

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abderrahmane Mira de Bejaia



Faculté de Technologie

Département de Génie Electrique

## Mémoire de fin d'études

En vue de l'obtention du diplôme de Master

En Automatique

Spécialité : Automatique et informatique industrielle

Thème :

*Reconnaissance Automatique des  
Expressions Faciales*

**Réalisé par :**

*M<sup>elle</sup> NOUIOUA Naouel*

*Mr MADI Moussa*

**Encadré par :**

*M<sup>me</sup> MEZZAH Samia*

***Devant le jury :***

*Dr. M. SADJI*

*Dr. A. ALLICHE*

**Année universitaire : 2019/2020**



# Remerciements

**Aucune œuvre humaine ne peut être réalisée sans la contribution d'autrui. Il est donc Important pour nous d'adresser nos sincères remerciements à toutes les personnes qui nous ont permises de réaliser notre mémoire.**

Nos remerciements à priori s'orientent envers Allah le Tout Puissant et le Miséricordieux qui nous a permis, par sa grâce et sa clémence, de poursuivre à terme la réalisation de ce travail.

On tient à exprimer nos profondes gratitudee à notre promotrice, Madame, pour toute l'attention qu'elle nous a apportée, Nous lui témoignons toute notre reconnaissance pour ses conseils, ses orientations et sa patience.

Il ne sera jamais remercié suffisamment le corps enseignant de la faculté Génie Electrique de l'Université Abderrahmane Mira de Bejaia, Nos vifs remerciements aux membres de jury de bien vouloir accepter d'évaluer notre travail.

Ainsi que tous ceux, famille(s) et amis(es) qui nous ont encouragés et soutenus pour réaliser ce travail.

# Dédicaces

## Je dédie ce mémoire à :

Ma très chère maman qui m'a soutenue tout au long de mon cursus scolaire, que je remercie énormément pour sa présence, son attention, ses conseils et encouragements, et que dieu la protège.

Mon frère Bachir.

Ma grande famille « grand et petit ».

Ma binôme Naouel ainsi que toute sa famille.

Mes chers amis Sofiane et Djalil

Mes chers amis (es) ainsi que leurs familles.

Toute la promotion Automatique et Système.

Toutes les personnes que je porte dans mon cœur et qui ont, sans le savoir participer de manière considérable à ma réussite.

À ceux que ma plume a oublié....

*Moussa. M*

# Dédicaces

Je dédie ce mémoire à mes parents auxquelles je dois ce que je suis aujourd'hui, grâce à leur amour, patients et innombrables sacrifices, je prie Dieu qu'ils restent toujours fiers de moi.

À mon frère et ma sœur pour tout l'amour que vous m'avez apporté, j'espère que je serais toujours le bon exemple de sœur aînée.

Mes chères tantes et oncles, cousins et cousines qui ont toujours été à mes côtés pour la joie comme la tristesse.

Ma grande famille « grand et petit ».

Mon binôme Moussa et sa chère famille, qui a été avec moi le long du chemin, merci pour ton soutien et ton encouragement.

À mes amis qui ont fait partie des plus beaux souvenirs qui ont marqué une étape importante de ma vie.

À tous ceux qui par un mot m'ont donné la force de continuer, je vous remercie de tout cœur.

*Naouel N*

## Liste des tableaux et figures

### Liste des tableaux

<b>Tableau I.1</b> : Les AUs fréquentes décodées par le manuel FACS sur la base d'images CK+ ..	7
<b>Tableau I.2</b> : Combinaison des AUs correspondant à chaque émotion .....	7
<b>Tableau III.1</b> : Détails de la base de données utilisée.....	27
<b>Tableau III.2</b> : Probabilités des 7 classes.....	38

### Liste des figures

<b>Figure I.1</b> : Les muscles du visage.....	4
<b>Figure I.2</b> : Diversité des expressions faciales.....	4
<b>Figure I.3</b> : Exemples d'expressions émotionnelles prototypiques.....	6
<b>Figure I.4</b> : Muscles impliqués dans l'activation de (a) AU1 et (b) AU2 .....	7
<b>Figure I.5</b> : AUs actives pour quelques expressions faciales .....	8
<b>Figure I.6</b> : Composantes du système d'analyse des expressions faciales .....	9
<b>Figure I.7</b> : Les caractéristiques d'apparence .....	11
<b>Figure II.1</b> : Neurone biologique (a) et neurone artificiel (b).....	15
<b>Figure II.2</b> : Réseau de neurones artificiels.....	15
<b>Figure II.3</b> : Architecture typique d'un réseau de neurone convolutif .....	17
<b>Figure II.4</b> : Exemple de résultat de la convolution d'une image d'entrée par 4 filtres. ....	17
<b>Figure II.5</b> : Exemple illustrant l'opération de convolution 3x3 .....	17
<b>Figure II.6</b> : Exemple de zéro padding.....	18
<b>Figure II.7</b> : Exemples de calcul du pooling sur une image 4x4.....	19
<b>Figure II.8</b> : Techniques de Transfer Learning.....	21
<b>Figure II.9</b> : Approches de transfer learning en Deep Learning.....	22
<b>Figure II.10</b> : Évolution des CNN pour la reconnaissance d'images ImageNet 2012 .....	23
<b>Figure III.1</b> : Architecture de ResNet50 .....	28
<b>Figure III.2</b> : Exemple de coefficient de 96 filtres de convolution entraînés de taille égale à 11x11x3 illustrant le concept d'apprentissage d'attributs .....	29
<b>Figure III.3</b> : Stratégie de tranfer learning testées .....	30
<b>Figure III.4</b> : Test de performances de quelques valeurs d'hyperparamètres .....	32
<b>Figure III.5</b> : Courbe du Fine tuning de ResNet.....	33
<b>Figure III.6</b> : Matrice de confusion de ResNet50 utilisé.....	34
<b>Figure III.7</b> : Quelques exemples de classification.....	35
<b>Figure III.8</b> : Image test.....	35
<b>Figure III.9</b> : représentation graphique des couches de la première section sous Matlab .....	35

<b>Figure III.10</b> : Les filtres de convolution de la section 1 .....	<b>36</b>
<b>Figure III.11</b> : Cartes de caractéristiques générées par les couches de la section 1. ....	<b>36</b>
<b>Figure III.12</b> : Cartes de caractéristiques des sections (a) 2, (b) 3, (c) 4 et 5 (d). ....	<b>37</b>
<b>Figure III.13</b> : Représentation graphique de vecteur de caractéristiques.....	<b>37</b>
<b>Figure III.14</b> : Courbe d'entraînement de ResNet (FC seulement) .....	<b>39</b>
<b>Figure III.15</b> : Comparaison entre les performances des stratégies d'apprentissage testées ..	<b>39</b>

# Table des matières

Remerciements.....	3
Dédicaces.....	4
Liste des tableaux et figures .....	6
Chapitre  I.....	2
I.1. Introduction.....	3
I.2. Physiologie et rôle des expressions faciales .....	3
I.3 Applications pratiques de la reconnaissance d’expressions faciales .....	5
I.4. Description et analyse des expressions faciales .....	5
I.5. Approches de reconnaissance automatique d’expressions faciales.....	9
I.5.1 Approches conventionnelles .....	9
II.5.1.1 Acquisition d’images et extraction du visage .....	10
I.5.1.2 Extraction des caractéristiques .....	10
I.5.1.3 Classification.....	11
I.5.1 Approches basées sur l’apprentissage profond .....	11
I.6 Conclusion.....	12
Chapitre II.....	13
II.1. Introduction .....	14
II.2. Principe du deep learning .....	14
II.3 Réseaux de neurones convolutifs (CNN) .....	16
II.3.1 La couche de convolution.....	16
II.3.3 La couche de pooling .....	19
II.3.4 Couches entièrement connectées.....	20
II.4 L’apprentissage d’un CNN.....	20
II.5 L'apprentissage par transfert.....	21
II.5.1 Stratégies de transfer learning.....	21
II.5.2 Quelques réseaux convolutifs pré-entraînés .....	23
Conclusion .....	25
Chapitre  III.....	26
III.1. Introduction.....	27
III.2. Base de données .....	27
III.3. Le réseau pré-entraîné choisi.....	28
III.4. Environnement du travail.....	31
III.5. Pseudo-code d’entraînement .....	33
III.6. Résultats des expériences .....	34
Conclusion .....	44
Conclusion générale .....	45
Bibliographie .....	71
Résumé.....	74

# **Introduction générale**

## Introduction générale

Notre visage est une partie complexe et hautement différenciée de notre corps. C'est un système de signalisation et de communication très complexe qui comprend plusieurs muscles autonomes structurellement et fonctionnellement. L'activité des muscles faciaux est fortement spécialisée pour l'expression, cela permet aux humains de partager des informations sociales entre eux et communiquer à la fois verbalement et non verbalement.

En raison de l'importance de la reconnaissance de l'expression faciale dans de nombreux domaines, tels que l'interaction homme-machine, le marketing, la médecine, la sécurité et l'éducation, divers algorithmes d'apprentissage automatique ont été développés. Le succès actuel des réseaux de neurones convolutifs (Convolutional Neural Networks CNN) dans la classification d'images s'est étendu au problème de la reconnaissance de l'expression faciale. Contrairement aux méthodes traditionnelles d'apprentissage par la machine où les caractéristiques sont définies par des méthodes traditionnelles, CNN apprend à les extraire directement de la base de données d'apprentissage.

L'objectif de ce projet de fin d'études est d'appliquer l'apprentissage profond à travers un CNN pour la reconnaissance des expressions faciales. Nous utilisons un réseau pré-entraîné appelé ResNet qui est un réseau convolutif profond initialement développé pour la classification d'images d'objets appartenant à 1000 classes différentes. Nous utilisons également l'approche d'apprentissage par transfert pour adapter le réseau initial à notre problème de classification des expressions faciales.

Ce mémoire est organisé en trois chapitres. Dans le premier chapitre, nous présentons les principales méthodes utilisées pour la description, l'analyse et la reconnaissance automatique des expressions faciales. Dans le second chapitre, nous présentons un aperçu sur le Deep Learning dans un premier lieu, ensuite nous détaillons les composantes des réseaux de neurones convolutifs et leurs architectures et les techniques qui permettent d'adapter un **réseau** pré-entraînés pour un problème de classification donné à un autre problème similaire. Le dernier chapitre est consacré à la description des différents outils utilisés dans notre étude, ainsi que les différents résultats obtenus.

# **Chapitre I**

## **Reconnaissance Automatique des Expressions Faciales**

## I.1. Introduction

Au cours de la dernière décennie, les recherches en vision par ordinateur ont montré beaucoup d'intérêt pour l'analyse et la reconnaissance automatique des expressions faciales dans des vidéos ou des images statiques. La plupart des méthodes d'analyse des expressions faciales tentent de classer les expressions en quelques grandes catégories émotionnelles, telles que la joie, la tristesse, la colère, la surprise, la peur et le dégoût [1]. Ce chapitre présente un aperçu sur les diverses approches de la reconnaissance automatique des expressions faciales et leurs applications.

## I.2. Physiologie et rôle des expressions faciales

L'expression faciale est le moyen le plus expressif pour l'être humain pour véhiculer des signaux de communication non verbaux dans les interactions face à face. Les expressions faciales jouent, donc, un rôle important dans les relations humaines et régulent les interactions des personnes avec l'environnement et d'autres personnes [2]. Grâce à ces **expressions**, il est possible de faire plusieurs déductions et de récupérer plusieurs informations comme [1]:

- L'état affectif que ce soit les émotions (peur, colère, joie, surprise, tristesse, dégoût) ou bien certaines humeurs.
- L'activité cognitive comme la concentration, l'ennui ou la perplexité.
- Le tempérament et la personnalité.

Du point de vue physiologique, l'expression faciale est contrôlée par 44 muscles, répartis de chaque côté du visage (figure I.1) [1]. Ces muscles, également appelés muscles mimétiques, font partie du groupe des muscles de la tête, qui contiennent en outre des muscles du cuir chevelu, des muscles de la mastication responsables du déplacement de la mâchoire et de la langue. Les muscles du visage sont innervés par le nerf facial, qui se ramifie dans le visage et son activation provoque des contractions ce qui se traduit par divers mouvements observables. Les expressions faciales résultent des mouvements des muscles faciaux tirant la peau et changeant temporairement la forme des yeux, des sourcils et des lèvres, conduisant à l'apparition de plis, et de sillons dans différentes parties de la peau.

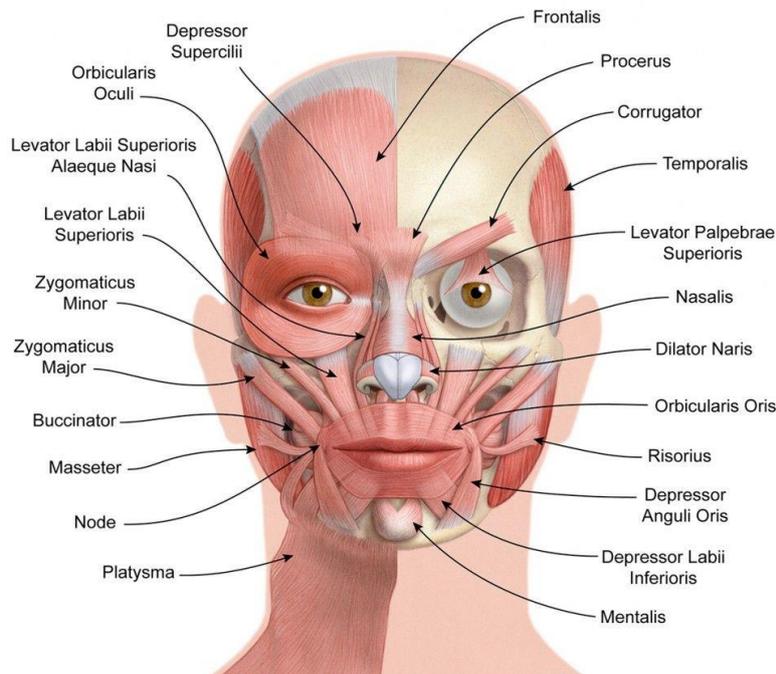


Figure I.1 : Les muscles du visage

L'enchevêtrement des muscles au niveau du visage permet une grande mobilité. Les points d'attache des muscles aux os du crâne sont relativement peu nombreux et beaucoup sont liés directement les uns aux autres. Les humains sont, en principe, capables de produire des milliers d'ensembles différents d'expressions faciales (Figure I.2). De plus, les muscles faciaux peuvent se contracter avec différentes intensités et un temps variable jusqu'à la contraction maximale, augmentant encore davantage le nombre de schémas de mouvement qu'un visage peut générer en principe [2].

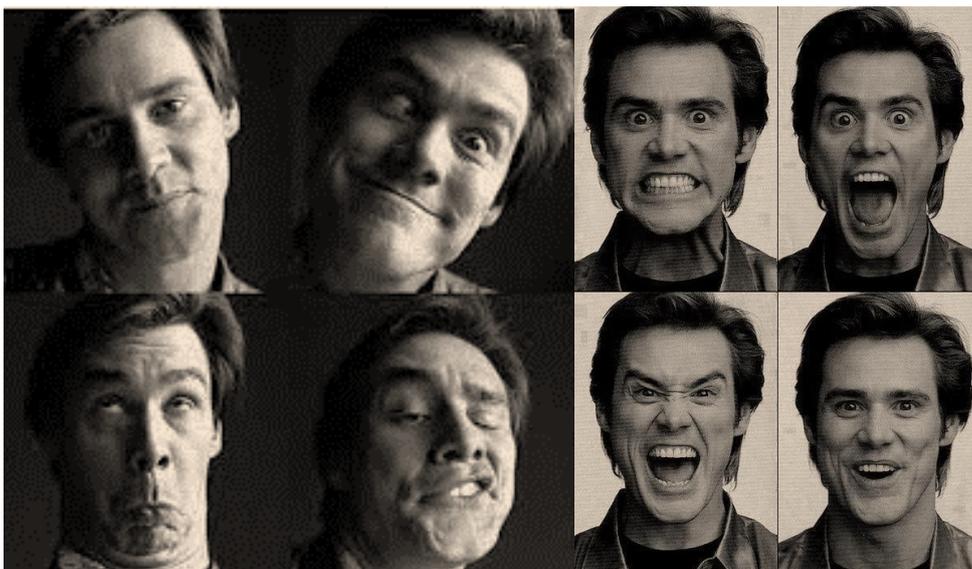


Figure I.2 : Diversité des expressions faciales

En pratique, cependant, il n'y a qu'un petit ensemble de configurations faciales distinctives que presque tout le monde associe à certaines expressions (tel que l'expression de joie), quels que soient le genre, l'âge, le profil social et culturel [1]. Ces expressions de base sont probablement dues à la façon dont le cerveau contrôle le visage ainsi qu'aux diverses contraintes anatomiques (par exemple, certains muscles sont plus ou moins susceptibles de se déplacer ensemble en raison de leur position relative les uns par rapport aux autres, ou de la manière dont ils sont attachés aux os du visage ou dont ils sont activés) [2].

