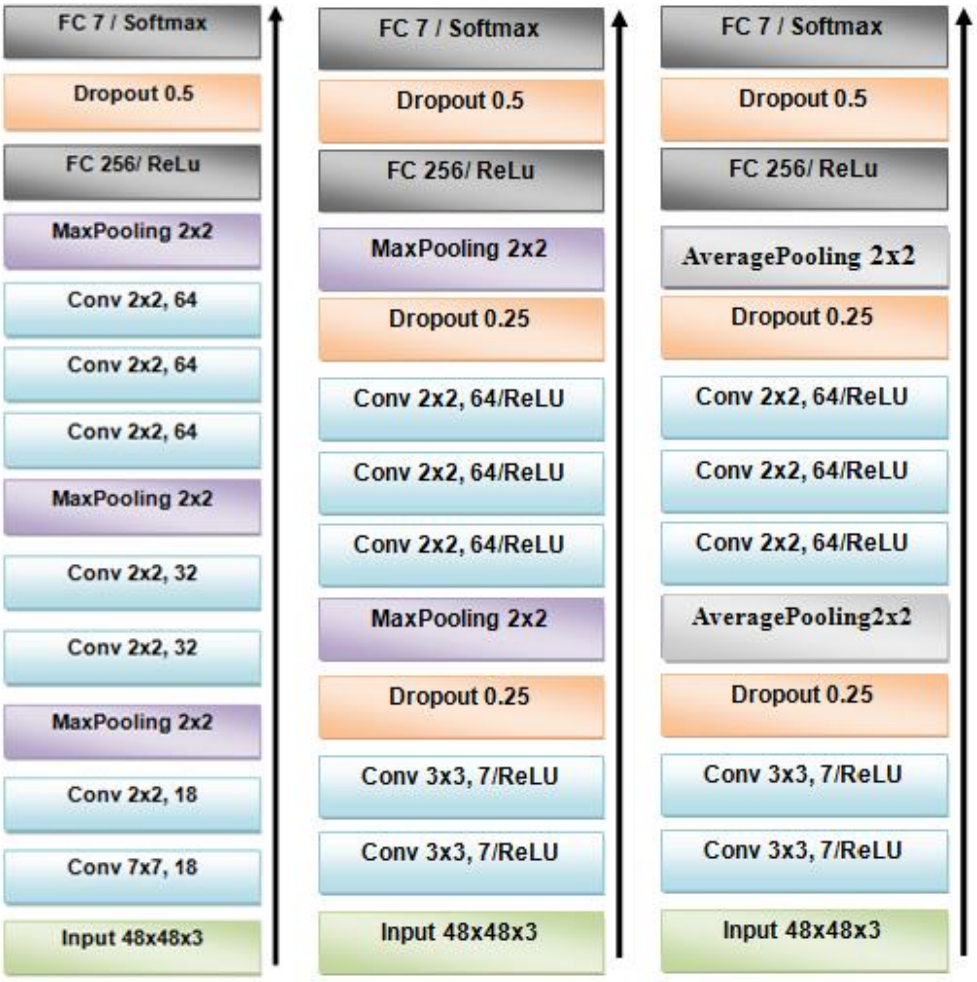

Figure 4-6 : De gauche à droite, implémentation de VGG_v1, VGG_v2 et VGG_v3.

4.5.2 Résultats

Pour mieux observer l'évolution du taux de validation et de test pour les sept architectures (voir tableau 4-2), nous allons tester les architectures pour un nombre d'époques de 200, et cela pour la base de données RaFD.

On a choisis de les tester avec la DB RaFD car comme l'illustre le tableau (4-1), la RaFD est celle qui a donnés les meilleurs taux d'apprentissage (pour 20 époques). Cela dit, une fois les architectures les plus performantes choisies, nous les testerons avec les trois DB.

DB	RaFD		JAFFE		Fer2013	
Taux d'apprentissage	Acc	Loss	Acc	Loss	Acc	Loss
LeNet_v1	0.9805	0.0866	0.6784	0.7557	0.9883	0.0607
LeNet2_v2	0.9796	0.0713	0.7836	0.5834	0.8766	0.3421
VGG_v1	0.9814	0.0756	0.6257	0.9781	0.7118	0.7531

Tableau 4-1 : Taux d'apprentissage de LeNet_v1, LeNet_v2 et VGG_v1 pour 20 époques, un batch size de 120 et l'optimiseur Adam.

Architecture	Val_acc	Val_loss	test_acc	test_loss
LeNet_v1	0.9365	0.1588	0.9642	0.4190
LeNet_v11	0.9841	0.0582	0.9928	0.2473
LeNet_v2	0.9762	0.1205	0.9714	0.4165
LeNet_v22	0.9841	0.0409	0.9785	0.1901
VGG_v1	0.9603	0.0706	0.9857	0.3260
VGG_v2	0.9524	0.1017	0.9714	0.1980
VGG_v3	0.9683	0.0904	0.9785	0.1720

Tableau 4-2 : Taux de validation et de test pour toutes les architectures et ce pour 200 époques, un batch size de 120 et avec l'optimiseur Adam.

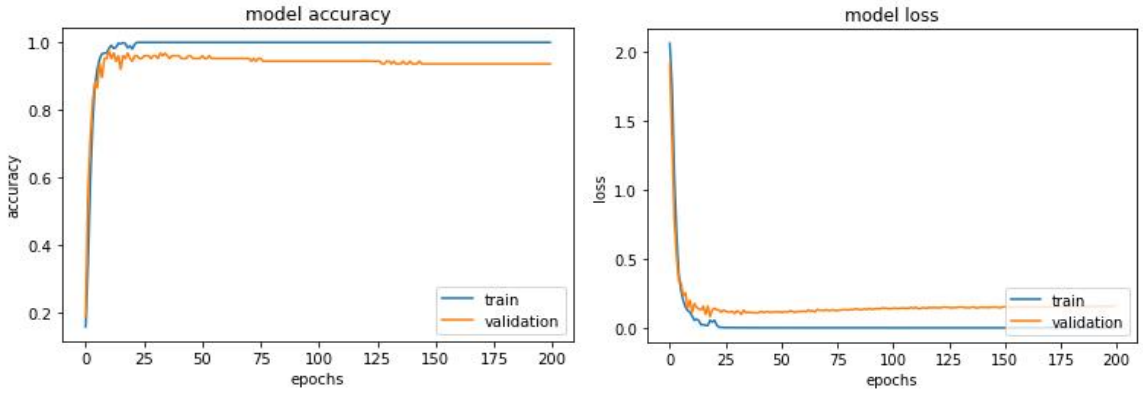

Figure 4-7 : Evolution des taux de validation pour LeNet_v1.

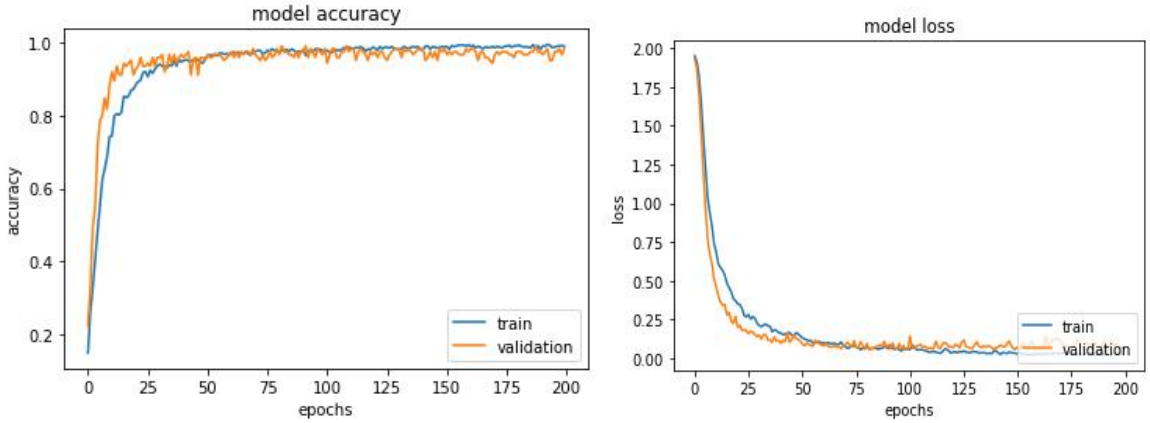

Figure 4-8 : Evolution des taux de validation pour LeNet_v11.

