Université 20 Aout 1955- Skikda

Faculté des Sciences

Département Informatique

Mémoire

En vue de l'obtention du diplôme de Master en Informatique

Filière : Informatique

Spécialité : Intelligence Artificielle

Vision Transformers pour la Reconnaissance des Emotions à partir des Expressions Faciales

Présenté par

BOUGHAMOU Imane

SOUAMES Rayane

Encadré Par : Dr. HAZMOUNE Samira

Année universitaire 2023/2024

Remerciment

Merci à Allah de nous avoir accordé la capacité d'écrire et de réfléchir, la force de croire en nos ambitions, la patience d'atteindre nos rêves et la joie de lever nos mains vers le ciel en disant "Alhamdoulillah".

En premier lieu, ce travail n'aurait pas été aussi abouti sans l'aide et l'encadrement précieux de notre encadrante, Dr. Samira Hazmoune. Nous lui exprimons notre gratitude pour la qualité exceptionnelle de son encadrement, sa patience et sa disponibilité tout au long de la préparation de ce mémoire.

Nous sommes également reconnaissants envers chaque membre du jury pour l'honneur qu'ils nous font en acceptant de juger notre travail. Notre gratitude sincère va aussi aux professeurs du département d'informatique qui ont contribué à notre formation au cours des cinq dernières années détudes.

Avant de conclure cette page de remerciements, nous tenons à exprimer notre reconnaissance envers nos parents pour leur patience et leur soutien inconditionnel, sans lesquels nous n'aurions pas pu mener à bien notre travail. Enfin, nous adressons nos remerciements à nos familles, nos collègues et à toutes les personnes qui ont contribué directement ou indirectement à nous encourager et à nous soutenir moralement à la fin de ce travail.

Merci à tous ceux qui ont contribué à cette réussite.

Dédicace

Je dédie ce travail

À ma chère maman Merci d'avoir été là pour moi à chaque étape de ma vie, de m'avoir soutenue et encouragée sans relâche.Ta présence chaleureuse et ton soutien sans faille ont été mes piliers dans les moments difficiles. Ta force et ta gentillesse ont toujours été un exemple pour moi, me motivant à persévérer et à viser toujours plus haut. Ce mémoire est ma façon de te remercier pour tout ce que tu as fait pour moi.

À mon cher papa À toi, qui m'as toujours inspiré et soutenu à chaque étape de ma vie, même après ton départ. Ta sagesse et ton amour inconditionnel ont profondément marqué mon chemin et continuent d'illuminer ma route.Ce mémoire est dédié à toi, en souvenir de ton influence positive. Tes valeurs et ton amour resteront gravés dans mon cur pour toujours. À travers ces pages, je veux rendre hommage à ta mémoire et te remercier pour tout ce que tu as fait pour moi. Tu seras toujours présent dans mon cur et dans mes pensées.

À mon cher frère À travers nos rires partagés et les moments où tu m'as soutenu, ta présence dans ma vie est un cadeau précieux qui m'inspire chaque jour à être meilleur. Merci d'être toujours là pour moi.

A mon binôme Je souhaite exprimer toute ma gratitude envers Imane, à la fois mon binôme et ma meilleure amie, pour son soutien précieux tout au long de ce projet. Sa contribution a été indispensable à chaque étape, nous permettant de relever les défis et d'atteindre nos objectifs grâce à notre solide collaboration. Merci infiniment pour ton engagement sans faille.

RAYANE

Je dédie ce travail à **"Allah"** le tout puissant, mon créateur, mon pilier solide, ma source d'inspiration, de sagesse, de connaissance et de compréhension. Il a été la source de ma force tout au long de ce programme.

À mes très chers parents : À mes parents, qui ont contribué à ma réussite par leur amour et leur soutien tout au long de ma carrière scolaire, pour tous les sacrifices qu'ils ont faits et leurs précieux conseils. Pour toute leur aide et leur présence dans ma vie, je leur dédie cette humbl oeuvre comme expression de mes sentiments et de ma gratitude éternelle.

À mes frères et soeurs : Aucune dévotion ne pourra jamais exprimer suffisamment ce que je ressens pour vous, je voudrais simplement dire merci beaucoup, je vous aime.

À mes chers amis : Comme témoignage de l'amitié sincère que j'ai eue et des bons moments passés ensemble. Je vous dédie ce travail ; je vous souhaite un avenir radieux rempli de bonnes promesses.