Pour les types d'expressions faciales, on distingue entre les macros et les micros expressions [2]. Les macros expressions sont, généralement, évidentes à l'œil nu et durent entre 0,5 et 4 secondes et se produisent dans les interactions quotidiennes. Les micro expressions durent moins d'une demi-seconde et se produisent lorsqu'on essaye de dissimuler ou réprimer inconsciemment l'état émotionnel.

### **I.3 Applications pratiques de la reconnaissance d'expressions faciales**

L'analyse de l'impact de toute situation, contenu, produit ou service, censé susciter des réponses faciales volontaires ou involontaires, est d'un intérêt majeur, car cette analyse permet de détecter l'état cognitif et affectif de la personne impactée et par conséquent comprendre et même prédire ses intentions, ses actions et ses réactions. De ce fait, l'analyse des expressions faciales trouve de nombreuses applications dans des domaines divers tels que [1] :

1. Marketing : des applications pour mesurer la satisfaction des clients, prévoir les produits qui les intéressent.
2. Médecine : aide à la détection de certaines maladies psychologiques, aide à l'apprentissage des émotions pour les enfants autistes.
3. Sécurité : détection du stress et de comportements suspects.
4. Interaction Homme-Machine : robot d'accompagnement, voiture intelligente.
5. Éducation : apprentissage à distance et le développement de jeux interactifs.

### **I.4. Description et analyse des expressions faciales**

Basée sur le travail de Darwin ainsi que sur celui d'Ekman [3], la recherche dans le domaine de l'analyse des expressions faciale se concentre sur deux approches majeures pour décrire les expressions faciales qui sont :

- Description par message : vise à décrire les expressions faciales en termes d'un ensemble d'étiquettes affectives discrètes prototypiques telles que l'expression de joie, surprise, peur, tristesse, dégoût et colère (figure II.3), ou d'autres ensembles d'étiquettes émotionnelles.

- Description par signe : vise à décrire les expressions affichées en termes de composantes musculaires activées appelées unités d'action (Action Unit AU) en utilisant un système de description spécial appelé Facial Action Coding System (FACS) [4].

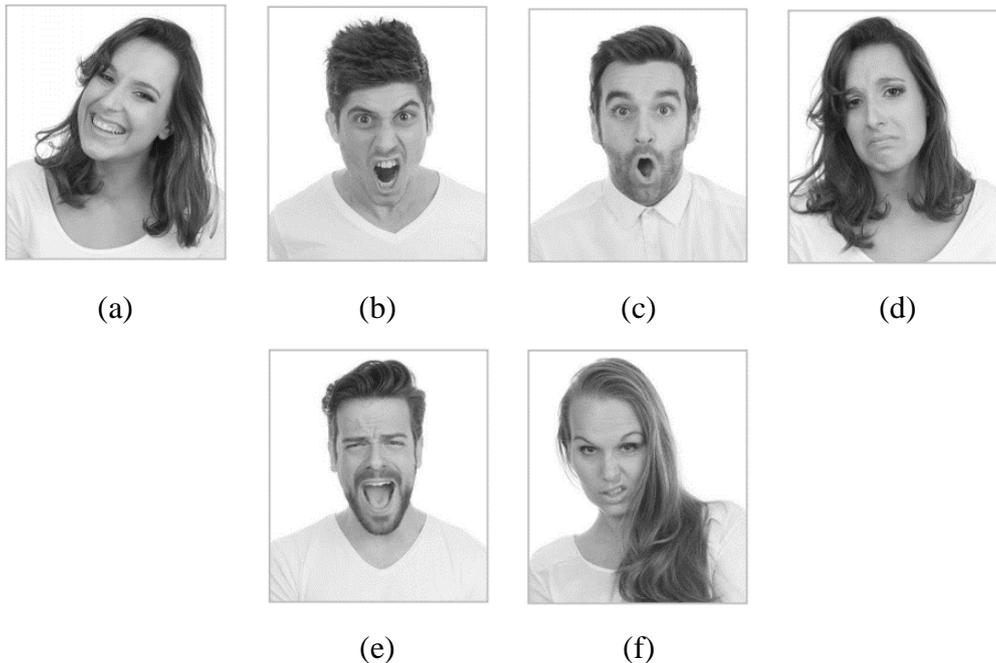


Figure I.3 : Exemples d'expressions émotionnelles prototypiques :  
(a) joie, (b) colère, (c) surprise, (d) tristesse, (e) peur, (f) dégoût.

L'étude de l'expression faciale repose sur l'analyse du mouvement de la peau (par exemple : les sourcils, les lèvres, les joues) et les rides (par exemple : sur le front, entre les sourcils ou sur le nez) et utilise souvent le système de codage FACS. Le système a été initialement développé en analysant des séquences vidéos d'individus et en associant les changements d'apparence faciale avec les contractions des muscles sous-jacents. Cette étude a permis la définition de plus de soixante unités d'action distinctes dont les plus fréquentes sont données par le tableau I.1. Par exemple (figure I.4):

- L'élévation des coins intérieurs des sourcils (contraction du muscle frontalis) correspond à AU 1.
- L'abaissement des coins internes des sourcils (activation des muscles corrugator, procerus, depressor supercillii) correspond à AU 4.

En utilisant FACS, il est possible de décrire presque n'importe quelle expression faciale anatomiquement possible, en la décomposant en AUs spécifiques. Les unités d'actions

impliquées dans quelques expressions émotionnelles prototypiques sont données par le tableau I.2 et illustrées par la figure I.5.

AU	Description	Exemple
1	Remontée de la partie interne des sourcils	
2	Remontée de la partie externe des sourcils	
4	Abaissement et rapprochement des sourcils	
5	Ouverture entre la paupière supérieure et les sourcils	
6	Remontée des joues	
7	Tension de la paupière	
9	Plissement de la peau du nez vers le haut	
12	Étirement du coin des lèvres	
15	Abaissement des coins externes des lèvres	
16	Ouverture de la lèvre inférieure	
20	Étirement externe des lèvres	
23	Tension refermante des lèvres	
26	Ouverture de la mâchoire	

Tableau I.1 : Les AUs fréquentes de FACS [1]

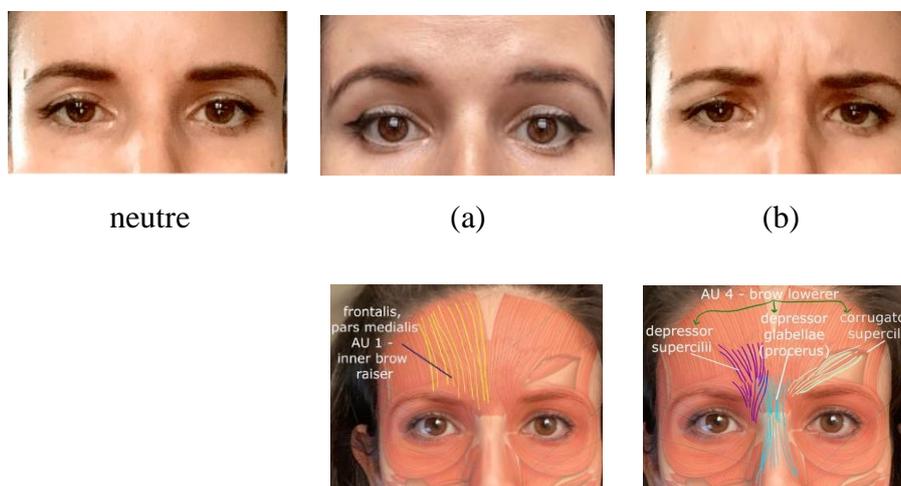


Figure I.4 : Muscles impliqués dans l'activation de (a) AU1 et (b) AU2

Émotion	Action Units
Joie	6+12
Tristesse	1+4+15
Surprise	1+2+5+26
Peur	1+2+4+5+20+26
Colère	4+5+7+23
Dégoût	9+15+16

Tableau I.2 : Combinaison des AUs correspondant à des expressions prototypiques



Figure I.5 : AUs actives pour quelques expressions faciales

En, général, les expressions faciales peuvent être analysées de trois manières différentes[1]:

- Par suivi de l'activité électromyographique du visage (facial electromyography fEMG).
- Par expert humain entraîné qui observe et classe manuellement l'activité faciale.
- Par analyse automatique des expressions faciales à l'aide des algorithmes de vision par ordinateur.

## I.5. Approches de reconnaissance automatique d'expressions faciales

### I.5.1 Approches conventionnelles

Diverses approches conventionnelles de reconnaissance automatique des expressions faciales ont été étudiées au cours des décennies. Ces approches implémentent un processus pour l'extraction de caractéristiques et pour la classification, où chaque étape est effectuée séparément. Un système conventionnel de reconnaissance automatique des expressions faciales est généralement composé de trois modules principaux, comme illustré par la Figure I.6. Le premier module consiste à extraire la région du visage dans les images ou les séquences d'images d'entrée. Le deuxième module consiste à extraire les caractéristiques représentant les changements faciaux causés par les expressions faciales. Le module de classification détermine une similarité entre l'ensemble des caractéristiques extraites et un ensemble de caractéristiques de référence. D'autres modules de prétraitement de données peuvent être utilisés entre ces modules principaux pour améliorer les résultats de détection, d'extraction de caractéristiques et de classification [5]

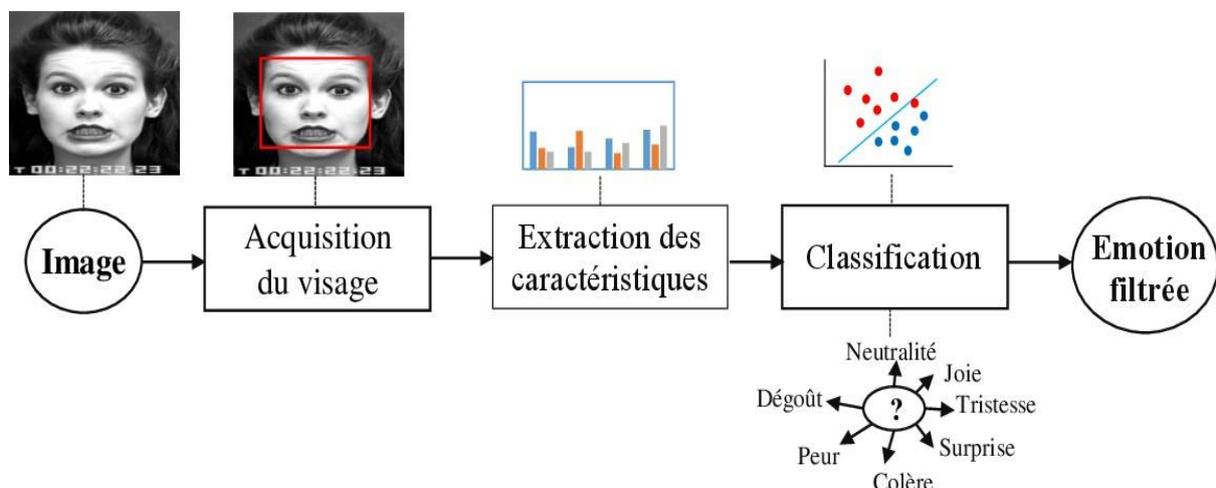


Figure I.6 : Composantes du système d'analyse des expressions faciales

### II.5.1.1 Acquisition d'images et extraction du visage

Le système d'acquisition est généralement équipé d'un appareil photo ou **caméra** qui permet d'acquérir une image 2D du visage. L'image acquise sera traitée pour détecter et localiser la région du visage et ses limites. Les algorithmes de détection de visages regroupent les différentes méthodes en quatre catégories : des méthodes basées sur la connaissance, des approches invariantes, des méthodes d'appariement de modèles, des méthodes basées sur l'apparence [6].

- A. Les méthodes basées sur la connaissance utilisent des règles prédéfinies basées sur la connaissance humaine afin de détecter un visage (par ex. un visage comprend deux yeux, un nez et une bouche avec des positions relatives clairement définies entre elles).
- B. Les approches basées sur les caractéristiques invariantes visent à trouver des caractéristiques de structure du visage robustes aux conditions de pose et d'éclairage.
- C. Les méthodes basées sur la correspondance avec les modèles utilisent des modèles de visage pré-stockés. Les valeurs de corrélation entre le modèle et l'image d'entrée sont calculées et la présence du visage est déterminée en utilisant ces valeurs de corrélation,
- D. Les méthodes basées sur l'apparence utilisent l'apprentissage automatique et les techniques statistiques pour modéliser le visage. Les modèles appris se présentent généralement sous la forme de distributions et de fonctions discriminantes, et les dits modèles sont utilisés pour distinguer les objets faciaux ou non-faciaux.
- E. Les méthodes basées sur des modèles rigides et comprennent les variations de boosting. Les principaux algorithmes de cette catégorie comprennent l'algorithme de détection de visage Viola-Jones (VJ) et ses variations.

### I.5.1.2 Extraction des caractéristiques

Le point commun de ces approches est de détecter la région du visage et d'extraire des caractéristiques. Les expressions faciales sont définies principalement par la contraction des muscles faciaux qui produisent des changements dans l'apparence et la forme du visage. De ce fait, les méthodes d'extraction des caractéristiques pour l'analyse d'expression peuvent être séparées en deux types d'approches : les méthodes basées sur les caractéristiques géométriques et les méthodes basées sur l'apparence [7]

Les caractéristiques géométriques représentent la forme et l'emplacement des composants du visage (y compris la bouche, les yeux, les sourcils et le nez) et les relations entre eux. Les caractéristiques d'apparence représentent les changements d'apparence (texture de la peau) du visage, tels que les rides et les sillons et peuvent être extraites sur tout le visage ou sur

des régions spécifiques du visage (figure I.7). Les caractéristiques sont extraites en utilisant des filtres tels que les filtres Gabor ou des descripteurs tels que les histogrammes de modèle binaire local (LBP), la transformation de caractéristique invariante à l'échelle (SIFT), l'histogramme de gradient orienté (HOG) ou les fonctionnalités robustes accélérées (SURF) [2]. Ces descripteurs sont généralement conçus pour coder l'existence de diverses caractéristiques de bas niveau telles que les coins, les jeux de couleurs, la texture de l'image du visage, etc. [8]



Figure I.7 : Les caractéristiques d'apparence

### I.5.1.3 Classification

La dernière étape d'un système de reconnaissance automatique d'expressions faciales est la détermination de l'expression faciale en fonction des caractéristiques extraites. Certains systèmes classent directement les expressions tandis que d'autres classent les expressions en reconnaissant d'abord des unités d'action (AUs) particulières. De nombreux classifieurs ont été appliqués à la reconnaissance d'expression tels que:

- Réseaux de neurone (Neural Networks, NN),
- Machines à vecteurs de support (Support Vector Machine, SVM),
- Analyse Discriminante Linéaire (Linear discriminant analysis, LDA),
- K-plus proche voisin (K Nearest Neighbor, KNN),
- Régression logistique multinomiale (Multinomial Regression Logistic, MRL),
- Modèles de Markov cachés (Hidden Markov Model, HMM),
- Réseaux bayésiens (Bayesian Network, BN), et d'autres.

### I.5.1 Approches basées sur l'apprentissage profond

Les approches de reconnaissance d'expressions faciales conventionnelles caractérisent d'abord les traits du visage, puis les utilisent pour l'inférence. Comme ces deux étapes sont effectuées séparément, une performance sous-optimale est obtenue, en particulier sur les ensembles de données les plus difficiles où de nombreuses sources de variabilité sont présentes [15]. Il est plus avantageux de réaliser conjointement ces deux étapes pour de meilleures performances de reconnaissance optimales.

L'apprentissage profond ou Deep Learning permet d'apprendre des représentations de

caractéristiques discriminantes pour la reconnaissance automatique des expressions faciales en concevant une architecture hiérarchique composée de multiples transformations non linéaires basées sur des réseaux de neurones de différents types tels que les réseaux convolutifs et réseaux récurrents (Recurrent Neural Network RNN). Pour la tâche de classification, ces réseaux sont couplés à une couche de classification. Ainsi les paramètres d'apprentissage d'un classifieur se fait conjointement avec l'apprentissage des représentations des caractéristiques. En automatisant le processus d'extraction et de classification des caractéristiques directement à partir des données brutes, la dépendance vis-à-vis des modèles basés sur la géométrie du visage et d'autres techniques de prétraitement est fortement réduite.

## **I.6 Conclusion**

Dans ce chapitre, nous avons d'abord présenté les méthodes de description et d'analyse des expressions faciales et l'importance de leurs reconnaissance automatique. Nous avons ensuite exposé les différentes approches des systèmes de reconnaissance conventionnels et récents. L'étude des méthodes conventionnelles a permis de mettre en évidence les différentes difficultés inhérentes à la reconnaissance automatique des expressions faciales en utilisant des caractéristiques liées à la géométrie ou l'apparence du visage. En revanche, les méthodes basées sur le deep learning permettent de contourner les limitations des méthodes conventionnelles et seront détaillées dans le prochain chapitre.