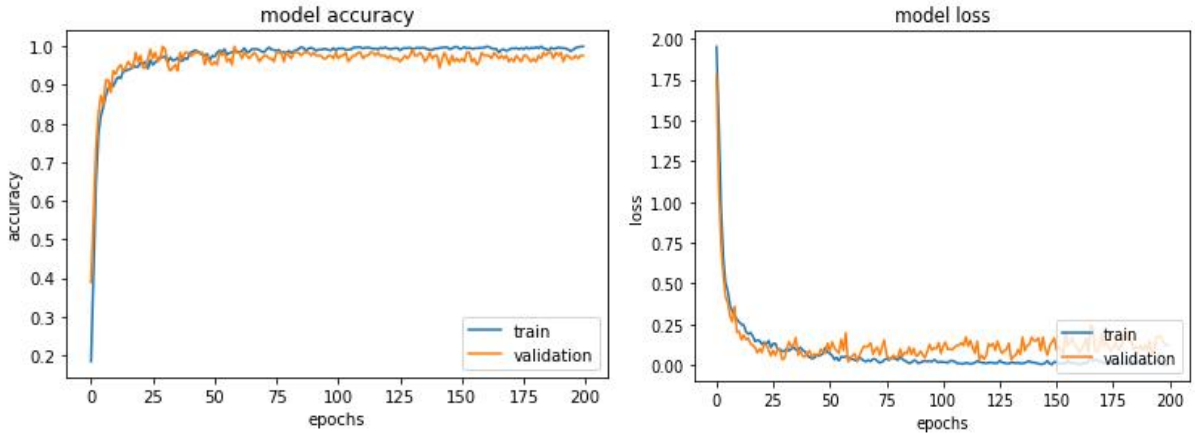

Figure 4-9 : Evolution des taux de validation pour LeNet_v2.

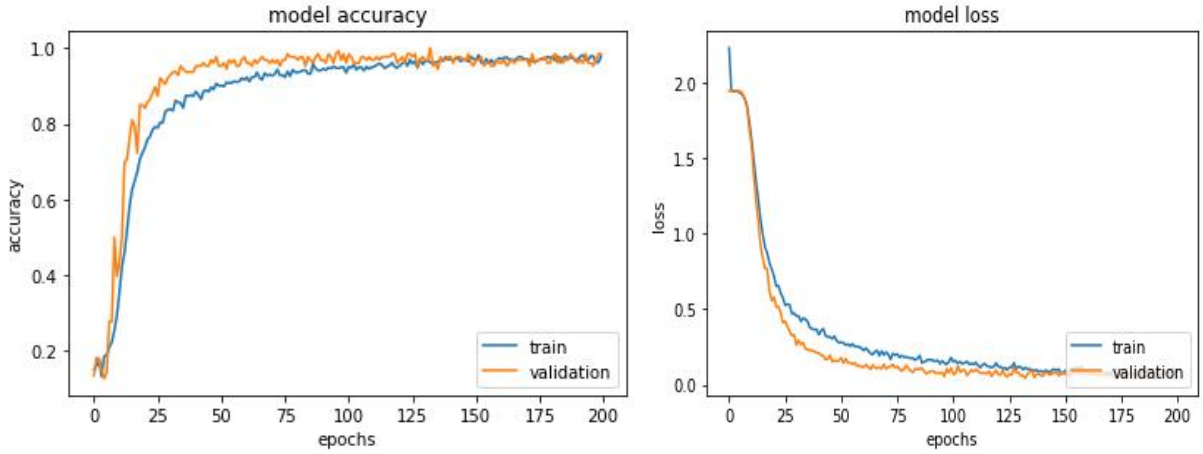

Figure 4-10 : Evolution des taux de validation pour LeNet_v22.

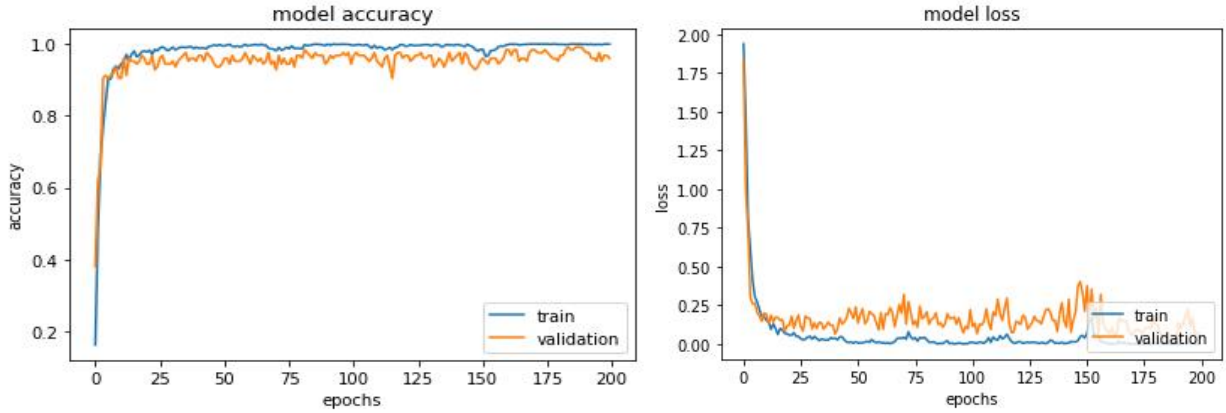

Figure 4-11 : Evolution des taux de validation pourVGG_v1.

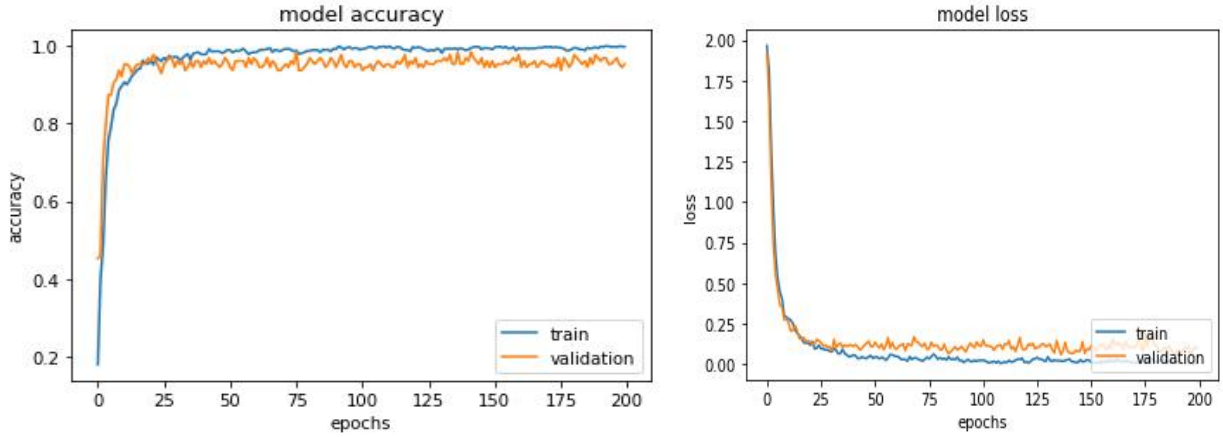

Figure 4-12 : Evolution des taux de validation pour VGG_v2.