IMANE

Résumé

Au cours de la dernière décennie, les systèmes intelligents ont fourni des services publiques, entraînant un changement important dans l'environnement de vie des gens, mais il reste encore de nombreux problèmes à résoudre pour s'assurer qu'ils travaillent ensemble de manière sûre, fiable et efficace. Notre mémoire portait sur la compréhension des émotions, une caractéristique essentielle qui reflète les aspects humains du comportement humain.

Nous avons utilisé le modèle pré-entraîné Vision Transformer (ViT), qui est l'une des dernières technologies basées sur la structure du Transformer pour atteindre des performances exceptionnelles dans les tâches de reconnaissance d'image, dépassant de nombreux modèles traditionnels dans ce domaine. ViT a s'appuie sur la division de l'image en petites parties et son traitement à l'aide de la technologie de Transformer pour comprendre le contexte plus en profondeur et améliorer la précision de la prédiction.

Nous avons entraîné le modèle plusieurs fois pour obtenir les meilleurs résultats dans la classification des émotions à travers les expressions faciales. Nous avons mené une série d'expériences sur l'ensemble de données Extended Cohn Kanade (CK+) pour déterminer les paramètres optimaux pour la structure Vit, où la précision de classification a atteint 100%. Nous avons également évaluer notre système sur un autre jeu de données OAHEGA, où la précision de classification était de 90%. Ces résultats sont très satisfaisants, notamment lorsqu'on les compare à certains travaux récents de la littérature.

Mots clés : Vision Transformer ViT, Reconnaissance des émotions, Expressions faciales, CK+, OAHEGA.

Abstract

Over the past decade, intelligent systems have provided public services, leading to a significant change in people's living environment. However, there are still many issues to be resolved to ensure that they work together in a safe, reliable, and efficient manner. Our paper focused on understanding emotions, a fundamental characteristic that reflects the human aspects of human behavior.

We used the pre-trained Vision Transformer ViT model, which is one of the latest technologies based on the Transformer structure to achieve exceptional performance in image recognition tasks, surpassing many traditional models in this field. ViT relies on dividing the image into small parts and processing them using Transformer technology to understand the context more deeply and improve prediction accuracy.

We trained the model multiple times to obtain the best results in classifying emotions through facial expressions. We conducted a series of experiments on the Extended Cohn-Kanade CK+ dataset to determine the optimal parameters for the ViT structure, where the classification accuracy reached 100%. We also evaluated our system on another dataset, OAHEGA, where the classification accuracy was 90%. These results are very satisfactory, especially when compared to traditional models in the field. These results are very satisfactory, especially when compared to some recent works in the literature.

Keywords : Vision Transformer ViT, Emotion Recognition, Facial Expressions, CK+, OAHEGA.

ملخص

على مدى العقد الماضي، قدمت الأنظمة الذكية خدمات عامة، مما أدى إلى تغيير كبير في بيئة حياة الناس، ولكن لا يزال هناك العديد من المشكلات التي يجب حلها لضمان عملها معًا بطريقة آمنة وموثوقة وفعالة. تناولت مذكرتنا فهم العواطف، وهي خاصية أساسية تعكس الجوانب البشرية للسلوك البشري.

لقد استخدمنا النموذج المدرب مسبقًا (ViT)، وهو إحدى أحدث التقنيات القائمة على بنية المحولات لتحقيق أداء استثنائي في مهام التعرف على الصور، متفوقًا على العديد من النماذج التقليدية في هذا المجال. يعتمد ViT على تقسيم الصورة إلى أجزاء صغيرة ومعالجتها باستخدام تقنية المحولات لفهم السياق بعمق أكبر وتحسين دقة التنبؤ.

قمنا بتدريب النموذج عدة مرات للحصول على أفضل النتائج في تصنيف العواطف من خلال تعابير الوجه. أجرينا سلسلة من التجارب على مجموعة البيانات الموسعة "كون كاندي (CK+)" لتحديد المعايير المثلى لبنية ViT ، حيث وصلت دقة التصنيف إلى 100%. كما قمنا بتقييم نظامنا على مجموعة بيانات أخرى"OAHEGA" ، حيث كانت دقة التصنيف 90 %. هذه النتائج مرضية للغاية، خاصة عند مقارنتها ببعض الأعمال الحديثة في الأدبيات.