# **Chapitre II**

**Les réseaux de neurones convolutifs**

## II.1. Introduction

Le deep learning est un type d'intelligence artificielle dérivé du machine learning où la machine est capable d'apprendre par elle-même. L'architecture de deep learning la plus populaire est les réseaux de neurones convolutifs (CNN) qui sont une catégorie de réseaux de neurones qui se sont avérés très efficaces dans des domaines tels que la reconnaissance et la classification d'images. Les CNN ont réussi à identifier les visages, les objets, panneaux de circulation etc. c'est la raison pour laquelle nous avons décidé de travailler avec cette structure pour notre projet.

## II.2. Principe du deep learning

Le deep learning est un ensemble d'algorithmes d'apprentissage automatique qui tentent d'apprendre à plusieurs niveaux, correspondant à différents niveaux d'abstraction. Il a la capacité d'extraire des caractéristiques à partir des données brutes grâce aux multiples couches de traitement composé de multiples transformations linéaires et non linéaires et apprendre sur ces caractéristiques petit à petit à travers chaque couche.

Le deep Learning s'appuie sur un réseau de neurones artificiels s'inspirant du cerveau humain (figure II.1). Le neurone artificiel reprend l'architecture et le fonctionnement du neurone biologique où les synapses assurent les connexions avec les autres neurones, les dendrites sont les entrées, les axones sont les sorties et le noyau active les sorties selon les stimulations en entrées selon l'expression suivante :

$$\hat{y} = f(w.x + b)$$

Où

- $x$  représente le vecteur d'entrées ;
- $\hat{y}$  représente la sortie (prédiction) ;
- $w$  et  $b$ , sont, respectivement, le poids et le biais (les paramètres) influençant le fonctionnement du neurone à travers la fonction d'activation associé  $f$ .

Cependant, un seul neurone ne permet pas de répondre à des problèmes complexes. Un réseau neuronal est l'association, en une structure plus ou moins complexe, de plusieurs dizaines voire centaines ou de milliers de neurones. Les principaux réseaux se distinguent principalement par l'organisation et le nombre de couches, le nombre de neurones et type d'apprentissage.

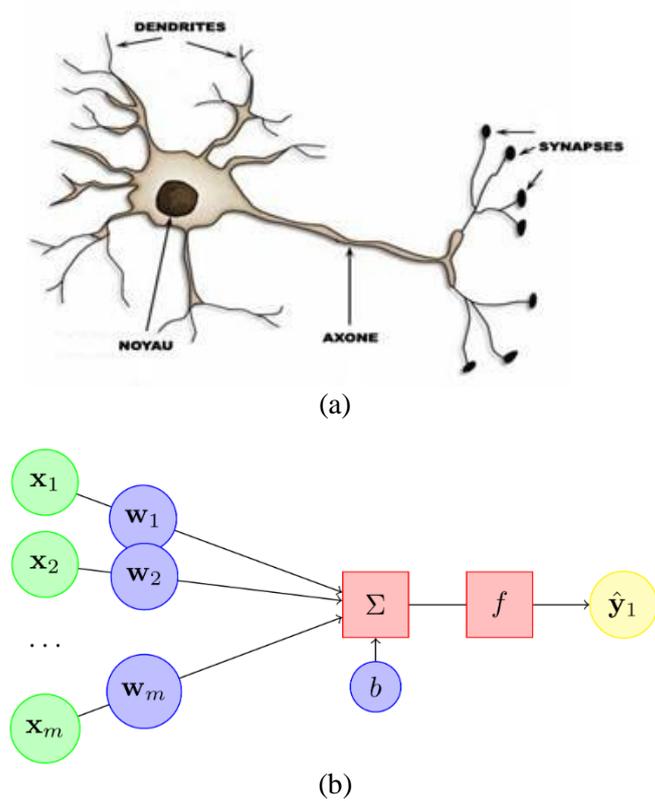


Figure II.1 : Neurone biologique (a) et neurone artificiel (b)

Le réseau de neurones est, donc, composé de plusieurs couches de neurones, chacune recevant et interprétant les informations de la couche précédente (figure II.2). Ainsi, le réseau multicouche, appelé aussi perceptron multicouche (Multi Layer Perceptron MLP) apprendra par exemple à reconnaître les lettres avant de s'attaquer aux mots dans un texte, ou détermine s'il y a un visage sur une photo avant de découvrir de quelle personne il s'agit.

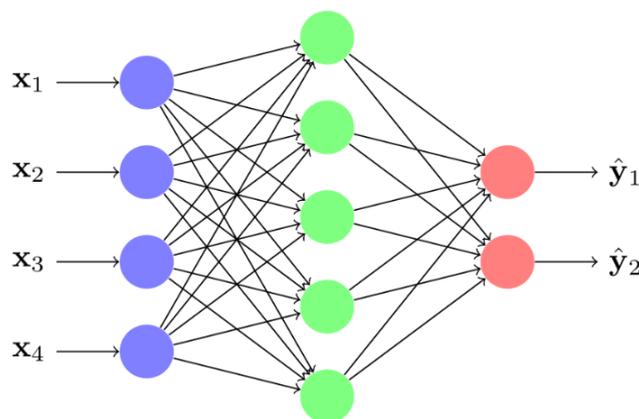


Figure I.2: Réseau de neurones artificiels

Bien qu'il existe un grand nombre de variantes d'architectures profondes. Il n'est pas toujours possible de comparer les performances de toutes les architectures, car elles ne sont pas toutes évaluées sur les mêmes ensembles de données. Les structures de deep learning les plus utilisées sont les réseaux récurrents et les réseaux convolutifs [22]. Dans ce qui suit nous

présenterons les réseaux convolutifs utilisés dans notre projet.

### II.3 Réseaux de neurones convolutifs (CNN)

Le but d'un CNN est d'apprendre des fonctionnalités d'ordre supérieur dans les données via des convolutions. Ils sont bien adaptés à la reconnaissance d'objets et de classification d'images. Ces structures permettent d'identifier des visages, des individus, des panneaux de signalisation et de nombreux autres aspects des données visuelles.

L'architecture typique d'un CNN est donnée par la figure II.3. Cette architecture comprend en plus de la couche d'entrée, plusieurs couches de convolution et de pooling. Le sommet du réseau est un réseau complètement connecté (Full Connected FC) qui permet la classification des caractéristiques extraites par les couches de convolutions.

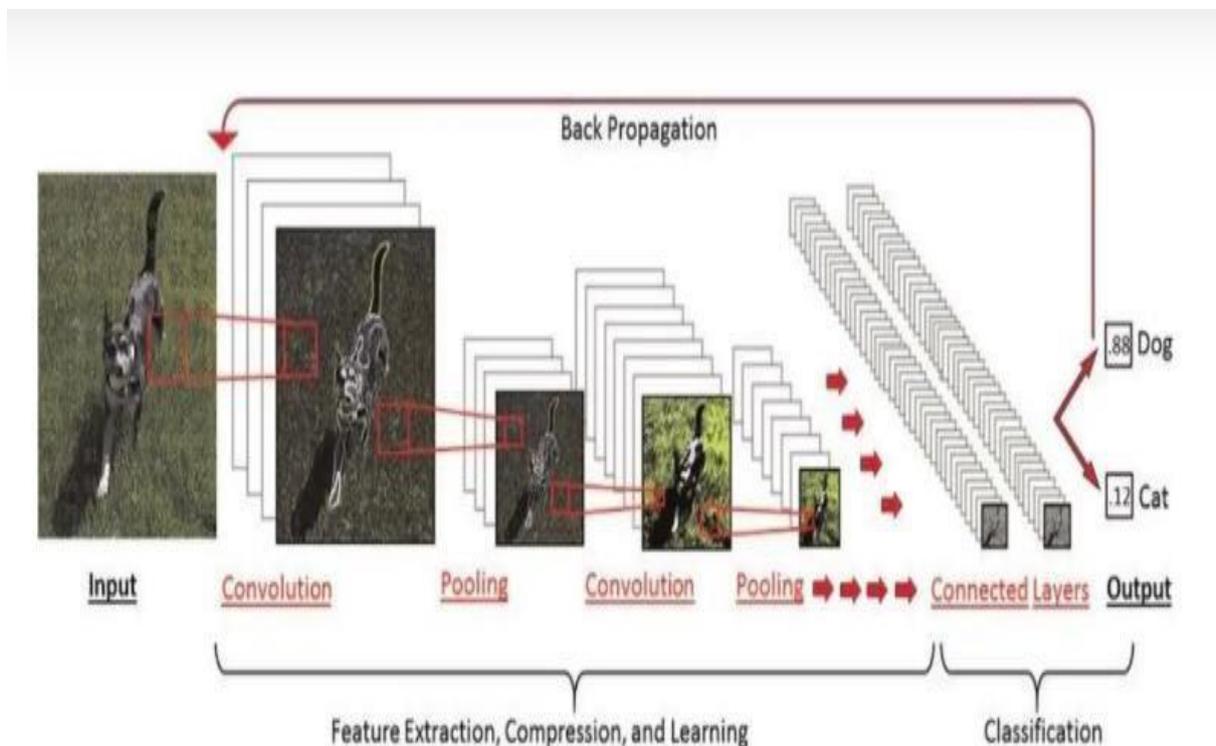


Figure II.3 : Architecture typique d'un réseau de neurone convolutif

La couche d'entrée est constituée de la matrice de données à classifier, représentée généralement par la matrice de l'image, dans le cas des images monochromes, ou les 3 matrices de l'image pour le cas d'images RGB.

#### II.3.1 La couche de convolution

La couche de convolution sert à l'extraction de caractéristiques à partir d'images. Elle génère de nouvelles matrices (images) appelées cartes de caractéristiques (features map). Cette carte accentue les caractéristiques uniques de l'image d'origine à travers les filtres de

convolution dont les paramètres sont fixés par le processus d'apprentissage. La figure II.4 montre un résultat de convolution, où la marque encadrée \* indique l'opération de convolution et la marque  $\phi$  est la fonction d'activation. Les icônes carrées en niveaux de gris entre ces opérateurs indiquent les filtres de convolution. La couche de convolution génère le même nombre de cartes d'entités que les filtres de convolution.

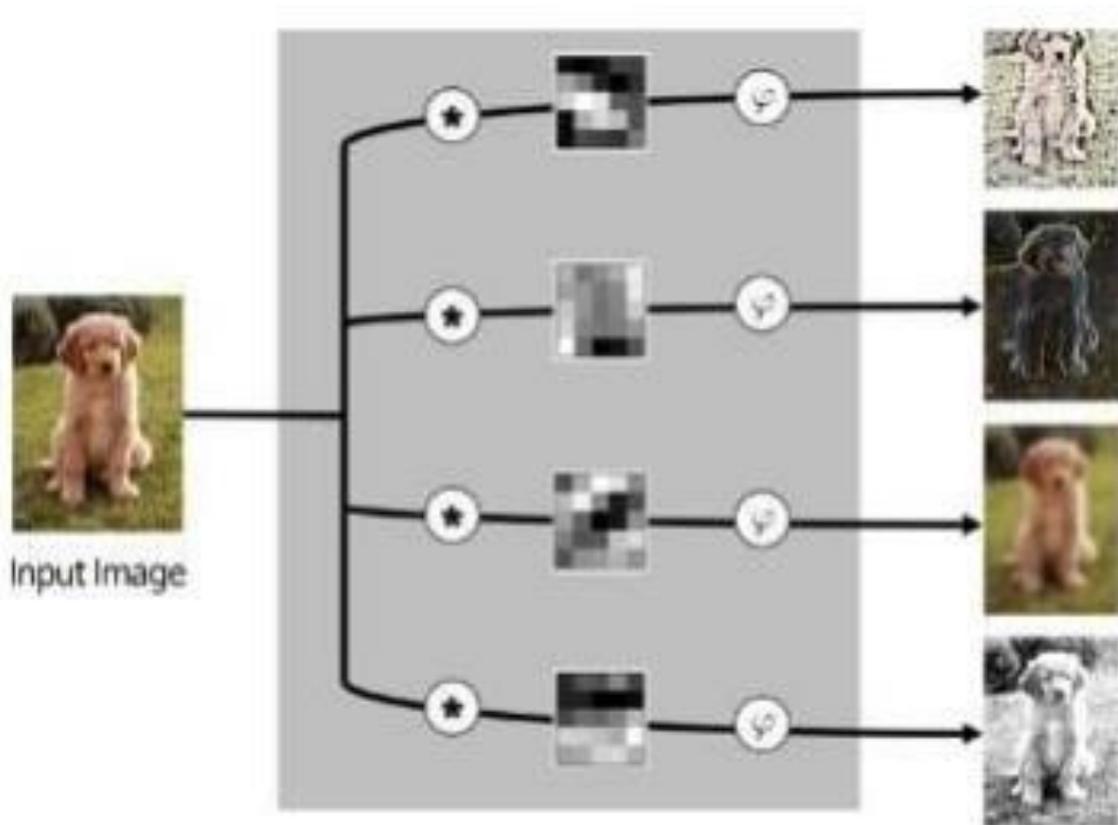


Figure II.4 : Exemple de résultat de la convolution d'une image d'entrée par 4 filtres.

Une couche de convolution est caractérisée par la taille et le nombre des filtres de convolution. La figure II.5 illustre l'opération de convolution. En considérant une image ou une carte de caractéristiques de  $5 \times 5$  pixels (matrice bleue) et un filtre de  $3 \times 3$  (matrice rouge), la convolution est la somme des produits des éléments qui se trouvent aux mêmes positions des deux matrices pour chaque étape où le filtre de convolution est décalé rapport à la matrice de données selon un pas (stride) horizontalement et verticalement. Pour conserver la même dimensionnalité des matrices d'entrée et de sortie, le zero padding peut être utilisé (Figure II.6).

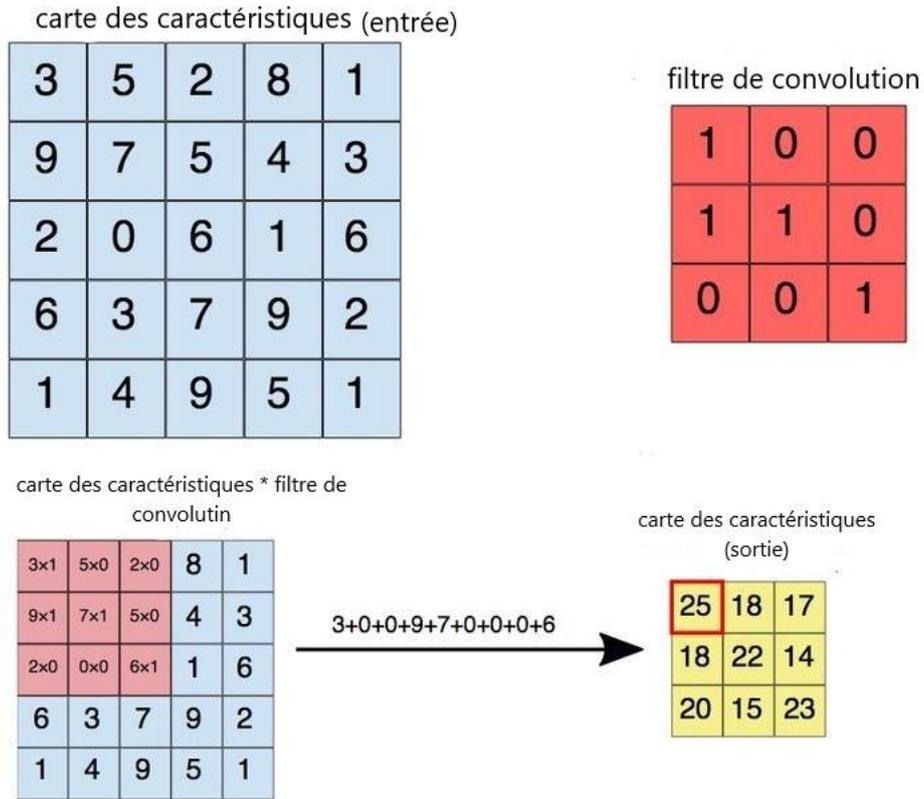


Figure II.5 : Exemple illustrant l'opération de convolution 3x3

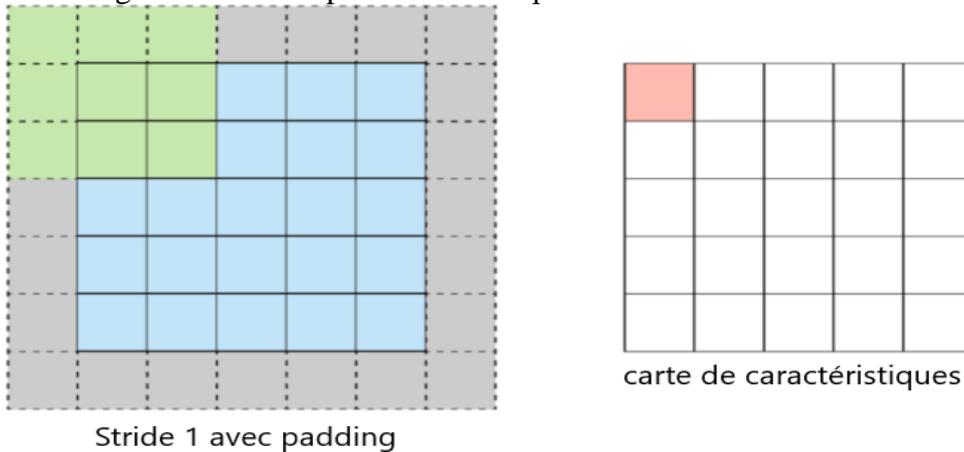


Figure II.6 : Exemple de zéro padding

### II.3.2. Non-linéarité

Pour améliorer les performances d'un réseau de neurones, des fonctions d'activation non linéaires sont utilisées. Ainsi, les valeurs dans les cartes de caractéristiques finales ne sont pas réellement des sommes. La fonction unité linéaire rectifiée (ou ReLU pour Rectified Linear Unit) est la fonction d'activation la plus utilisée dans les réseaux convolutifs. Cette fonction est définie par :

$$f(x) = \begin{cases} x & \text{si } x > 0 \\ 0 & \text{ailleurs} \end{cases}$$

### II.3.3 La couche de pooling

Après une opération de convolution, un pooling est généralement effectué pour réduire la dimensionnalité des cartes de caractéristiques. Cela permet de réduire le nombre de paramètres, ce qui à la fois raccourcit le temps d'entraînement et empêche le surapprentissage (overfitting)[1]. Les couches Pooling sous-échantillonnent chaque carte de caractéristiques indépendamment, en réduisant la hauteur et la largeur, tout en conservant la profondeur intacte. Il existe plusieurs types de pooling (figure II.7) :