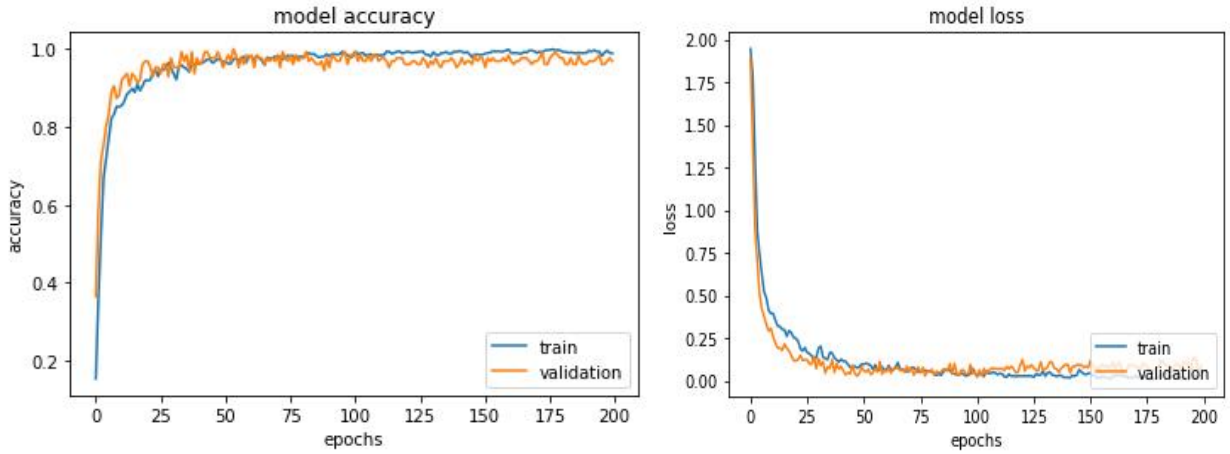

Figure 4-13 : Evolution des taux de validation pour VGG_v3.

4.5.3 Analyse des résultats

La partie la plus délicate dans l'implémentation d'un CNN est la conception de son architecture. En effet, selon le choix des composantes architecturales du CNN : le nombre de couches de convolution ainsi que de filtres, l'ajout ou non de couches dropout, le type de pooling (average pooling ou max pooling)…etc, on peu soit obtenir un setup dont l'erreur diminue selon le nombre d'époques, signe d'un bon modèle, soit l'inverse, dans ce cas, les performances seront restreintes à celles données durant les premières époques. Les architectures LeNet_v11, LeNet_v22, VGG_v2 et VGG_v3 ont été les plus performantes et c'est avec ces architectures qu'on étudiera l'influence des DB, du batch size ainsi que du choix de l'optimiseur.

On observe l'influence du nombre d'époques (nombre de passages complets dans l'ensemble de données d'apprentissage) qui plus il est grand, plus les performances de nos architectures s'améliorent. En outre, il ne doit ni être trop petit afin d'éviter le manque d'apprentissage, ni trop grand pour ne pas causé un sur-apprentissage [45].

Par ailleurs, on remarque que l'ajout de couches dropout ainsi que des couches ReLU ont amélioré les résultats. Ce constat est plus visible pour VGG_v1, l'ajout de ces dernières (pour obtenir l'architecture VGG_v2) ainsi que l'utilisation de l'average pooling au lieu du max pooling (pour obtenir l'architecture VGG_v3) ont nettement amélioré les résultats.

4.6 Modification de la DB

Dans cette section nous allons mettre en entrées des architectures choisies les trois bases de données citées dans le chapitre 3. Chacune d'elles contient un nombre d'images différent et une résolution différente, sachant que la RaFD est la plus récente et la plus riche, la JAFFE ne se compose que de modèles femmes et la Fer2013 contient plus de 35 000 images. Les résultats des test ainsi que l'évolution des taux de validation sont mentionnés ci-dessous.

4.6.1 Résultats

DB	RaFD		JAFFE		Fer2013	
Taux	val_acc	val_loss	val_acc	val_loss	val_acc	val_loss
LeNet_v11	0.9841	0.0582	0.7000	1.2928	0.4536	2.9483
LeNet_v22	0.9841	0.0409	0.6500	1.1894	0.5909	1.2118
VGG_v2	0.9524	0.1017	0.6500	1.6608	0.5445	1.9825
VGG_v3	0.9683	0.0904	0.6000	2.0257	0.5608	2.1461

Tableau 4-3 : Taux de validation pour LeNet_v1, LeNet_v2 et VGG_v2 et VGG_v3 pour les trois DBs : La RaFD, la JAFFE et la Fer2013, un nombre d'époques de 200, l'optimiseur Adam et un batch size de 120.

DB	RaFD		JAFFE		Fer2013	
Taux	Test_acc	Test_loss	Test_acc	Test_loss	Test_acc	Test_loss
LeNet_v11	0.9928	0.2473	0.9090	0.2647	0.4618	2.9455
LeNet_v22	0.9785	0.1901	0.7727	0.6212	0.5618	1.2431
VGG_v2	0.9714	0.1980	0.9090	0.4537	0.5559	1.9375
VGG_v3	0.9785	0.1720	0.7727	0.7599	0.5611	2.1534

Tableau 4-4 : Résultats des tests pour LeNet_v11, LeNet_v22, VGG_v2 et VGG_v3 pour les trois DBs : RaFD, JAFFE et Fer2013 en utilisant l'optimiseur Adam, un nombre d'époques de 200 et un batch size de 120.

Les figures suivantes illustrent l'évolution des taux de validation et d'entrainement de l'architecture LeNet_v22. On remarque nettement la différence entre les trois DB et la RaFD et celle qui a donné les meilleurs résultats.

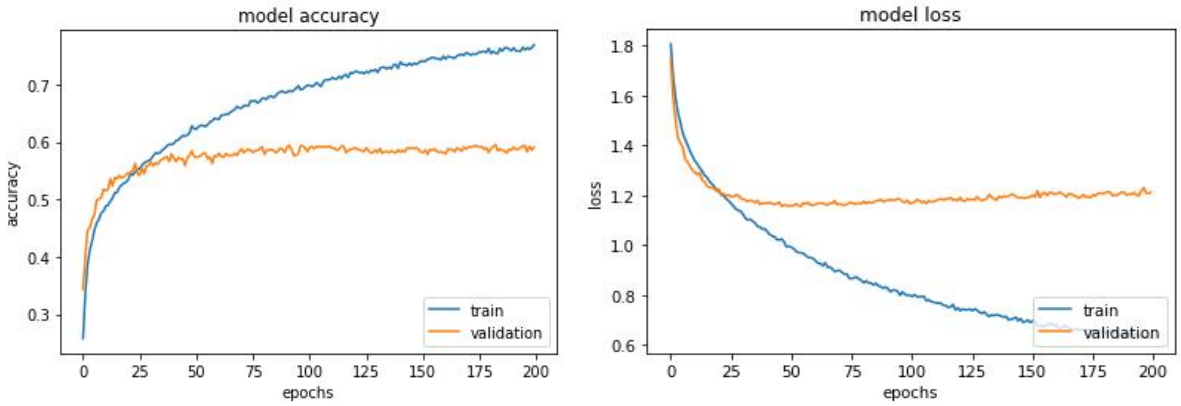

Figure 4-14 : Evolution des taux de validation et d'entrainement de l'architecture LeNet_v22 avec la Fer2013 pour 200 époques, l'optimiseur Adam et un batch size de 120.