الكلمات المفتاحية: " المحولات البصرية" (ViT)، التعرف على العواطف، تعابير الوجه، +CK، OAHEGA.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

BERT Bidirectional Encoder Representations from Transformers. 29, 30

CK+ Extended Cohn Kanade. IV, V

CNN Convolutional Neural Network. 21, 22

DL Deep Learning. 17

DPT Dense Prediction Transformers. 34

FFN Feed Forward Network. 28, 29

GELU Gaussian Error Linear Unit. 16

GPT Generative Pretrained Transformer. 29, 30

GRU Gated Recurent Units. 21

HMM Hidden Markov Model. 12

IA Intelligence Artificielle. 2, 5

ILSVRC Imagenet Large Scale Visual Recognition Challenge. 24

KNN K-Nearest Neighbors. 11

LSTM Long Short Term Memory. 18–20

ML Machine Learning. 6

MLP Multilayer Perceptron. 13, 14, 32, 33

MSA Multi head Self Attention. 32

PAD Plaisir Activation dominance. 48

RNN Recurrent Neural Network. 18–20

SVM Support Vector Machine. 10

TALN Traitement Automatique du Langage Naturel. 30

TL Transfer Learning. 35

ViT Vision Transformer. 31, 32, IV, V, VIII

INTRODUCTION GÉNÉRALE

Contexte et problématique

La reconnaissance des émotions à partir des expressions faciales dans les images est un domaine de recherche essentiel pour comprendre les états émotionnels des individus. Cela revêt une importance particulière dans divers domaines tels que la psychologie, la santé mentale et les interfaces homme-machine. En identifiant et en comprenant les émotions exprimées par les visages, il devient possible d'améliorer la communication, de détecter les signaux de détresse et de concevoir des systèmes interactifs plus intuitifs et adaptatifs.

Dans cette optique , les systèmes de reconnaissance des émotions basés sur l'Intelligence Artificielle (IA) ont émergé comme des outils prometteurs pour améliorer la compréhension et l'interprétation des expressions faciales dans les images. L'IA, et plus particulièrement, l'apprentissage automatique combiné à des techniques de vision par ordinateur, permet d'extraire des informations émotionnelles précieuses à partir des visages capturés par des caméras ou des dispositifs d'imagerie. Cela ouvre la voie à de nouvelles possibilités pour mieux comprendre les émotions humaines, soutenir la santé mentale et développer des applications interactives plus empathiques et personnalisées.

Les Transformers ont montré qu'ils peuvent accomplir beaucoup dans le traitement du langage naturel (NLP). Cependant, leur utilisation pour le traitement des images est relativement nouvelle. Le Vision Transformer (ViT), apparu en 2020, est un modèle spécialement conçu pour classer des images. Contrairement aux réseaux de neurones convolutifs classiques (CNN), ViT ne fait pas appel à de couches de convolution mais divise plutôt les images

en séquences de petits morceaux (patches). Cette téchnique lui permet de s'adapter facilement à des tailles d'images variées sans sacrifier la puissance de traitement des convolutions. De plus, ViT utilise des mécanismes d'attention multi-têtes (multi-head attention) pour se concentrer sur les parties importantes de l'image, ce qui lui permet d'apprendre des représentations d'image efficaces. ViT a obtenu des résultats impressionnants sur différents tests de classification d'images, ce qui montre à quel point les modèles basés sur les Transformers peuvent être performants dans le domaine de la vision par ordinateur.

Objectifs et contribution

L'objectif principal de ce travail est de mettre en place un modèle basé sur ViT pour améliorer la performance des systèmes de classification des émotions à partir des expressions faciales. Afin d'atteindre cet objectif, nos contributions sont résumées en ce qui suit :

— Mettre en place un modèle ViT pour classifier les émotions à partir des expressions faciales, en distinguant différentes émotions : la joie, la tristesse, la colère, la surprise, la peur et le dégoût.

— Effectuer une étude expérimentale, sur les ensembles de données OAHEGA et CK+, pour optimiser les paramètres du modèle et évaluer leur impact sur la précision du système.

— Mener une analyse comparative pour démontrer l'efficacité des modèles ViT pour la classification des émotions à partir des expressions faciales, en les comparant à d'autres travaux précédents sur OAHEGA et CK+.

Plan de mémoire

Ce mémoire comprend une introduction générale, trois chapitres distincts, et une conclusion générale.

— Chapitre 1 : Il offre un aperçu général des techniques de l'apprentissage automatique et de l'apprentissage profond. Il met en évidence le ViT, la méthode choisie pour cette étude.