- *max pooling* : qui revient à prendre la valeur maximale de la sélection. C'est le type le plus utilisé, car il est rapide à calculer (immédiat), et permet de simplifier efficacement l'image.
- *mean pooling* : (ou average pooling) qui revient à calculer la somme de toutes les valeurs et diviser par le nombre de valeurs. On obtient ainsi une valeur intermédiaire pour représenter ce lot de pixels.
- *sum pooling* : c'est la moyenne sans avoir divisé par le nombre de valeurs (on ne calcule que leur somme)

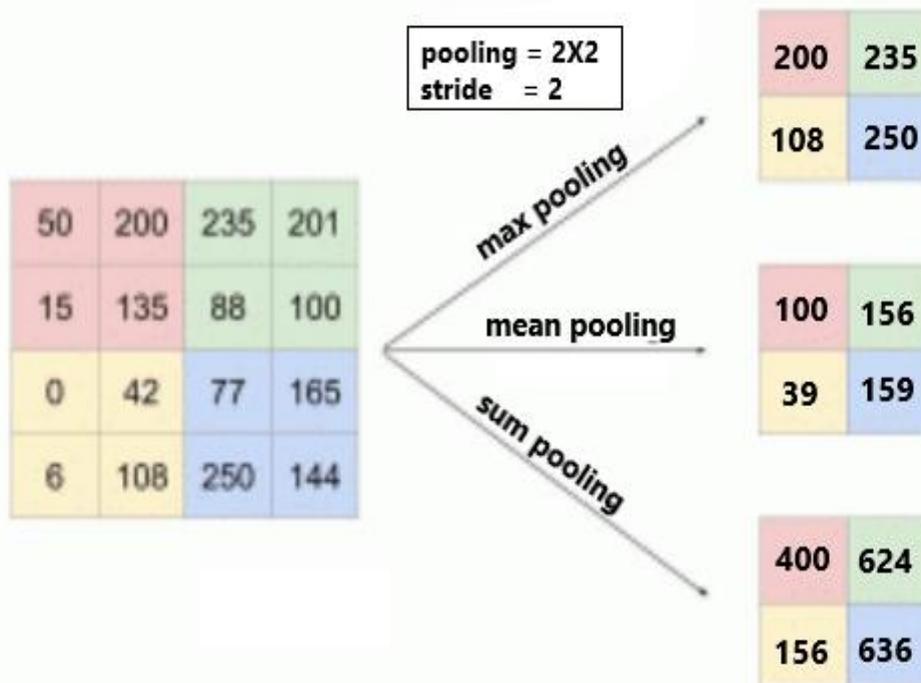


Figure II.7 : Exemples de calcul du pooling sur une image 4×4

Contrairement à l'opération de convolution, l'opération de pooling n'utilise pas des poids et fait glisser une fenêtre sur son entrée et prend simplement une valeur dans cette fenêtre selon le type du pooling.

### II.3.4 Couches entièrement connectées

La couche entièrement connectée est un perceptron multicouche traditionnel qui utilise généralement, la fonction d'activation softmax dans la couche de sortie. La fonction softmax maintient la somme des valeurs de sortie à un et limite également les sorties individuelles à des valeurs de 0 à selon la formule mathématique suivante :

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{i=1}^K e^{z_i}}$$

$\mathbf{z}$  : vecteur d'entrée de  $\mathbf{k}$  nombres réels.  $z = \{ z_1, z_2, \dots, z_k \}$

$\sigma(\mathbf{z})$  : vecteur de sortie de  $\mathbf{k}$  nombres réels strictement positifs et de somme 1.

Le terme « entièrement connecté » implique que chaque neurone dans la couche précédente est connecté à chaque neurone sur la couche suivante. La sortie des couches de convolution et de pooling représente les fonctions de haut niveau de l'image d'entrée. Le but de la couche entièrement connectée est de pouvoir utiliser ces fonctions pour classer l'image d'entrée dans différentes classes en fonction de l'ensemble de données d'apprentissage.

### II.4 L'apprentissage d'un CNN

Les réseaux de neurones artificiels, apprennent par principe de rétro propagation de données, les données sont propagées depuis l'entrée, puis activées grâce à une fonction d'activation, pour enfin avoir en sortie un résultat primaire. Une fois le résultat obtenu, il est comparé avec le résultat désiré, une erreur alors peut être calculée grâce à une fonction de coût  $J$  (cost function).

La rétropropagation du gradient de l'erreur (ou back propagation) est un algorithme d'optimisation permettant d'ajuster les paramètres d'un réseau de neurones multicouches pour mettre en correspondance les entrées (l'image) et les sorties (le classement désiré). La caractéristique principale de la rétropropagation est sa méthode itérative et récursive pour le calcul des mises à jour des paramètres du réseau (poids et biais) avec les formules suivantes :

$$w_i = w_i - \tau \frac{\partial J}{\partial w_i}$$

$$b_i = b_i - \tau \frac{\partial J}{\partial b_i}$$

Où  $\tau$  est le taux d'apprentissage

Il existe plusieurs méthodes qui accélèrent l'optimisation d'apprentissage des modèles en deep learning en appliquant des taux d'apprentissage, tel que l'algorithme de gradient adaptatif Adam (adaptive moment estimation) et RMSprop (Root Mean Square Propagation).

Un des principaux problèmes reliés à l'apprentissage profond est que celui-ci demande une quantité phénoménale de données pour l'entraînement. Ce problème n'est pas négligeable dans la majorité des cas d'application, car les données annotées appropriées pour l'application en question n'existent souvent pas en quantité suffisante. Il faut donc manuellement les collecter et les étiqueter, en s'assurant de leur qualité. À noter qu'il est possible de mettre en œuvre des techniques d'augmentation du jeu d'entraînement s'il est trop réduit. Il s'agit de créer des variations d'une même image par de légères transformations pour augmenter l'ensemble des données d'entraînement. Ainsi, une série d'opérations de rotation, zoom, décalage de largeur et de hauteur, cisaillement, retournement horizontal et remplissage sont appliquées à l'ensemble de données d'entraînement.

L'entraînement des réseaux CNN dépend de beaucoup d'hyperparamètres mais le plus important c'est le taux d'apprentissage ; pour fixer sa valeur, plusieurs valeurs sont généralement testées et la valeur qui donne les meilleures performances sera sélectionnée.

## II.5 L'apprentissage par transfert

Pour utiliser un CNN, 2 solutions se présentent :

- 1- Construire un réseau personnalisé en empilant les différentes couches puis l'entraîner en utilisant une base de données annotées. Créer un nouveau réseau de neurones convolutif est coûteux en termes d'expertise, de matériel et de quantité de données annotées nécessaires. L'entraînement peut prendre plusieurs semaines pour les meilleurs CNN destinés à la classification d'image, avec de nombreux GPU (Graphics Processing Units) travaillant sur des centaines de milliers d'images annotées.
- 2- Pour des usages pratiques, il est possible d'exploiter la puissance des CNN existants pré-entraînés en les adaptant à une nouvelle classification d'image avec du matériel accessible et une quantité raisonnable de données annotées. Toute la complexité de création de CNN peut être évitée en adaptant des réseaux pré-entraînés disponibles publiquement.

Le principe de la 2<sup>ème</sup> solution consiste à exploiter l'apprentissage acquis sur un problème de classification général pour l'appliquer de nouveau à un problème particulier. La technique de prendre un réseau existant pré-entraîné et l'utiliser comme point de départ pour former un modèle pour une tâche similaire, est appelée apprentissage par transfert (Transfer Learning).

### II.5.1 Stratégies de transfer learning

La "connaissance" sur la classification d'images contenue dans un réseau pré-entraîné

peut-être exploitée de deux façons, en utilisant le réseau (figure II.8) :

- soit comme un extracteur automatique de caractéristiques des images, matérialisé par le vecteur de caractéristiques issu de la partie convolutive.
- Soit comme un modèle initial, qui est ensuite réentraîné plus finement (Fine Tuning) pour traiter le nouveau problème de classification.

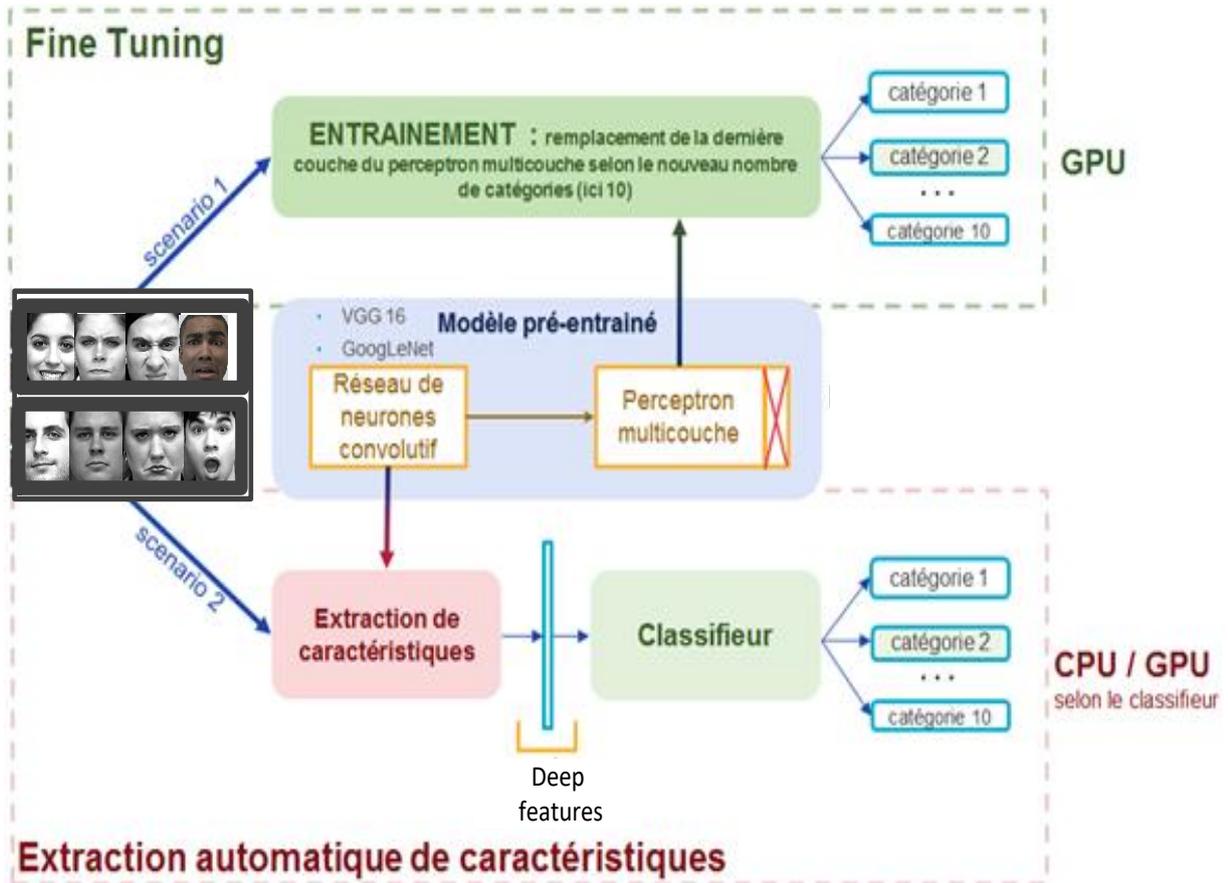


Figure II.8: Techniques de Transfer Learning

Ainsi, la mise au point d'un réseau avec l'apprentissage par transfert est généralement beaucoup plus rapide et plus facile que la formation d'un réseau à partir de zéro. L'approche est couramment utilisée pour la détection d'objets, la reconnaissance d'images, la reconnaissance vocale et d'autres applications.

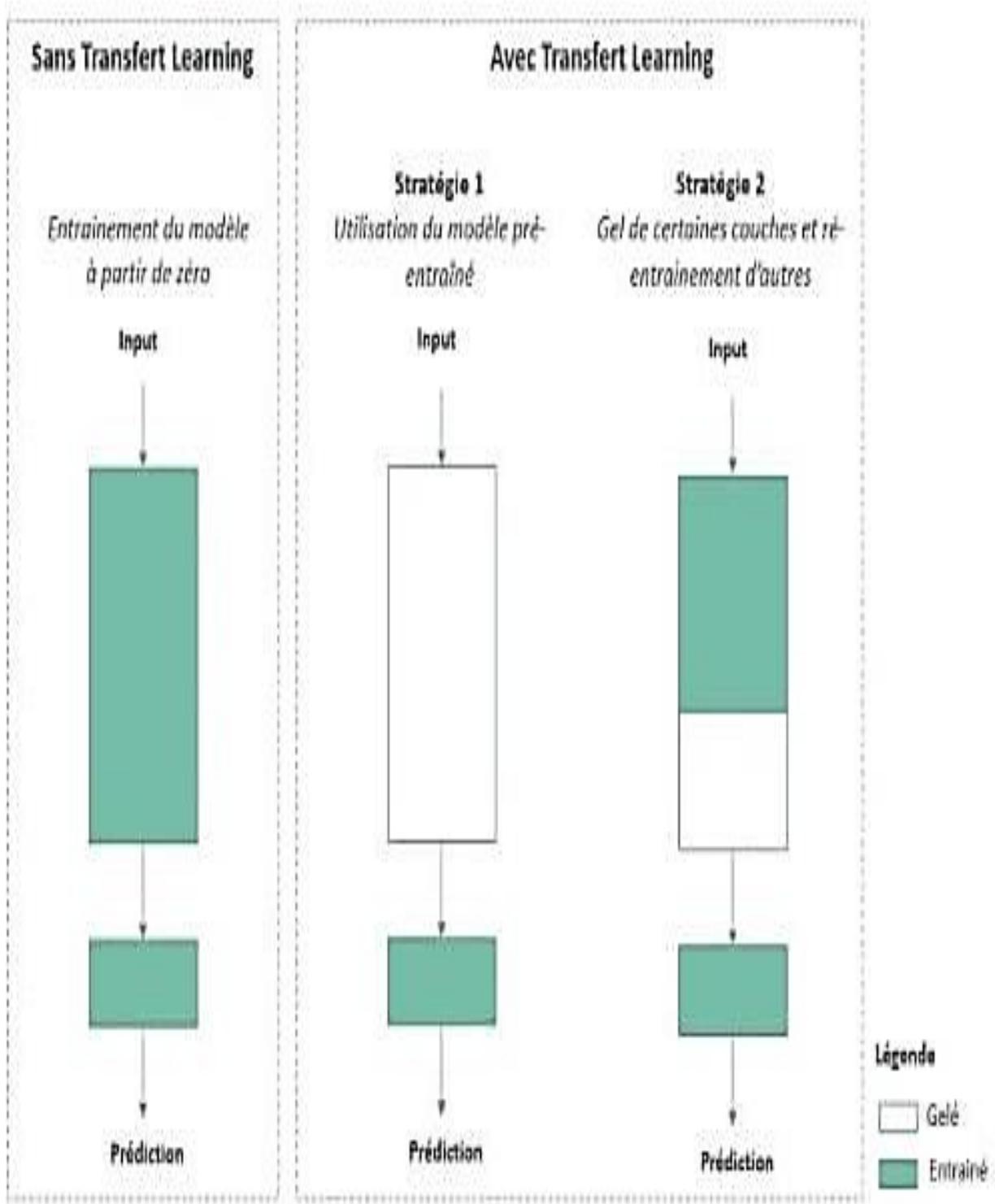


Figure II.9 : Approches de transfert learning en Deep Learning

### II.5.2 Quelques réseaux convolutifs pré-entraînés [16]

Il existe plusieurs réseaux de classification d'images dont les plus répandus ont été entraînés avec plus d'un million d'images et peuvent classer les images en un ou plusieurs milliers de catégories d'objets ( tels que « clavier », « tasse à café », « chat », « voiture », « joueur de baseball », etc). ImageNet est une base de données d'images publique évolutive qui

comporte un total de 14 millions d'images et 22 mille catégories visuelles [24]. Cette base de données a été largement utilisée dans la recherche d'algorithmes de reconnaissance et a été principalement utilisée pour entraîner des réseaux de deep learning sur le sous-ensemble sélectionné pour l'ImageNet Challenge (ILSVRC ImageNet Large Scale Visual Recognition Competition) fournit par exemple 1.2 millions d'images classées en 1000 catégories [25]. Ainsi, de nombreux CNN de référence existent en plusieurs variantes. La figure II.10 présente l'évolution et montrant leur nombre de paramètres (poids et biais) comme taille du cercle. Les différentes couleurs distinguent les différentes familles d'architectures (les couleurs mélangées signifient que les concepts de deux familles d'architecture sont combinés). Ainsi, de nombreux CNN de référence existent en plusieurs variantes tels que :