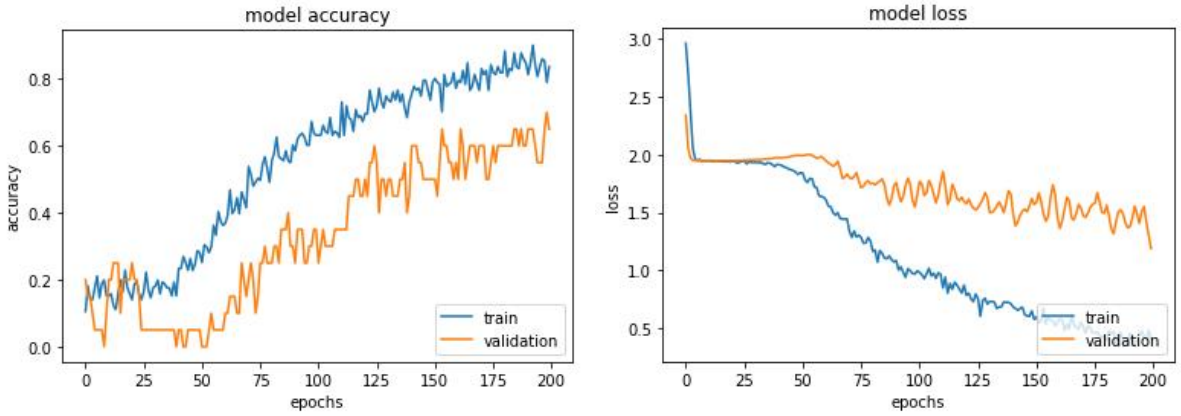

Figure 4-15 : Evolution des taux de validation et d'entrainement de l'architecture LeNet_v22 avec la JAFFE pour 200 époques, l'optimiseur Adam et un batch size de 120.

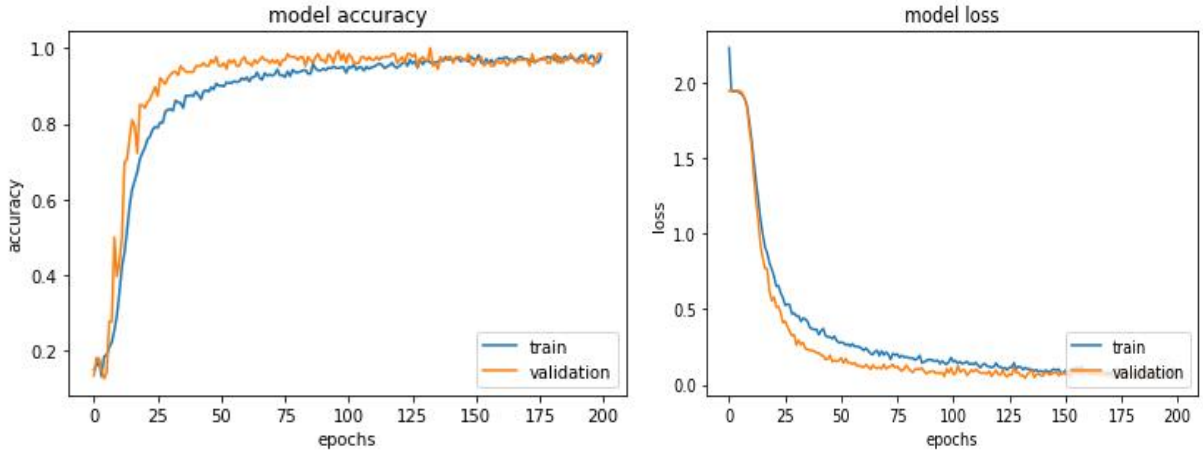

Figure 4-16 : Evolution des taux de validation et d'entrainement de l'architecture LeNet_v22 avec la RaFD pour 200 époques, l'optimiseur Adam et un batch size de 120.

4.6.2 Analyse des résultats

Le choix de la base de données est très important étant donné qu'il influence les résultats de classification. En effet le choix des modèles mimant les émotions dans une dataset est primordial et cela pour plusieurs raisons :

- **Le niveau de texture qui varie selon les visages et qui dépend de l'âge :** Un bébé dont la peau est lisse ne va pas exprimer ses émotions de la même manière qu'une personne âgée dont le visage est parsemé de rides [16].

- **L'origine des modèles :** La morphologie des visages varie selon les origines (comme les asiatiques qui ont des yeux petits) mais pas que, selon P.Ekman, l'expression des émotions varie aussi selon les cultures. Son expérience sur des japonais et des américains montrent que les japonais sont plus réservés en ce qui concerne l'expression de leurs émotions devant des étrangers, ce qui donne une dimension culturelle aux expressions faciales [16].

La base de données RaFD est celle qui nous a donné les meilleurs résultats. Comme dit dans le chapitre 2, les CNN extraits les caractéristiques de type FAC, et étant donné que les modèles de la RaFD ont été entrainés par des spécialistes du FAC, cela fait d'elle une DB très complète. En outre, les modèles sont d'âges différents, ce qui permet aux réseaux d'être encor plus performants.

4.7 Modification de l'optimiseur

Les optimiseurs jouent un rôle déterminant dans les performances des CNNs. En effet ce sont eux qui ont le rôle de mettre à jour les poids des réseaux selon une fréquence de mise à jour déterminée par le batch size. Le SGD est parmis les optimiseuers les plus anciens , nous allons pour notre part le teter à coté du Adam, du Rmsprop et du Adagrad.

4.7.1 Résultats

	Adam		SGD		Rmsprop		Adagrad	
Taux	Val_acc	Val_loss	Val_acc	Val_loss	Val_acc	Val_loss	Val_acc	Val_loss
LeNet_v11	0.9841	0.0582	0.8492	0.3728	0.9683	0.1048	0.9762	0.0813
LeNet_v22	0.9841	0.0409	0.8254	0.6836	0.9524	0.0732	0.9762	0.0831
VGG_v2	0.9524	0.1017	0.9603	0.1594	0.9444	0.1702	0.9683	0.0651
VGG_v3	0.9683	0.0904	0.9127	0.2511	0.9683	0.1204	0.9841	0.0443

Tableau 4-5 : Taux de validation pour LeNet_v11, LeNet_v22, VGG_v2 et VGG_v3 pour les optimiseurs Adam, SGD, Rmsprop et Adagrad pour la DB RaFD, un nombre d'époques de 200 et un batch size de 120.

	Adam		SGD		Rmsprop		Adagrad	
Taux	test_acc	test_loss	test_acc	test_loss	test_acc	test_loss	test_acc	test_loss
LeNet_v11	0.9928	0.2473	0.8428	0.4541	0.9714	0.3551	0.9642	0.2267
LeNet_v22	0.9785	0.1901	0.7714	0.7626	0.9714	0.1962	0.9642	0.1678
VGG_v2	0.9714	0.1980	0.9499	0.2181	0.9857	0.2939	0.9857	0.1929
VGG_v3	0.9785	0.1720	0.8857	0.3470	0.9928	0.1846	0.9857	0.2598

Tableau 4-6 : Résultats des tests pour LeNet_v11, LeNet_v22, VGG_v2 et VGG_v3 pour les optimiseurs Adam, SGD, Rmsprop et Adagrad pour la DB RaFD, un nombre d'époques de 200 et un batch size de 120.

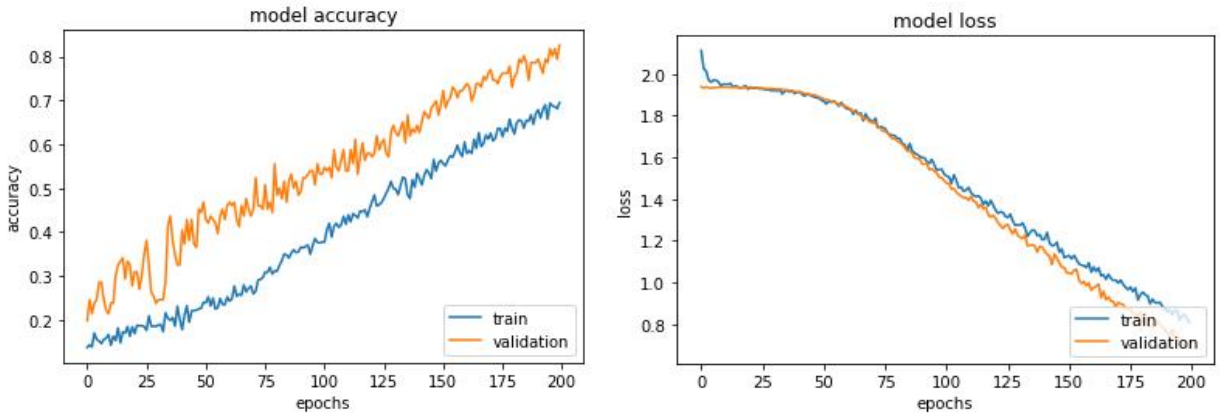

Figure 4-17 : Evolution des taux de validation et d'entrainement de l'architecture LeNet_v22 avec la DB RaFD, l'optimiseur SGD, pour 200 époques d'époques de 200 et un batch size de 120.

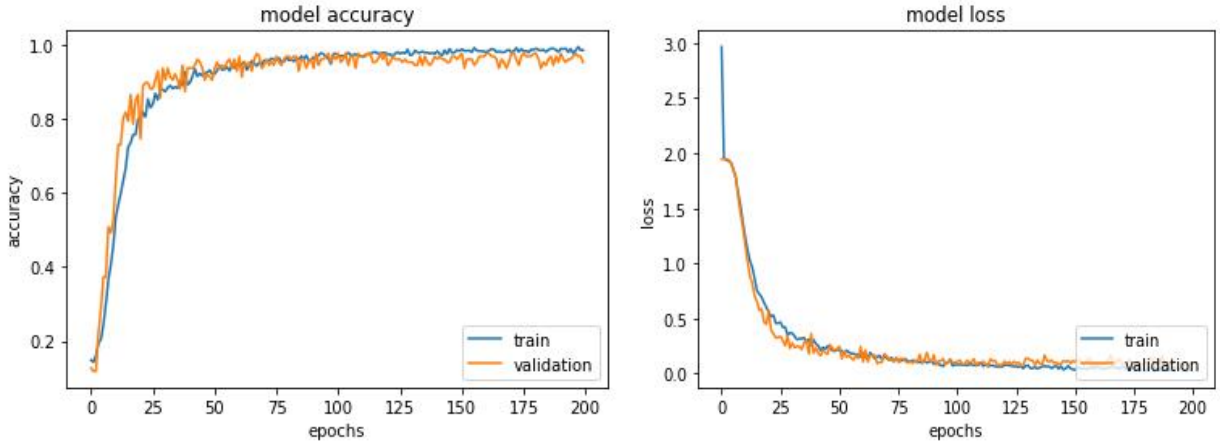

Figure 4-18 : Evolution des taux de validation et d'entrainement de l'architecture LeNet_v22 avec la DB RaFD, l'optimiseur Rmsprop, pour 200 époques d'époques de 200 et un batch size de 120.

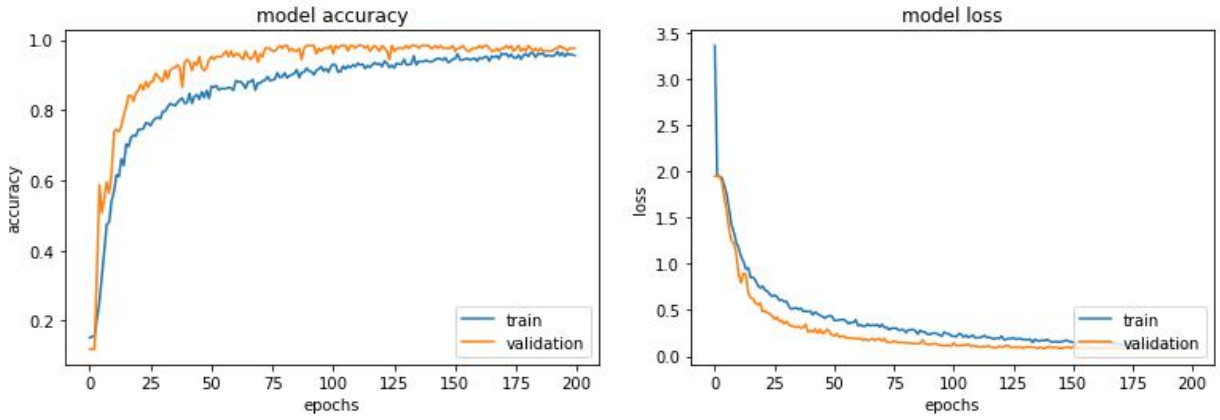

Figure 4-19 : Evolution des taux de validation et d'entrainement de l'architecture LeNet_v22 avec la DB RaFD, l'optimiseur Adagrad, pour 200 époques d'époques de 200 et un batch size de 120.

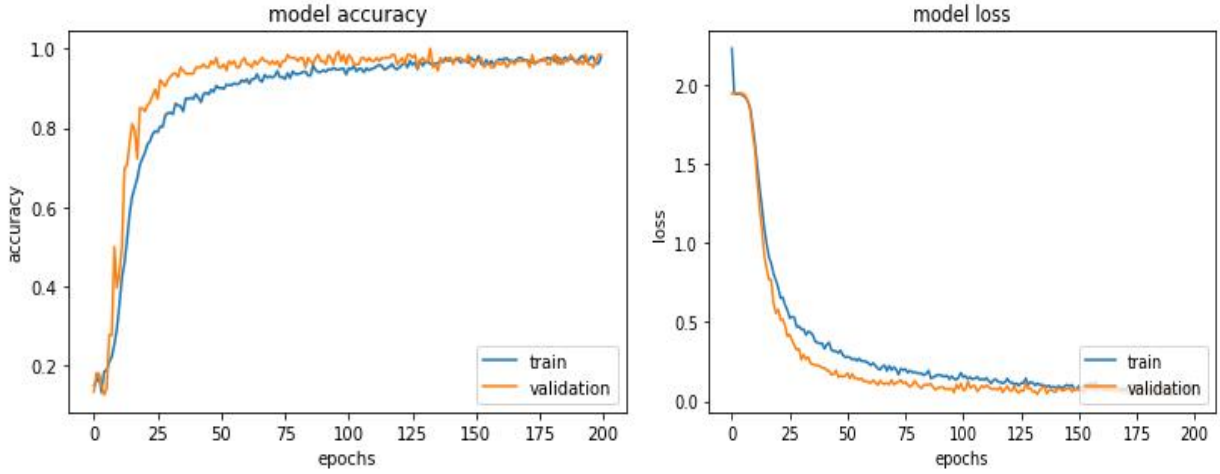

Figure 4-20 : Evolution des taux de validation et d'entrainement de l'architecture LeNet_v22 avec la DB RaFD, l'optimiseur Adam et pour 200 époques d'époques de 200 et un batch size de 120.