- ✓ **LeNet** : Les premières applications réussies des réseaux convolutifs ont été développées par Yann LeCun dans les années 1990. Parmi ceux-ci, le plus connu est l'architecture LeNet utilisée pour lire les codes postaux, les chiffres, etc.
- ✓ **AlexNet** : Le premier travail qui a popularisé les réseaux convolutifs dans la vision par ordinateur était AlexNet, développé par Alex Krizhevsky, Ilya Sutskever et Geoff Hinton. Ce CNN a emporté ImageNet challenge en 2012 en surpassant nettement ses concurrents. Le réseau avait une architecture très similaire à LeNet, mais était plus profond, plus grand et comportait des couches convolutives empilées les unes sur les autres (auparavant, il était commun de ne disposer que d'une seule couche convolutive toujours immédiatement suivie d'une couche de pooling) .
- ✓ **GoogLeNet** : C'est un modèle développé par Google. Sa principale contribution a été le développement d'un module d'inception qui a considérablement réduit le nombre de paramètres dans le réseau (4Million, par rapport à AlexNet avec 60Millions). En outre, ce module utilise le global Average pooling ce qui élimine une grande quantité de paramètres. Il existe également plusieurs versions de GoogLeNet, parmi elles, Inception-v4 et Xception.
- ✓ **ResNet** : développé par Kaiming He et al. Ce réseau été le vainqueur de ILSVRC 2015 (ImageNet Large Scale Visual Recognition Challenge). Il présente des sauts de connexion et une forte utilisation de la batch normalisation. Il utilise aussi le global AVG pooling au lieu du PMC à la fin. Ce dernier est le modèle dont notre système s'inspire, plus de détails dans le chapitre prochain.

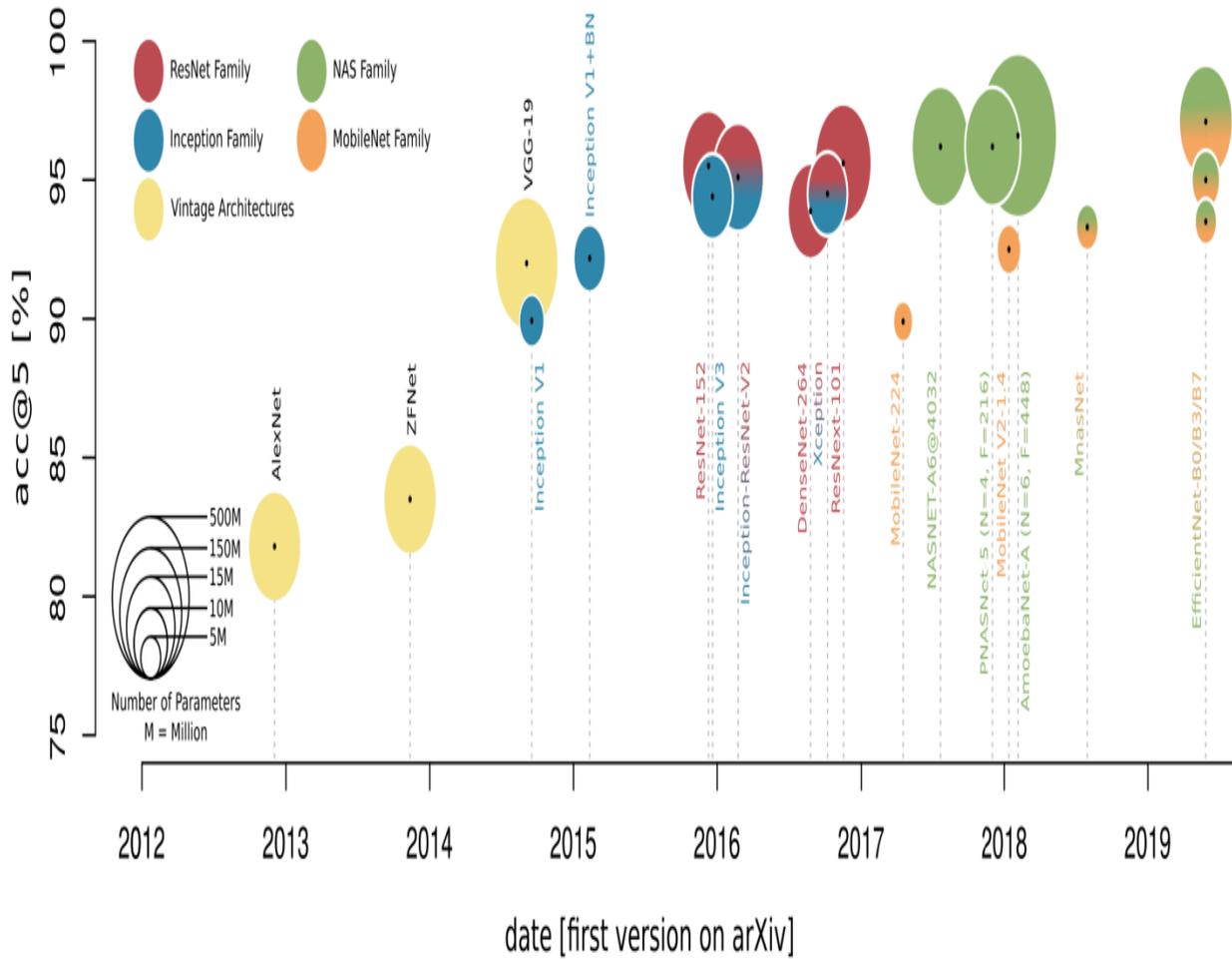


Figure II.10 : Évolution des CNN pour la reconnaissance d'images ImageNet 2012 [26]

Ces CNN pré-entraînés sont disponibles pour toutes les plateformes de développement d'application à deep learning (TensorFlow, Theano...) et aussi Matlab. Le Deep Learning Toolbox de Matlab, que nous utiliserons dans ce projet, offre une multitude de fonctions qui permettent de réutiliser ces réseaux [26].

## Conclusion

Dans ce chapitre, nous avons présenté les concepts principaux du réseau de neurones à convolution. Nous avons présenté également le concept d'apprentissage par transfert qui est une méthode efficace pour réutiliser un CNN pré-entraîné et l'adapter à un autre problème de classification au lieu de développer un modèle à partir de zéro. L'utilisation d'un CNN avec l'apprentissage par transfert pour la reconnaissance automatique des expressions faciales sera expliquée dans le prochain chapitre.

# **Chapitre III**

## **Implémentations et Résultats**

### III.1. Introduction

L'objectif de ce chapitre est de présenter les étapes de l'implémentation de l'approche utilisée afin de reconnaître automatiquement des expressions faciales. Nous commençons tout d'abord par la présentation du réseau choisi et de la base de données utilisée. Puis, nous présentons les étapes de l'ajustement du modèle pré-entraîné et les performances de classification obtenues.

### III.2. Base de données

Il existe plusieurs bases de données pour la reconnaissance des expressions faciales telles que [27]:

- JAFFE (JAPanese Female Facial Expressions) ,
- Extended Cohn Kanade (CK+),
- FER 2013 (Facial Expression Reconition 2013 database),
- AffectNet ,
- MMI (MMI, 2017),
- AFEW
- Karolinska Directed Emotional Faces (KDEF).

Le CK+ est l'ensemble de données le plus utilisé pour la détection des expressions faciales. Les performances des modèles CNN entraînés avec cet ensemble de données sont supérieures à 95% en raison du fait que les images sont capturées dans un environnement contrôlé (laboratoire) et les expressions sont exagérées (overacted). L'ensemble de données contient 593 images (taille 249x303) de 123 sujets. Les images traduisent les états neutres et les six émotions de base suivantes : colère, dégoût, peur, bonheur, tristesse et surprise. JAFFE est un ensemble de données contrôlées en laboratoire avec moins d'images que CK+ (taille d'image 112x112). Les modèles CNN testés à l'aide de cette base de données ont des performances supérieures à 90%. KDEF est également une base de données contrôlée en laboratoire qui se concentre sur la reconnaissance des émotions de cinq différents angles de caméra (taille d'image 314x415). Le FER2013 est le deuxième ensemble de données largement utilisé après CK+. La différence est que les images FER 2013 sont collectées du Web et ne sont pas contrôlées en laboratoire, les expressions ne sont pas exagérées, donc plus difficile à reconnaître (taille d'image 112x112). Dans cette base de données, il y a 28079 images d'entraînement appartenant à sept classes de base (colère, dégoût, peur, neutre, bonheur,

tristesse et surprise). Certaines compilations ont été créées afin d'améliorer la précision de FER 2013 basé sur les réseaux CNN en ajoutant des images contrôlées en laboratoire de CK +, JAFFE et KDEF.

Dans notre projet, nous utilisons une compilation d'images disponible sur Kaggle [28], composée d'images de FER 2013, CK+, JAFFE et KDEF. Les images FER2013 constituent la plus grande partie de l'ensemble de données (environ 90% pour chaque émotion). Les expressions faciales appartiennent aux sept classes suivantes : angry (colère), disgust (dégoût), fear (peur), happy (heureux), neutral (neutre), sad (triste) et surprise (surprise). La distribution des classes sur l'ensemble des données est : colère 10%, dégoût 2%, peur 3%, heureux 26%, neutre 35%, triste 11% et surprise 13%. Le tableau III.1 donne le nombre d'images utilisées pour l'entraînement et la validation du CNN utilisé.

Catégorie	Entrainement	Validation
angry	2878	720
disgust	564	141
fear	869	218
happy	7085	1772
Neutral	9773	2444
Sad	3030	758
Surprise	3616	904

Tableau III.1: Détails de la base de données utilisée.

### III.3. Le réseau pré-entraîné choisi

Par défaut de ressources matérielles qui permettent de mener des expériences afin d'étudier les performances de plusieurs réseaux pré-entraînés et choisir le mieux adapté à notre application, nous avons décidé de choisir un CNN dont les performances ont été vérifiées avec des études de recherche. Selon les résultats de comparaison, entre trois CNN (VGG16, ResNet50, InceptionV3) pour la reconnaissance des expressions faciales, effectuée par Melinte et al. [29], le réseau ResNet50 a affiché la meilleure précision de classification.

ResNet est un réseau convolutif profond qui utilise en partie des couches convolutives identité (taille des filtres de convolution est 1x1) et des connexions de contournement afin de surmonter le problème d'évanouissement (disparition) des gradients (Vanishing gradients). En effet, le gradient peut devenir extrêmement petit lorsqu'il est rétro-propagé à travers un réseau profond constitué de grand nombre de couches. Les connexions de contournement (raccourcis), qui sont un chemin alternatif de rétropropagation du gradient, résolvant ainsi le problème de

l'annulation du gradient. La couche d'entrée du ResNet est de taille 224x224 composée de 3 canaux (c-à-d ce réseau utilise des images RVB de 224x224 pixels).

Pour notre projet, nous utilisons le réseau ResNet50 qui est constitué de 50 couches réparties principalement sur 5 sections. Les 4 dernières sections comportent un bloc convolutif 3x3 et 2 blocs identité. La description détaillée de la structure et des caractéristiques des éléments de chaque bloc est donnée par la figure III.1. Dans cette figure, chaque couche de convolution est représentée par un rectangle qui précise la taille et le nombre des filtres de convolution, par exemple la première section est composée d'une seule couche de convolution qui utilise 64 filtres de taille 7x7 et de pas égale à 2 et padding égale à 3.

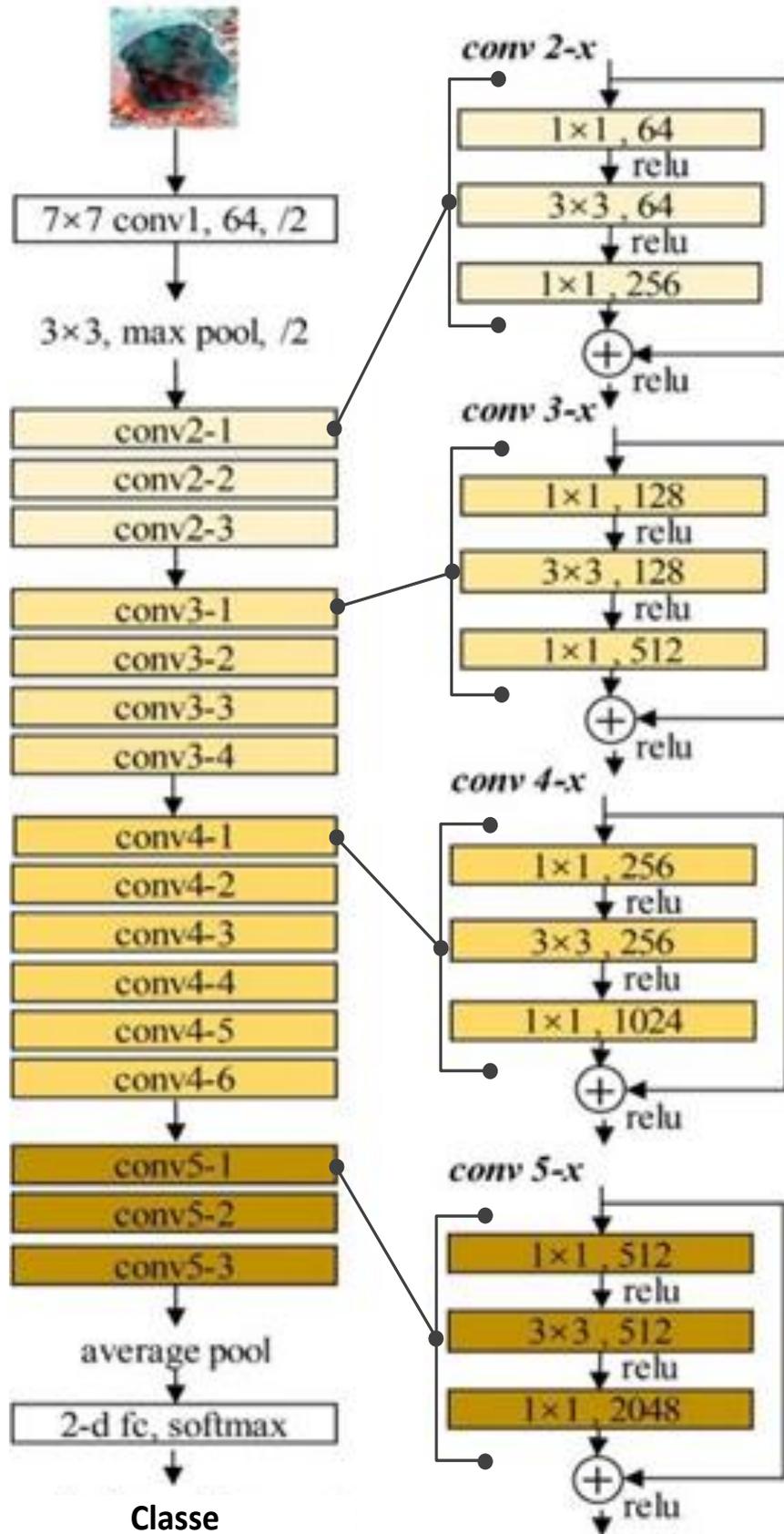


Figure III.1: Architecture de ResNet50.

Il faut noter également que ResNet normalise la réponse de chaque résultat de convolution. La technique de Batch Normalization peut améliorer fortement la convergence lors de l'entraînement. La couche de Batch Normalization normalise, en moyenne et en variance, la sortie de la couche à laquelle elle est associée.

Bien que la base de données ImageNet ne contient pas des catégories explicites de visages humains (comporte seulement 3 catégories de personnes (plongeur, marié et joueur de baseball), beaucoup d'études ont montré, que les CNN ont appris à reconnaître les visages humains [15]. À titre d'exemple, le visage de certains animaux et des humains peut être similaires et l'apprentissage effectué sur les catégories d'animaux pour servir à reconnaître les faces et classer leurs expressions, car le concept est que ces réseaux ont également appris des attributs (petites formes ou détails) dans les ensembles de données à grande échelle utilisées (figure III.2).