4.7.2 Analyse des résultats

L'optimiseur Adam est celui qui a fournit de meilleurs résultats, suivi du Rmsprop. Parmi les désavantages de Adagrad est la diminution de la capacité d'apprentissage en fonction de l'évolution de l'algorithme. La cause revient à la sommation des valeurs de gradients précédents, plus ya de gradient plus la sommation est grande et plus la valeur du quotient $\frac{\alpha}{\sqrt{\sum_i^t g_t^2}}$ tendra vers zéro.

Le Rmsprop quant à lui restreint cette sommation par l'accumulation non pas de tous les gradients depuis le début comme Adagrad, mais ne somme que les gradients les plus récents et c'est pour cela qu'il a donné de meilleurs résultats que Adagrad. Même chose pour Adam [39].

En somme, SGD est l'optimiseur à avoir donné les résultats les moins bons. On remarque aussi que c'est le plus lent à converger. Ce problème est du au fondement même de la méthode et revient souvent pour cet optimiseur [W14].

4.8 Modification du batch size

Le batch size est un hyper-paramètre qui contrôle le nombre d'échantillons d'entraînement à traiter avant la mise à jour des paramètres internes du modèle. A titre d'exemple, si le nombre de données d'entrainement est de cent, si le batch size est de vingt, la mise à jour des paramètres se fera cinq fois pour la même époque.

4.8.1 Résultats

Batch size	60		120		240	
Taux de validation.	val_acc	val_loss	val_acc	val_loss	val_acc	val_loss
LeNet_v11	0.9603	0.0891	0.9841	0.0582	0.9921	0.0205
LeNet_v22	0.9524	0.0928	0.9841	0.0409	0.9603	0.0893
VGG_v2	0.9286	0.2270	0.9524	0.1017	0.9762	0.0632
VGG_v3	0.9524	0.2075	0.9683	0.0904	0.9286	0.1073

Tableau 4-7 : Taux d'entrainement des quatre architectures pour la RaFD, 200 époques, l'optimiseur Adam et un batch size différent (60, 120, 240).

Batch size	60		120		240	
Résultats des tests	Test_acc	Test_loss	Test_acc	Test_loss	Test_acc	Test_loss
LeNet_v11	0.9857	0.2833	0.9928	0.2473	0.9642	0.1863
LeNet_v22	0.9785	0.2440	0.9785	0.1901	0.9857	0.1716
VGG_v2	0.9857	0.2381	0.9714	0.1980	0.9714	0.2302
VGG_v3	0.9785	0.3700	0.9785	0.1720	0.9857	0.2404

Tableau 4-8 : Résultats de test des quatre architectures pour la RaFD, 200 époques, l'optimiseur Adam et un batch size différent (60, 120, 240).

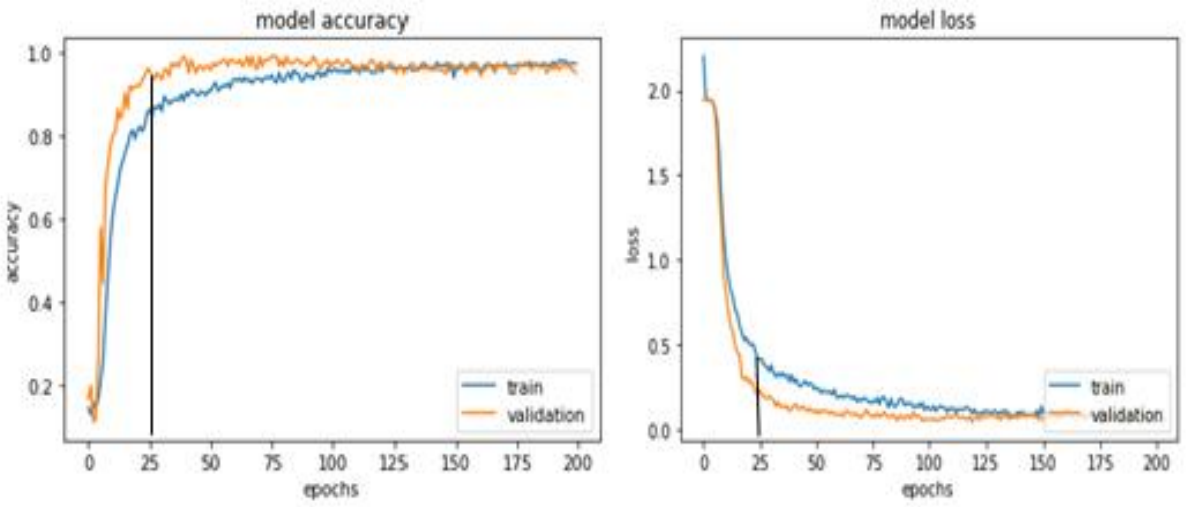

Figure 4-21 : Evolution des taux de validation de LeNet_v22 pour 200 époques, en utilisant la RaFD, l'optimiseur Adam et un batch size de 60.

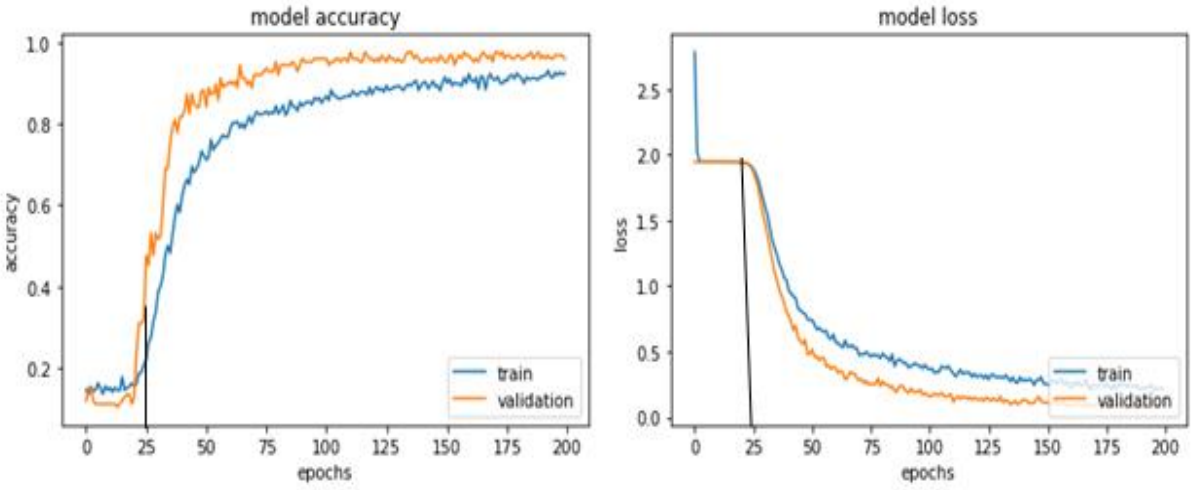

Figure 4-22 : Evolution des taux de validation de LeNet_v22 pour 200 époques, en utilisant la RaFD, l'optimiseur Adam et un batch size de 240.

4.8.2 Analyse des résultats

En comparant les résultats pour un batch size de 60 et un batch size de 240, nous observons deux changements :

- D'un coté nous remarquons aux deux graphs de l'évolution des taux d'entrainement et de validation pour un batch size de 60 varient rapidement.
- D'un autre coté, on observe que les performances finales sont en somme meilleures pour LeNet_v22 avec un batch size de 240, que pour LeNet_v22 avec un batch size de 60.

Le batch size joue donc sur l'apprentissage des réseaux, plus le batch size est petit, plus le réseau est entrainé. Le batch size de 60 et donc trop petit et a causé un sur-apprentissage du réseau. En somme, les résultats sont meilleurs pour batch size de 240 et de 120.

4.9 Conclusion

Dans cette section nous avons décrit les démarches à suivre afin d'obtenir une architecture convolutive la plus performante possible. Nous avons modifié plusieurs paramètres un à un afin d'observer au mieux l'influence de chacun d'eux sur les résultats de validation ainsi que sur ceux des tests. Même si les quatre réseaux utilisés ont montré de très bonnes performances (erreurs quasi nulles lors de l'apprentissage), l'erreur reste présente pour les tests même si elle reste raisonnable. Cela dit, pour les architectures choisies d'être le plus fidèles possible aux architectures originales (VGG16 et LeNet), ce sont les meilleures performances que l'on puisse obtenir d'elles, tout en gardant un nombre d'époques moyen mais assez grand pour voir l'évolution des performances de nos architectures.

En partant donc des architectures originales, les meilleures performances que l'on puisse obtenir pour les modèles choisis, toutes modifications confondues sont :

	Val_acc	Val_loss	Test_acc	Test_loss
LeNet_v11	0.9841	0.0582	0.9928	0.2473
LeNet_v22	0.9603	0.0893	0.9857	0.1716
VGG_v2	0.9524	0.1017	0.9714	0.1980
VGG_v3	0.9683	0.0904	0.9785	0.1720

Tableau 4-9 : Taux de validation et de test des architectures les plus performantes.

Conclusion générale

Le travail que nous avons présenté traite en somme du domaine de la reconnaissance des émotions en utilisant des architectures convolutives capables d'extraire les expressions faciales afin de les affilier à la classe d'émotion de base correspondante. Pour ce fait, nous avons testé quatre types d'architectures : Deux inspirées de l'architecture LeNet, deux inspirées de l'architecture VGG16 et les résultats qu'on a obtenus ont été très satisfaisants, du moins pour la base de données RaFD. Au-delà du type d'architecture utilisée, nous avons constaté que plusieurs paramètres influencent les performances d'un réseau de neurones, à savoir : la base de données utilisée, le nombre d'époques, le batch size ainsi que le choix de l'optimiseur. Néanmoins, on a aussi observé que le temps d'apprentissage était lent, surtout quand le modèle était volumineux et quand la base de données était grande comme c'était le cas pour la Fer2013, une des solutions pour ce fait, est l'exécution du programme sur Google Collaboratory qui fourni un accès libre à des GPUs accélérant le temps de calcul.

Notre programme est satisfaisant, mais il n'est pas parfait et peut être amélioré de plusieurs manières, pour ce fait nous prévoyons les perspectives suivantes :

- Entrainer les différentes architectures avec d'autre bases de données comme la CK+ afin d'améliorer leurs performances.

- Le système mis au point permet de détecter les émotions que pour des images dont les visages sont de face, une des solutions est d'entrainer notre programme sur une base de données différentes contenant des illustrations de visages sous tous les angles et avec toutes les rotations possibles ce qui induit l'utilisation d'une méthode de détection de visage autre que l'algorithme de Viola et Jones.

- Même si les émotions de base sont universelles, les expressions faciales peuvent avoir une relation avec la culture, i.e, une personne peut mimer de fausses expressions faciales afin de dissimuler ou de simuler une émotion pour être en accord avec les codes de conduite dictés par la société. Pour combler à ce manque, nous proposons la conception d'un système de reconnaissance bimodal alliant notre système de reconnaissance faciale à une méthode de mesure des signaux physiologiques comme le rythme cardiaque.

Bibliographie

[1] W.Handouzi, "Traitement d'information mono-source pour la validation objective d'un modèle d'anxiété: Application au signal de pression sanguine volumique", 2014.

[2] F.abdat," Reconnaissance automatique des émotions par données multimodales: Expression faciale et signaux physiologiques", 2010.

[3] J.F.Coget, C.Haag, et A.M.Bonnefous, "Le rôle de l'émotion dans la prise de décision intuitive : Zoom sur les réalisateurs-décideurs en période de tournage".

[4] M.Courgeon, MARC, " Modèles Informatiques des Emotions et de leurs Expressions Faciales pour l'Interaction Homme-Machine Affective Temps Réel".

[5] M.Batty, M.J.TAYLOR, "Early processing of the six basic facial emotional expressions", 2003.

[6] WEHRLE, Thomas, et al, "Studying the Dynamics of Emotional Expression Using SynthesizedFacial Muscle Movements".

[7] M.Guidetti,"L'expression vocale des émotions : Approche interculturelle et développementale", 1991.

[8] T. Bänziger, D. Grandjean, P. J. Bernard, G. Klasmeyer & K. R. Scherer, "Prosodie de l'émotion : Etude de l'encodage et du décodage"

[9] Y.Attabi," Reconnaissance automatique des émotions à partir du signal de parole, 2015.

[10] R. A. Calvo, S. D'Mello, J. M. Gratch, et A. Kappas, The Oxford Handbook of Affective Computing. Oxford University Press", 2015.

[11] S.Gil, "Comment étudier les émotions en laboratoire ?", 2009.

[12] A.Nugier," Histoire et grands courants de recherche sur les émotions", 2009.

[13] M. Turk et A. Pentland, "Eigenfaces for recognition ", 1991.

[14] K. Mahboub, "Modélisation des processus émotionnels dans la prise de décision ",2012

[15] Vincent Roy, "Psychologie des émotions".

[16] S.Gharsalli," Reconnaissance des émotions par traitement d'images", 2016

[17] M.Gendre,"Influence des émotions sur l'organisation biomécanique des mouvements volontaires d'approches et d'évitement, cas de l'initiation du pas et de l'élévation latérale de la jambe", 2015.

[18] I.Bakker, Theo van der Voordt, Peter Vink , Jan de Boon, Pleasure," Arousal, Dominance: Mehrabian and Russell revisited", 2014.

[19] Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth," The World of Emotions Is Not Two-Dimensional", 2007.

[20] K.Ghanem,"Reconnaissance des Expressions Faciales à Base d'Informations Vidéo ; Estimation de l'Intensité des Expressions Faciales", 2010.

[21] K.Lekdioui,"Reconnaissance d'états émotionnels par analyse visuelle du visage et apprentissage machine", 2018.

[22] Shichuan Du, Yong Tao, and Aleix M. Martinez," Compound facial expressions of emotion", 2013.

[23] P. Philippot, "Emotion et psychothérapie", 2007.

[24] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, et Thomas S. Huang," A survey of affect recongnition methods :Audio, visuals and spontaneous expressions", 2009.

[25] Y.Lecun, B.Broser, J.S.Denker, D.Henderson, R.E.Howard, W.HUBBARD, L.D.Jackel, "Backapropagation applied to handwritten zip code recongnition", 1989.

[26] C.Szegedy, W.Liu, Y.Jia, P.Sermanet, S.Reed, D.Anghelov, D. Erhan, V.VANHOUCKE, A.Rabinovich, "Going deeper with convolution", 2014.

[27] Y.Bengio," Learning Deep Architectures for AI", 2009.

[28] G. Gelly, "Réseaux de neurones récurrents pour le traitement automatique de la parole ", 2017

[29] A. Graves , N. Jaitly, " Towards end-to-end speech recognition with recurrent neural networks ", 2014.

[30] A.Khan, A.Sohail, U. Zahoora, and A. S.Qureshi," A Survey of the Recent Architectures of Deep Convolutional Neural Networks".

[31] E.Kauderer-Abrams,"Quantifying Translation-Invariance in Convolutional Neural Networks",2017.