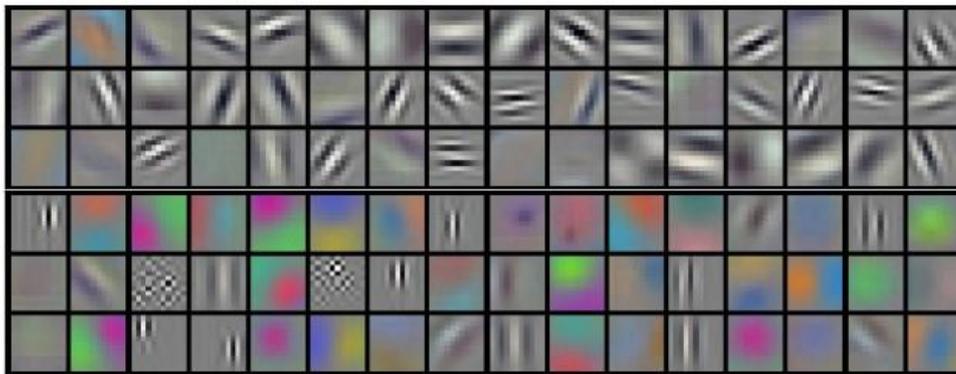


Figure III.2 : Exemple de coefficient de 96 filtres de convolution entraînés de taille égale à  $11 \times 11 \times 3$  illustrant le concept d'apprentissage d'attributs.

#### III.4. Environnement du travail

Pour notre projet, nous utilisons Deep Learning Toolbox de Matlab avec un PC Intel i5 Quadcore @3.90Ghz, 8Ghz de RAM et la carte graphique de NVIDIA GeForce GTX 1050 Ti qui possède 768 coeurs CUDA fonctionnant à une fréquence de 1.5 GHz. Rappelons que CUDA (Compute Unified Device Architecture) est une technologie de GPGPU (General-Purpose Computing on Graphics Processing Units), utilisant un processeur graphique (Graphic processor Unit GPU) pour exécuter des calculs à la place du processeur (CPU). En effet, ces processeurs comportent couramment de l'ordre d'un millier de circuits de calcul fonctionnant typiquement à 1 GHz, ce qui représente un potentiel très supérieur à un CPU si, et seulement si, le calcul peut s'effectuer en parallèle.

Afin de déterminer la meilleure stratégie de transfer learning en termes de précision de

classification, nous avons mené 3 expériences pour adapter un ResNet50 avec les différentes stratégies suivantes :

- 1- Expérience 1 : Fine tuning de tout le réseau avec des taux d'apprentissage différents pour la partie convolutifs et la nouvelle couche FC.
- 2- Expérience 2 : Geler toute la partie convolutive et entrainer le nouveau réseau entièrement connecté seulement.
- 3- Expériences 3 : Ré-entraîner une partie des sections convolutives et la couche FC avec des taux d'apprentissage différents. 4 expériences ont été réalisé en gelant les k premières sections avec k= 4,3,2 et 1.

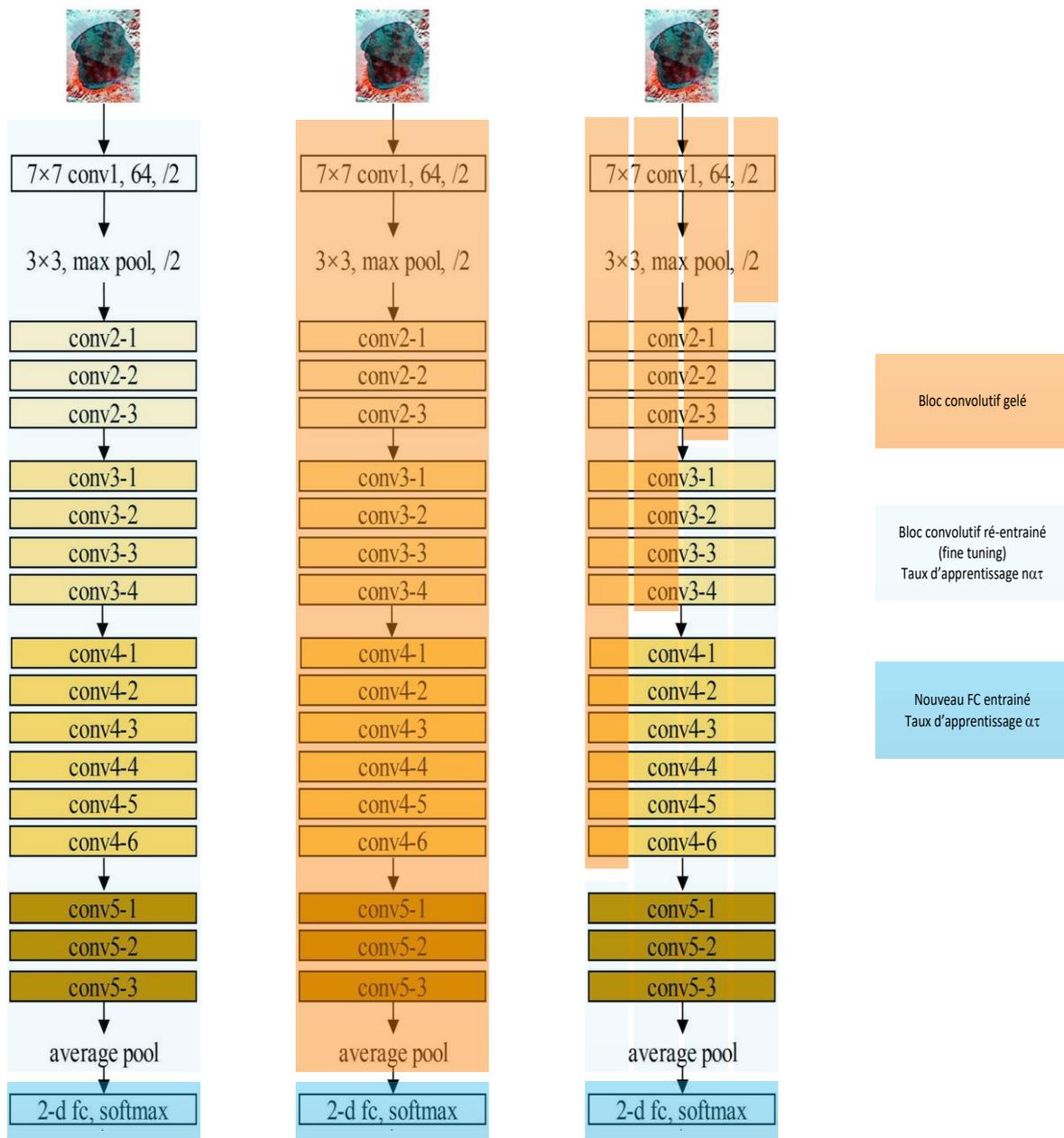


Figure III.3: Stratégie de tranfer learning testées.

### III.5. Pseudo-code d'entraînement

Les étapes de ce script sont les suivantes :

1. Charger la base de données (images d'entraînement et images de validation).
2. Charger le réseau pré-entraîné
3. Saisir la taille de la couche d'entrée (images de  $224 \times 224 \times 3$ )
4. Remplacer les dernières couches en quelques étapes :
  - Convertir le réseau convolutif en un réseau accessible et modifiable.
  - Trouver les couches à remplacer.
  - Remplacer la dernière couche entièrement connectée par une nouvelle couche entièrement connectée avec un nombre de sorties égal au nombre de classes de la base de données.
5. Extraire les couches et les connexions du graphe de couches et sélectionner les couches à figer en fixant le taux d'apprentissage à zéro.
6. Redimensionner les images d'entraînement pour les adapter à la couche d'entrée de ResNet.
7. Spécifier les options d'entraînement comme :
  - La taille du mini-batch.
  - Nombre d'époques d'entraînement.
  - Le taux d'apprentissage.
  - La fréquence de validation.
8. Lancer l'entraînement.

Pour le choix des principaux hyperparamètres d'entraînement, nous avons mené une série d'expériences pour tester des différentes valeurs de la taille de mini batch (16, 32) et taux d'apprentissage ( $10^{-5}$ ,  $10^{-6}$ ) et de  $n$  (10 et 100). L'algorithme d'optimisation du gradient est Adam, le nombre d'epochs est égale à 5 et les autres hyperparamètres sont fixés aux valeurs par défaut. Le taux d'apprentissage est diminué par facteur de 10 après l'époch 3. Après l'analyse des résultats obtenus, nous avons fixé les hyperparamètres aux valeurs suivantes :

- *Taille mini batch* = 32
- *Taux d'apprentissage initial*  $\tau = 5 \times 10^{-4}$
- $n = 10$

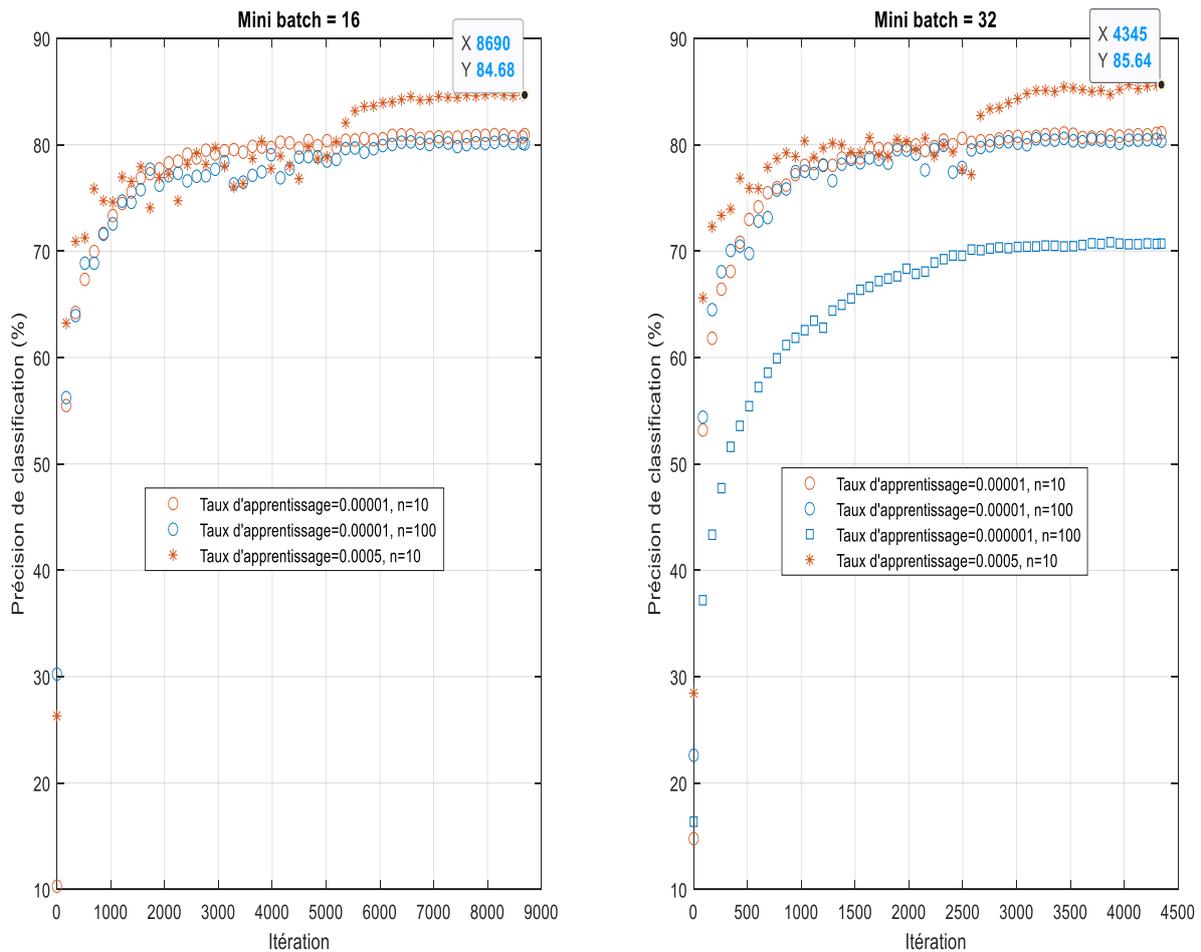


Figure III.4 : Test de performances de quelques valeurs d'hyperparamètres.

### III.6. Résultats des expériences

#### III.6.1. Fine tuning de toute la partie convolutive

Lors de cette expérience, nous avons entraîné la nouvelle couche FC avec un taux d'apprentissage égale à  $5 \times 10^{-3}$  et nous avons ajusté toute la partie convolutive avec un taux d'apprentissage égale à  $5 \times 10^{-4}$ . La figure donne l'évolution de la précision de classification des données d'entraînement et de validation pendant l'apprentissage. Ce modèle final a affiché une précision de plus de 86% sur les données de validation.

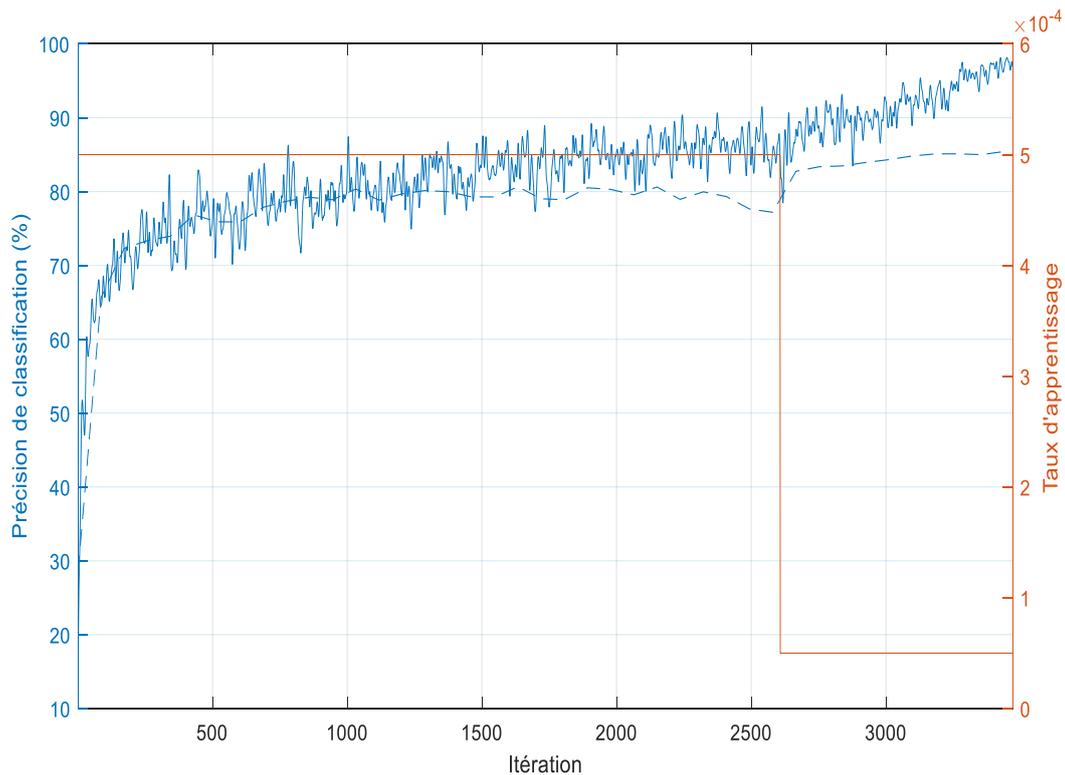


Figure III.5 : Courbe du Fine tuning de ResNet

La matrice de confusion ordonnée de la figure III.6 permet d'étudier la précision de classification de chaque classe. La première chose à remarquer était que la classe avec le score le plus élevé était la classe « heureuse ». Deux facteurs cruciaux ont permis une si bonne prédiction : une variance élevée et un ensemble de données volumineux, le premier étant le plus important ; Cela peut être vu du fait que la classe heureuse n'est pas la plus grande des échantillons d'entraînement, étant dépassée par la classe neutre. Nous remarquons aussi que le réseau confond entre les classes « happy » et « neutral » puisqu'il classe 87 images « happy » comme « neutral ». Pour la classe neutre, le réseau a réussi à bien classer 90,2% des données de validation (proche de la précision de classification de la classe « happy » qui est 91,4%), bien que cette classe détienne la plus grande part de l'ensemble de données. Nous remarquons une confusion de « neutral » avec « happy » et avec « sad ».

True Class	happy	1620	87	16	22	2	21	4	91.4%	8.6%
	neutral	51	2205	27	36	4	118	3	90.2%	9.8%
	surprise	19	43	799	19		5	19	88.4%	11.6%
	angry	20	59	14	607	6	11	3	84.3%	15.7%
	disgust	1	9		13	108	9	1	76.6%	23.4%
	sad	14	172	6	32	2	519	13	68.5%	31.5%
	fear	5	17	40	4	1	14	137	62.8%	37.2%
		93.6%	85.1%	88.6%	82.8%	87.8%	74.5%	76.1%		
		6.4%	14.9%	11.4%	17.2%	12.2%	25.5%	23.9%		
		happy	neutral	surprise	angry	disgust	sad	fear		
		Predicted Class								

Figure III.6 : Matrice de confusion de ResNet50 utilisé.

Les classes « angry », « sad » et « surprises » qui partageaient la même proportion de l'ensemble de données, sont classées avec des précisions différentes. Un nombre important d'images des classes « angry » et « sad » ont été classées comme « neutral ». Cela se produit parce que ces classes avaient une faible variance, en termes de forme de la bouche et des sourcils. La forme de la bouche n'a pas changé de manière significative, tandis que le déplacement des sourcils était difficile à distinguer par le modèle CNN.

Une confusion de classification est enregistrée également entre « fear » et « surprise ». En raison de similitude de ces émotions en termes de forme de la bouche et de faibles changements dans la forme des sourcils, « fear » a été trop mal classée comme surprise. Nous pensons que la classification erronée pourrait être diminuée en augmentant le nombre d'images et en supprimant la classe neutre afin de forcer le modèle à apprendre d'autres caractéristiques distinctes. La figure III.7 donne quelques exemples de classification. Cette figure présente, pour chaque exemple, la classe de l'image et la prédiction du réseau avec la probabilité de la classe prédite.



Figure III.7 : Quelques exemples de classification

À titre démonstratif, nous présentons les cartes de caractéristiques à la sortie des différentes couches du réseau pour l'exemple de l'image donnée par la figure III.8. Pour la première section de convolution (Figure III.9), la figure III.10 donne les 64 filtres de convolution de la section 1. L'application de ces filtres à l'image de test génère 64 cartes de caractéristiques présentées par la figure III.11.(a). La figure III.10.(b) , (c) et (d) montrent, respectivement, les effets de la normalisation, de l'activation ReLU et du max pooling appliqués au résultat de la couche de convolution précédente.

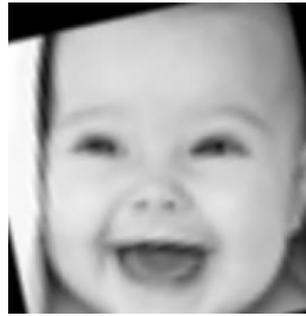


Figure III.8 : Image test

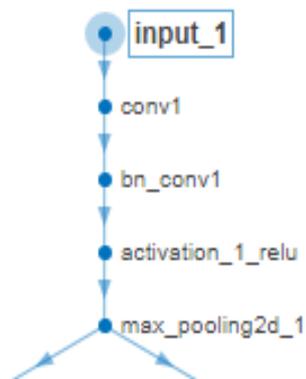


Figure III.9 : représentation graphique des couches de la première section sous Matlab

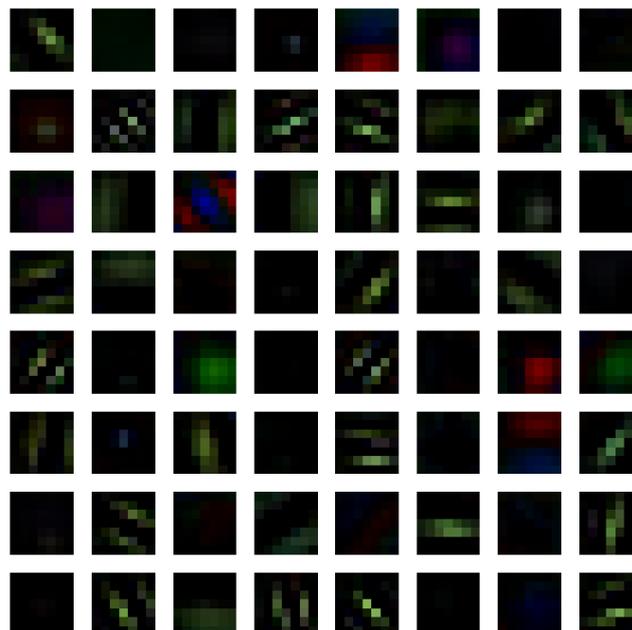


Figure III.10 : Les filtres de convolution de la section 1

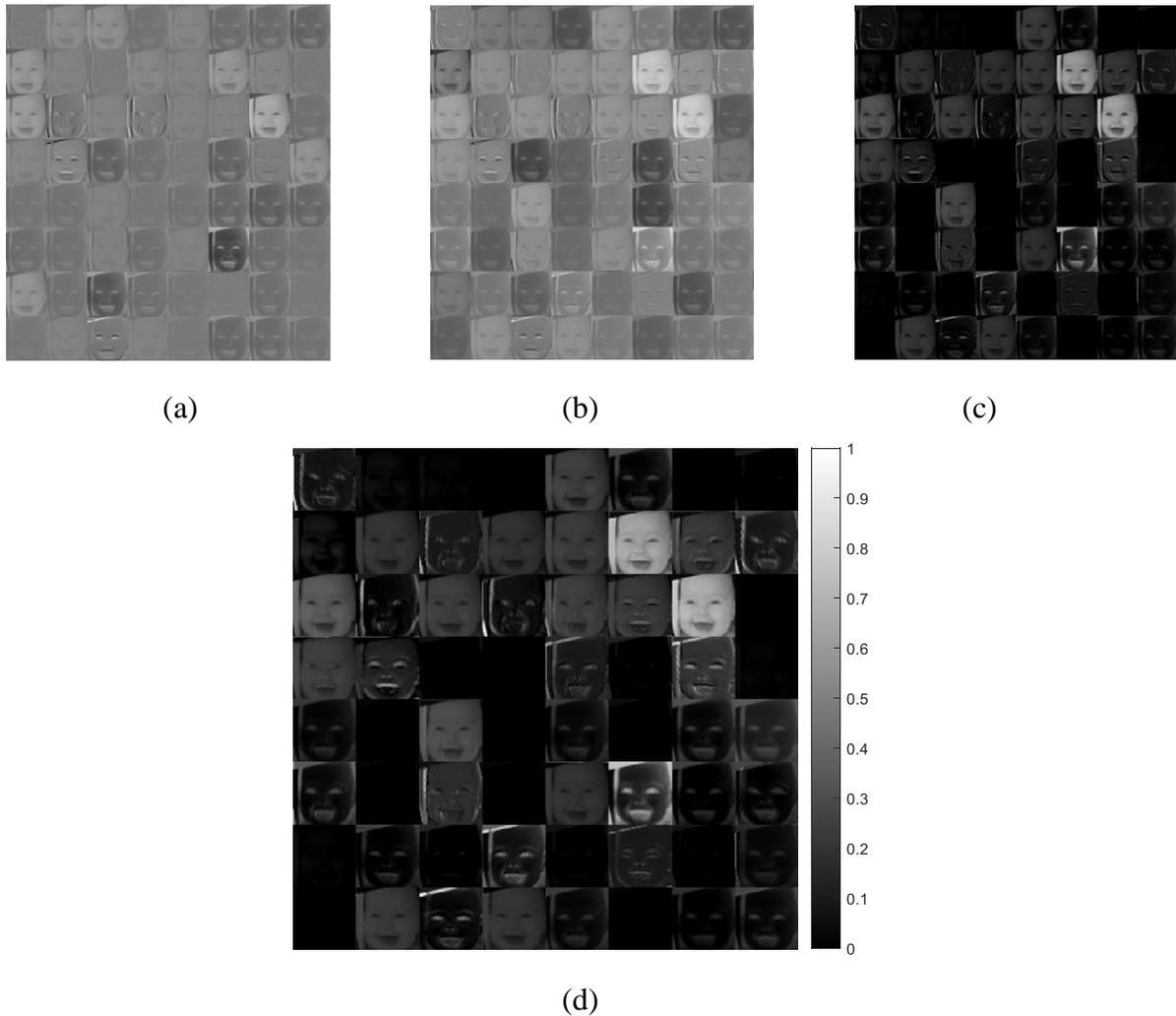
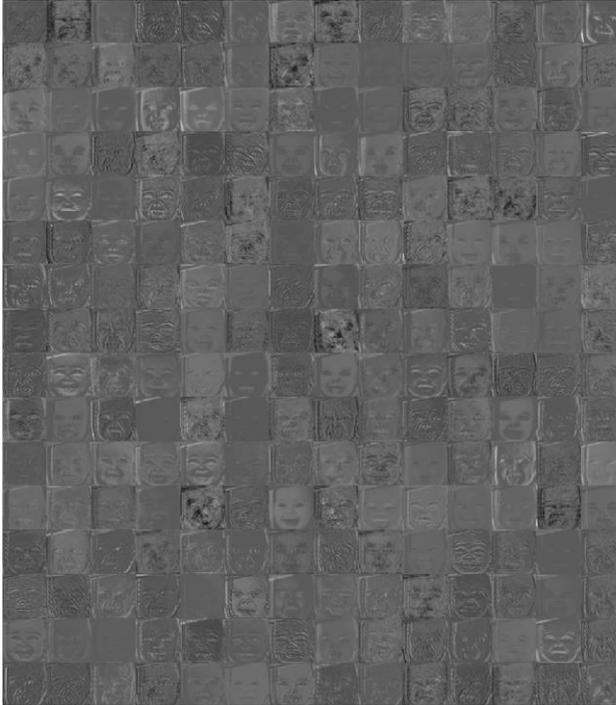
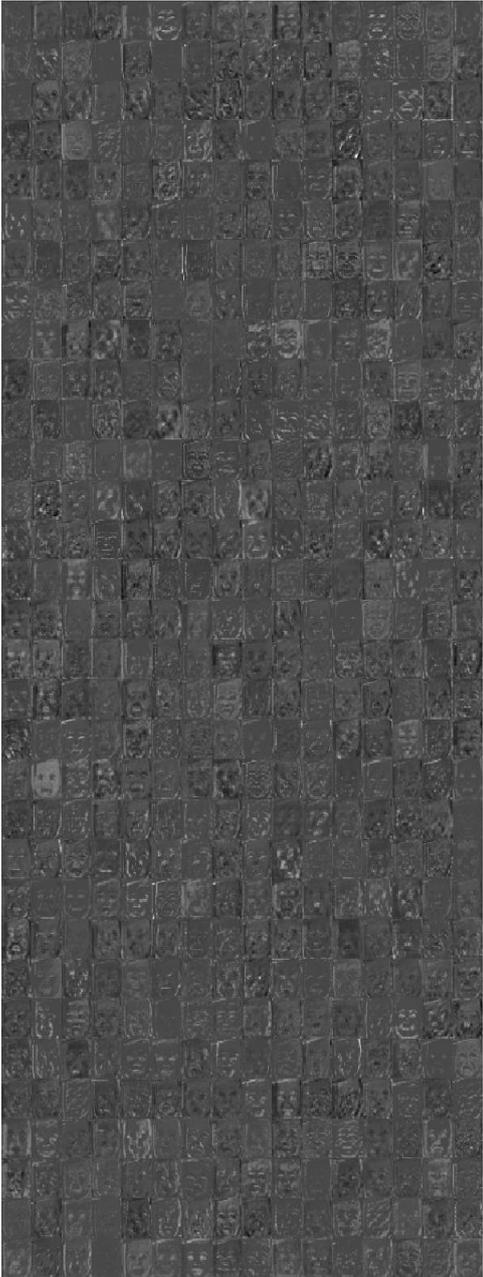


Figure III.11 : Cartes de caractéristiques générées par les couches de la section 1.

La figure III.12 présente les cartes de caractéristiques à la sortie des sections de convolutions. Nous remarquons que les filtres activent des zones particulières. Pour traiter une image labellisée « heureux », nous remarquons ici que les filtres se concentrent sur la bouche et les yeux qui sont des caractéristiques claires de la joie.



(a)



(b)

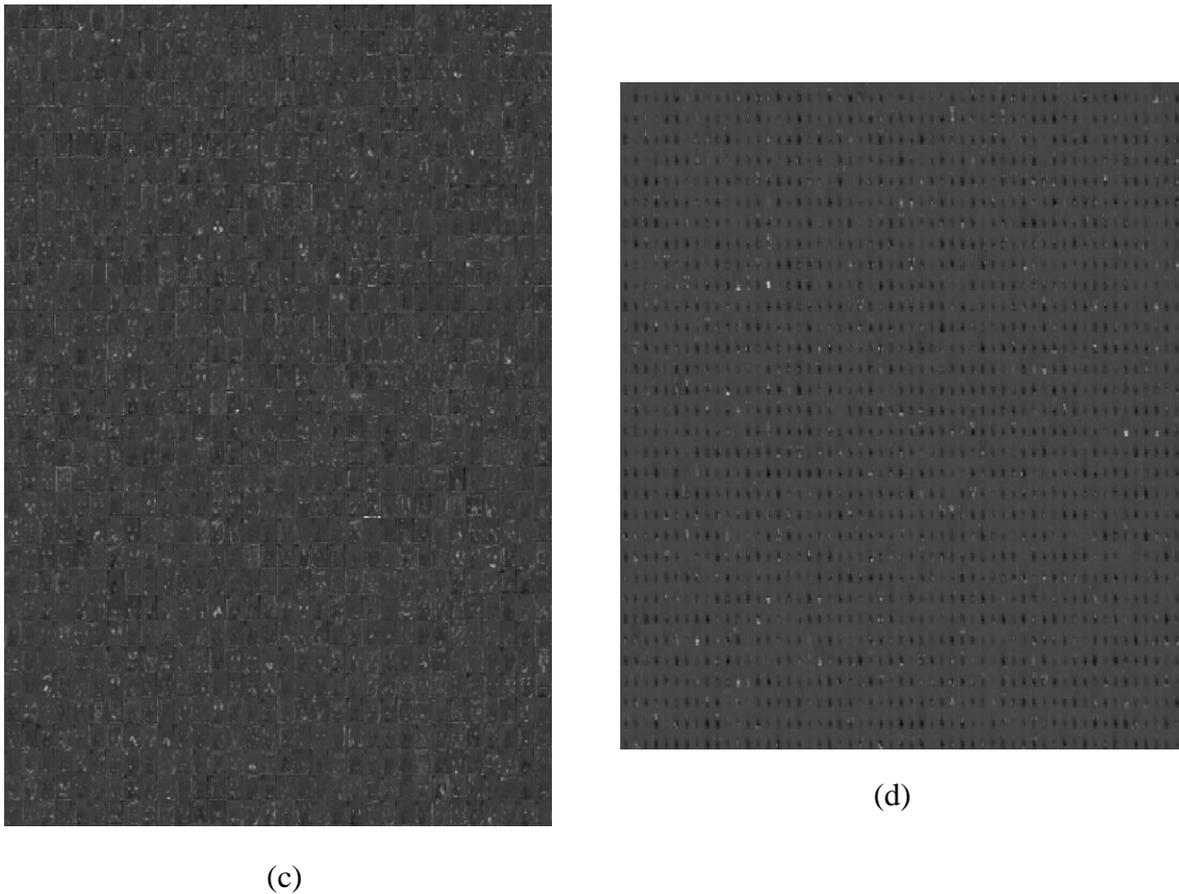


Figure III.12 : Cartes de caractéristiques des sections (a) 2, (b) 3, (c) 4 et 5 (d).

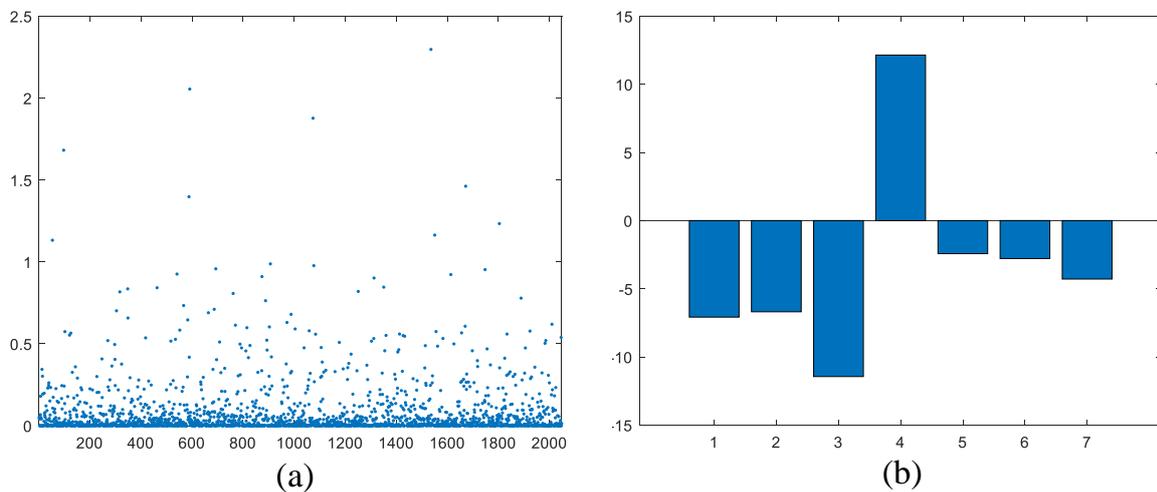


Figure III.13 : Représentation graphique de vecteur de caractéristiques

Au sommet du réseau, le vecteur de caractéristiques final est obtenu après la couche de average pooling (Figure III.13.a). Ce vecteur est appliqué à la couche FC et donne une sortie composée des 7 valeurs (7 est le nombre des classes) représentées par la figure III.13.b. Enfin, l'application de la fonction softmax attribue les probabilités de classes indiquées par la tableau

## III.2.

Classes	Probabilités
1	4.4754285e-09
2	6.6148096e-09
3	5.7145545e-11
4	0.99999905000
5	4.7076188e-07
6	3.2919431e-07
7	7.3074489e-08

Tableau III.2 : Probabilités des 7 classes

**III.6.2. Entraînement du nouveau FC seulement**

Lors de cette expérience, nous avons entraîné la nouvelle couche FC avec un taux d'apprentissage égale à  $10^{-4}$  et nous avons gelé toute la partie convolutive. La figure III.14 donne l'évolution de la précision de classification des données de traitement et de validation pendant l'apprentissage. Le modèle final a atteint une précision 66.9% sur les données de validation. L'analyse de la performance de classification de cette stratégie de transfer learning indique que l'entraînement de la couche FC seulement n'améliore pas la reconnaissance des expressions faciales et le fine tuning de la partie convolutive est nécessaire.