[32] P.Buyssens, A.Elmoataz, "Réseaux de neurones convolutionnels multi-échelle pour la classifiation cellulaire", 2016.

[33] Y. LeCun, Y. Bengio, et G. Hinton, "Deep learning ", 2015.

[34] H.Lee, R.Grosse ,R.Ranganath, A.Y. Ng, "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations", 2009.

[35] J.-C. Vialatte, "On convolution of graph signals and deep learning on graph domains", 2018.

[36] S. Lai, L. Xu, K. Liu, et J. Zhao,"Recurrent convolutional neural networks for text classification", 2015.

[37] N.Srivastava, G.Hinton, A. Krizhevsky, I.Sutskever, R.Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", 2014.

[38] L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, et B. An, "Can Cross Entropy Loss Be Robust to Label Noise? ", 2020.

[39] D. P. Kingma et J. Ba, "Adam: A Method for Stochastic Optimization", 2014.

[40] G.Hinton, N.Srivastava, K.Swersky, "Lecture 6a Overview o f mini - batch gradient descent".

[41] L.Bottou," Large-Scale Machine Learning with Stochastic Gradient Descent", 2010.

[42] I. Dagher, E. Dahdah, et M. Al Shakik, "Facial expression recognition using three-stage support vector machines ", 2019.

[43] O.Langner, R.Dotsch, G.Bijlstra, and D.H.J.Wigboldus, S.T. Hawk, Ad van Knippenberg, "Presentation and validation of the Radboud Faces Database", 2010.

[44] J. Lever, M. Krzywinski, et N. Altman," Points of significance: model selection and overfitting", 2016.

[45] A. Chamekh, "Optimisation des procédés de mise en forme par les réseaux de neurones artificiels ", 2008.

Webographie

[W1] « Muscles du visage - pour les nuls ».https://godiche.ru/education-et-langues/la-science/anatomie/19366-muscles-de-la-face.html (consulté le juill. 31, 2020).

[W2] « Le nerf facial ». http://www.medecine.unige.ch/ enseignement /apprentissage/ module3/ pec/ apprentissage /neuroana/facial/facial3.htm (consulté le juill. 31, 2020).

[W3] « Système nerveux autonome. Schéma montrant les deux parties et les deux... | Download Scientific Diagram », ResearchGate. https://www.researchgate.net/figure/Systeme-nerveux-autonome-Schema-montrant-les-deux-parties-et-les-deux-voies-du-SNA-A_fig4_267507690 (consulté le juill. 31, 2020).

[W4] « Figure 1. A-V emotion plane-Russell's circumplex model (Russell, 1980). », *ResearchGate*. https://www.researchgate.net/figure/A-V-emotion-plane-Russells-circumplex-model-Russell-1980_fig1_324807629 (consulté le août 09, 2020).

[W5] K. Bai, « A Comprehensive Introduction to Different Types of Convolutions in Deep Learning », Medium, févr. 11, 2019. https://towardsdatascience.com/a-comprehensive-introduction-to-different-types-of-convolutions-in-deep-learning-669281e58215 (consulté le mai 01, 2020).

[W6] V. Bushaev, « Understanding RMSprop — faster neural network learning », Medium, sept. 02, 2018. https://towardsdatascience.com/understanding-rmsprop-faster-neural-network-learning-62e116fcf29a (consulté le sept. 21, 2020).

[W7] « Figure 13-Architecture du réseau convolutionnel LeNet-5 pour la... », *ResearchGate*. https://www.researchgate.net/figure/Architecture-du-reseau-convolutionnel-LeNet-5-pour-la-reconnaissance-de-caracteres-69_fig12_324937802 (consulté le oct. 23, 2020).

[W8] F. Kınlı, « [Deep Learning Lab] Episode-3: fer2013 », Medium, mars 15, 2019. https://medium.com/@birdortyedi_23820/deep-learning-lab-episode-3-fer2013-c38f2e052280 (consulté le août 15, 2020).

[W9] « OpenCV ». http://www.ai.univ-paris8.fr/~chalencon/Vision/openCV.html (consulté le sept. 10, 2020).

[W10] « Plongez en détail dans la librairie NumPy »,OpenClassrooms. https:// openclassrooms.com/ fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science/4740941-plongez-en-detail-dans-la-librairie-numpy (consulté le sept. 10, 2020).

[W11] « NumPy ». https://numpy.org/ (consulté le sept. 10, 2020).

[W12] « Matplotlib: Python plotting — Matplotlib 3.3.1 documentation ». https://matplotlib.org/index.html (consulté le sept. 10, 2020).

[W13] P. SCHWARTZ, « Scikit-learn, une bibliothèque de machine learning- », Developpez.com. https://khayyam.developpez.com/articles/machine-learning/scikit-learn/ (consulté le sept. 16, 2020).

[W14] « Efficient Distributed SGD with Variance Reduction - IEEE Conference Publication ». https://ieeexplore.ieee.org/abstract/document/7837835 (consulté le oct. 16, 2020).

Résumé

Une expression faciale est une combinaison de mouvements faits grâce à la détente ainsi qu'à la contraction des muscles du visage, elle est une composante essentielle de la communication non verbale et traduit constamment l'état émotionnel de l'interlocuteur. Nous nous intéresserons aux six émotions de base décrites par EKMAN à savoir : La joie, la tristesse, la colère, le dégout, la peur ainsi que la surprise, à cela nous ajouterons le neutre. Ce travail propose de faire une reconnaissance automatique des émotions en utilisant des réseaux de neurones basés sur la convolution. Une étude comparative de différentes architectures convolutives a été faite, ayant fait leurs preuves dans le domaine de la reconnaissance d'images, quatre architectures inspirées de LeNet et VGG seront testées et optimisées. Des résultats prometteurs ont été obtenus, soit plus de 97 % de précision et ceux pour les quatre architectures.

Mots clé : Emotion, apprentissage profond, benchmark.

Abstract

A facial expression is a combination of movements made through relaxation as well as the contraction of the muscles of the face, it is an essential component of non-verbal communication and constantly reflects the emotional state of the interlocutor. We will be interested in the six basic emotions described by Ekman, namely: Joy, sadness, anger, disgust, fear as well as surprise, to which we will add the neutral. This work proposes to make an automatic recognition of emotions using neural networks based on convolution. A comparative study of different convolutional architectures has been made, having proved their worth in the field of image recognition, four architectures inspired by LeNet and VGG will be tested and optimized. Promising results were obtained, that is to say more than 97% of precision for the four architectures.

Key words : Emotion, deep learning, benchmark.

نبذة مختصرة

تعبيرات الوجه هي مزيج من الحركات التي يتم إجراؤها من خلال الاسترخاء وكذلك تقلص عضلات الوجه، وهي عنصر أساسي للتواصل غير اللفظي وتعكس باستمرار الحالة العاطفية للمحاور. سنهتم بالمشاعر الست الأساسية التي وصفها إيكمان وهي: الفرح ،الحزن ،الغضب ،الاشمئزاز ،الخوف وكذلك المفاجأة ، والتي سنضيف إليها الحياد. يقترح هذا العمل إجراء التعرف التلقائي على المشاعر باستخدام الشبكات العصبية على أساس التعلم العميق. تم إجراء دراسة مقارنة للبنى المختلفة ، بعد أن أثبتت قيمتها في مجال التعرف على الصور ، سيتم اختبار وتحسين أربعة هياكل مستوحاة من LeNet و VGG. تم الحصول على نتائج واعدة ، أي أكثر من 97٪ من الدقة وتلك الخاصة بالبنى الأربعة.

الكلمات المفتاحية: العاطفة ، التعلم العميق, تحليل مقارن.