**III.6.3. Fine tuning d'une partie des sections convolutifs**

Lors de ces quatre expériences, nous avons entraîné la nouvelle couche FC avec un taux d'apprentissage égale à  $5 \times 10^{-3}$  et nous avons gelé les premières sections convolutives par partie. La figure III.15 donne l'évolution de la précision de classification des données de validation pendant l'apprentissage pour les quatre cas et ceux des premières expériences. L'analyse de ces résultats montre que le fine tuning des sections convolutives améliore la reconnaissance des expressions faciales. Pour notre application la meilleure performance de classification par le fine tuning des 2 dernières sections ou plus.

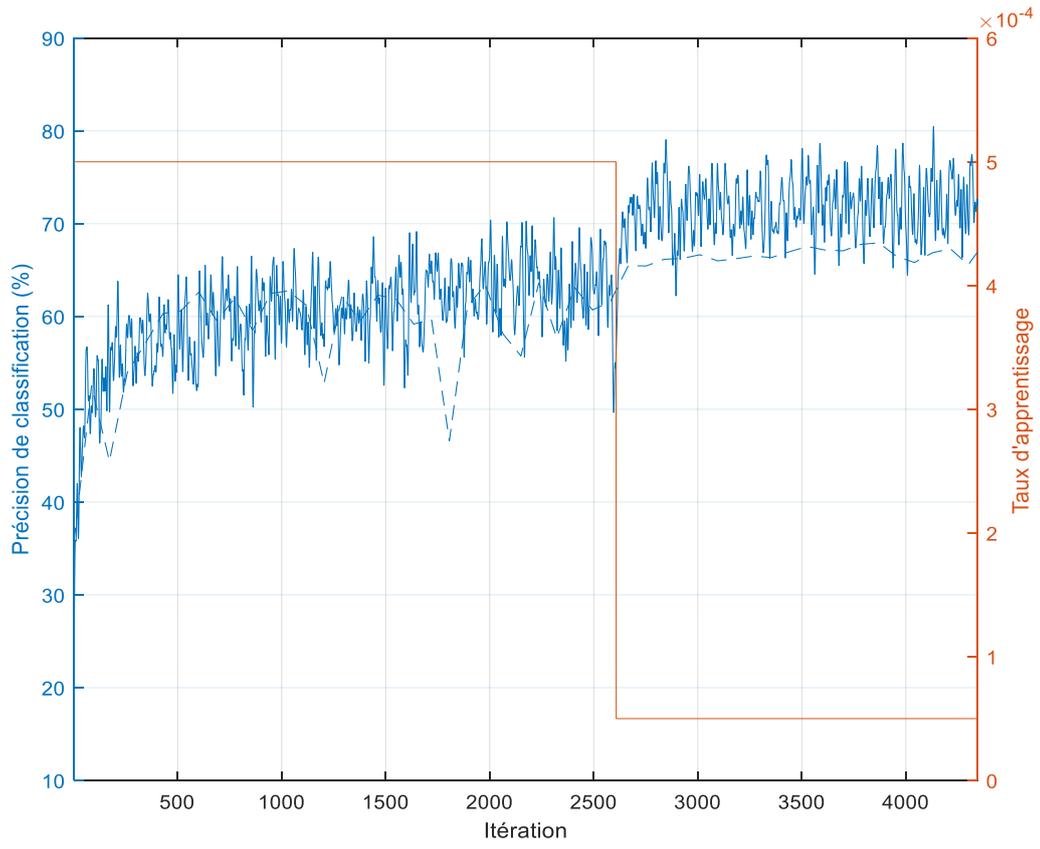


Figure III.14 : Courbe d'entraînement de ResNet (FC seulement)

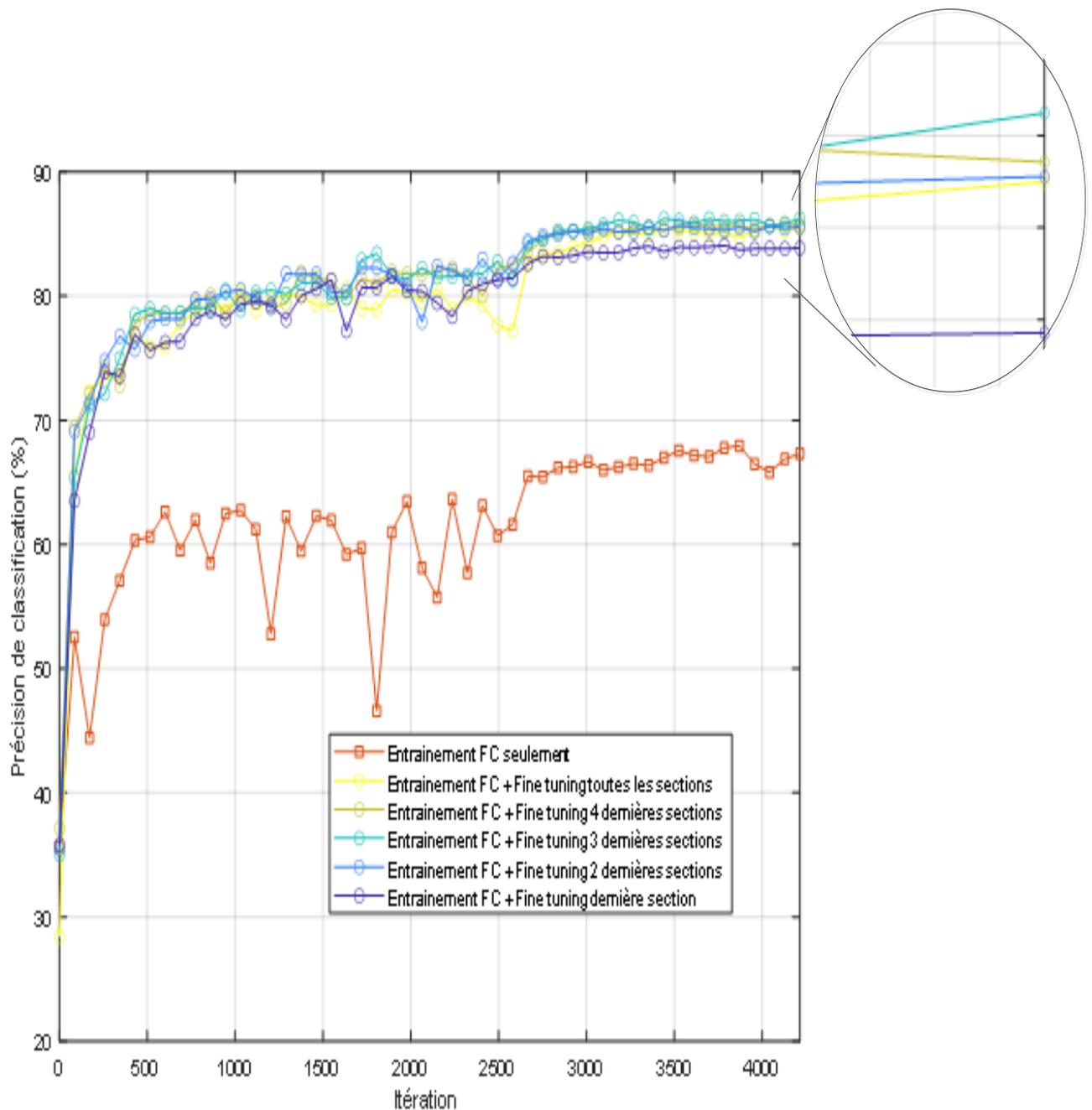


Figure III.15 : Comparaison entre les performances des stratégies d'apprentissage testées.

## Conclusion

Dans ce chapitre, nous avons présenté les résultats des différentes expériences réalisés pour appliquer un CNN à la reconnaissance automatique des expressions faciales. Les résultats montrent que le fine tuning des 2 dernières sections convolutives avec l'entraînement de la nouvelle couche entièrement connectée donne de meilleurs résultats en termes de précision de classification.

# Conclusion générale

## Conclusion Générale

Ce travail est un mémoire de Master réalisé à l'université de Béjaia. Le travail présenté ici s'inscrit dans le domaine de l'analyse automatique des expressions faciales. Ainsi, nous avons utilisé les méthodes de Deep learning pour la reconnaissance des expressions faciales et de les classifier. Ce travail nous a permis de nous confronter à des problèmes interdisciplinaires toujours en relation avec l'automatique et d'enrichir nos connaissances dans le domaine de l'apprentissage automatique.

Dans le chapitre 1, nous avons présenté des généralités sur les expressions faciales. Le deuxième chapitre est consacré au Machine Learning où nous avons présenté les méthodes du Deep Learning qui sont actuellement les plus utilisées dans le domaine de la reconnaissance faciale et d'objet. Leur efficacité nous a motivés à les utiliser dans le domaine de la reconnaissance des expressions faciales. et leur discussion

Le troisième chapitre est consacré à l'application des réseaux de neurones convolutifs (CNN) sur la reconnaissance des expressions faciales. Nous y présentons une étude comparative sur les résultats obtenus lors des différentes expériences réalisées pour appliquer un réseau CNN à la reconnaissance automatique des expressions faciales. Notre étude montre que le fine tuning, avec l'entraînement de la nouvelle couche entièrement connectée, donne de meilleurs résultats en termes de précision de classification.

Ce travail nous a permis de mettre en pratique nos connaissances sur les réseaux de neurones, le Deep Learning et la reconnaissance des expressions faciales.

# **Bibliographie**

## Bibliographie

- [1] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak , “Emotional Expressions [Reconsidered]: Challenges to Inferring Emotion From Human Facial Movements”. *Psychological Science in the Public Interest*, July 2019, Vol. 20(1) 1-68.
- [2] Dawood Al Chanti. “Automatic Analysis of Macro and Micro Facial Expressions : Detection and Recognition via Machine Learning. Signal and Image processing”, Thèse de doctorat, Université Grenoble Alpes, 2019. <https://tel.archives-ouvertes.fr/tel-02525707> consulté Septembre 2020.
- [3] Ekman, P. and Friesen, “Facial action coding system”. Consulting Psychologists Press, vol 3 , 90-28, 1977.
- [4] Ekman, P. and Friesen, “Manual for the facial action coding system”. Consulting Psychologists Press, .....
- [5] Ekman, P. and Friesen, W. V. (2003). *Unmasking the face : A guide to recognizing emotions from facial clues*. Ishk.
- [6] Ekman, P. and Friesen, W. (1978a). *Facial action coding system : a technique for the measurement of facial movement*. Palo Alto : Consulting Psychologists.
- [7] Valstar, M. and Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. In 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), pages 149–149. IEEE.
- [8] Viola, P and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.
- [9] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object
- [10] ). Ica and gabor representation for facial expression recognition. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–855. IEEE.
- [11] Smith, B. M., Zhang, L., Brandt, J., Lin, Z., and Yang, J. (2013). Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,

pages3484–3491.

- [12] Xiong,X.and DelaTorre,F.(2013).Supervised descent method and its applications to face alignment.In Proceedings of the IEEE conference on computer vision and pattern recognition, pages532–539.
  - [13] Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3476–3483.IEEE.
  - [14] Stathopoulou, I.-O. and Tsihrintzis, G. A. (2004). An improved neural-network- based face detection and facial expression classification system. In Systems, Man and Cybernetics,2004IEEE International Conference on volume1,pages666–671.IEEE.
  - [15] Fasel,B.andLuetin,J.(2003).Automatic facial expression analysis : asurvey. Pattern recognition, 36(1):259–275.
  - [16] Ghimire, D., Jeong, S., Lee, J., and Park, S. H. (2017). Facial expression recognition based on local region specific features and support vector machines. Multimedia Tools and Applications, 76(6):7803–7821.
  - [17] Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition In Computer Vision (ICCV),2015 IEEE International Conference on, pages 2983–2991.IEEE.
  - [18] Deng, L., & Yu, D. (2014). Deep learning: methods and applications. Foundations and Trends® in Signal Processing, 7(3–4), 197-387.
  - [19] Bengio, Y. (2009). Learning deep architectures for AI. Foundations and trends in Machine Learning, 2(1), 1-127
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.
- [20] J. Schmid Huber, “Deep Learning.,” Scholarpedia, vol. 10, no. 11, p. 32832, 2015
  - [21] Russakovsky, O., Deng, J., Su, H. et al., ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211–252 (2015). <https://doi.org/10.1007/s11263-015-0816-y>
  - [22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
  - [23] Thorsten Hoeser , Claudia Kuenzer , Object Detection and Image Segmentation with Deep Learning on Earth Observatio Data: A Review-Part I: Evolution and Recent Trends, Remote

Sens. 2020, 12, 1667, doi:10.3390/rs12101667.

- [24] Mark Hudson Beale, Martin T. Hagan , Howard B. Demuth, “Deep Learning Toolbox User's Guide”, 2020 by The MathWorks, Inc, online [https://ch.mathworks.com/help/pdf\\_doc/deeplearning/nnet\\_ug.pdf](https://ch.mathworks.com/help/pdf_doc/deeplearning/nnet_ug.pdf) (consulté septembre 2020)
- [25] <https://www.kaggle.com/qnkhuat/emotion-compilation> (consulté septembre 2020)
- [26] Daniel Octavian Melinte and Luige Vladareanu, “Facial Expressions Recognition for Human–Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer”, Sensors 2020, 20, 2393; doi:10.3390/s20082393

## Résumé

Pour les êtres humains, l'expression du visage est l'un des plus puissants et naturels moyens pour communiquer leurs émotions et leurs intentions. Un être humain est capable de détecter les expressions du visage sans effort, mais, pour une machine, cette tâche est très difficile. La reconnaissance automatique des expressions faciales est un problème intéressant. Les systèmes de reconnaissance automatique des expressions du visage peuvent être utilisés principalement pour l'interaction homme-machine. Dans ce travail nous présentons la mise en œuvre d'un système de reconnaissance des expressions faciales, les étapes principales de ce système sont : l'extraction des paramètres caractéristiques et la classification de l'expression faciale. Nous avons utilisé le principe des réseaux de neurones convolutifs qui sont une méthode très populaire du deep learning pour la détection et le suivi faciale, le tranfer learning pour. Un système pré-entraîné pour résumer le temps d'entraînement et avoir un système de classification efficace avec un nombre de données d'entraînement plus réduit.

## Abstract

For human beings, facial expression is one of the most powerful and natural ways to communicate their emotions and intentions. A human being is able to detect facial expressions effortlessly, but for a machine this task is very difficult. The automatic recognition of facial expressions is an interesting problem. The systems for automatic recognition of facial expressions can be mainly used for human-machine interaction. In this work we present the implementation of a facial expression recognition system, the main steps of this system are: the extraction of the characteristic parameters and the classification of the facial expression. We used the principle of convolution neural networks which are a very popular method of deep learning for facial detection and tracking, transfer learning for a pre-trained system to summarise the training time and have an efficient classification system with a smaller amount of training data.

## ملخص

يستطيع الإنسان اكتشاف تعابير الوجه . بالنسبة للبشر ، يعد تعبيرات الوجه من أقوى الطرق للتعبير عن مشاعرهم ونواياهم يمكن يعد التعرف التلقائي على تعابير الوجه مشكلة مثيرة للاهتمام بسهولة ، ولكن هذه المهمة صعبة للغاية بالنسبة للألة نقدم في هذا العمل تنفيذ نظام . استخدام أنظمة التعرف التلقائي على تعابير الوجه بشكل أساسي للتفاعل بين الإنسان والآلة . استخراج المعلومات المميزة وتصنيف تعبيرات الوجه : التعرف على تعبيرات الوجه ، والخطوات الرئيسية لهذا النظام هي استخدمنا مبدأ الشبكات العصبية الالتفافية التي تعد طريقة شائعة جداً للتعلم العميق لاكتشاف الوجه وتتبعه ، ونقل التعلم لنظام مدرب مسبقاً لتلخيص وقت التدريب ولديه نظام تصنيف فعال بكمية أقل من بيانات التدريب