— Chapitre 2 : Il décrit le fondement théorique des émotions. En effet, dans ce chapitre, nous définissons les émotions et explorons leurs divers types, la modélisation, les principales sources des données émotionnelles, ainsi que le processus de reconnaissance des émotions.

— Chapitre 3 : Il présente notre système en détail, ainsi que les différentes expérimentations réalisées et l'analyse des résultats obtenus. Nous débutons par décrire l'architecture de notre système de classification des émotions à partir des expressions faciales. Ensuite, nous menons plusieurs expérimentations pour ajuster les hyper paramètres de notre modèle et évaluer ses performances. Enfin, nous comparons les performances de notre modèle avec celles reportées dans des travaux similaires publiés dans la littérature.

CHAPITRE 1

TECHNIQUES D'APPRENTISSAGE EN INTELLIGENCE ARTIFICIELLE

1.1 Introduction

L' intelligence artificielle IA est une science qui a émergé il y a environ trois décennies en tant que discipline expérimentale. Actuellement, elle englobe un large spectre de domaines, de technologies et de méthodologies, et a engendré de multiples techniques qui sont largement intégrées dans divers secteurs des sciences et des technologies, tels que la médecine, le transport, lindustrie, léducation et la finance.

Dans ce chapitre, nous commencerons par expliquer ce que signifie l'intelligence artificielle.

Ensuite, nous examinerons les différentes techniques couramment employées pour détecter les émotions faciales à l'aide de l'intelligence artificielle. Ces techniques sont classées en deux types : le machine learning et le deep learning, dont les Transformers et plus particulièrement, les Vision Transformers (ViT) represente une technologie de pointe. Ces derniers sont sélectionnés pour concevoire notre modèle de classification des émotions à partir des

expressions faciales.

1.2 Définition de lintelligence artificielle

En 1956, le terme "Intelligence Artificielle" émergea après des années de recherches initiées dans le sillage de la Seconde Guerre mondiale. Son introduction lors d'une conférence au Dartmouth College fut attribuée à McCarthy [1]. Elle joue un rôle de plus en plus important grâce à ses capacités d'analyse et d'apprentissage, Se positionnant ainsi en l'un des domaines d'étude les plus récents entre les sciences et l'ingénierie. L'intelligence artificielle est une branche de l'informatique qui vise à créer des machines intelligentes, Concurrençant l'intelligence humaine, et qui englobe l'apprentissage automatique et l'apprentissage profond. Ceci s'explique par l'amélioration de l'efficience opérationnelle, l'accroissement de la productivité et l'optimisation de l'utilisation des ressources. Grâce à l'utilisation d'algorithmes avancés et de modèles prédictifs, cela favorise une anticipation des données et une autonomie décisionnelle. C'est un domaine en constante évolution, où il acquiert jour après jour les capacités cognitives humaines et dépasse ce que les meilleurs experts de leur domaine peuvent accomplir. Elle permet de prendre des décisions complexes, d'optimiser des processus et d'améliorer l'efficacité de nombreuses tâches telles que la traduction automatique, la sécurité informatique, la reconnaissance d'objets, la gestion des ressources humaines et l'assistance à la clientèle.

1.3 Apprentissage automatique (Machine Learning)

Arthur Samuel, un précurseur dans le domaine de l'intelligence artificielle et des jeux vidéo, a introduit en 1959 le concept révolutionnaire de "Apprentissage automatique". Il l'a décrit comme « le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmé » [2]. L'apprentissage automatique, ou Machine Learning (ML), est un sous-domaine clé de l'intelligence artificielle, utilisant des algorithmes et des méthodes statistiques pour créer des modèles à partir de données d'entraînement. Son objectif principal est d'améliorer les algorithmes afin de permettre aux machines d'apprendre à partir de vastes ensembles de données et de prendre des décisions précises. En permet-

tant aux systèmes d'évoluer grâce à l'expérience, sans avoir besoin d'une programmation explicite, l'IA soutenue par l'apprentissage automatique ouvre de nouveaux horizons. Ce domaine révolutionne la technologie de l'IA, offrant un potentiel immense pour résoudre des problèmes complexes et traiter des données variées. Son utilisation croissante dans des domaines comme la médecine, les systèmes de recommandation et la vision par ordinateur [3], montre son rôle crucial dans l'évolution de l'intelligence artificielle. Les principaux types d'apprentissage automatique incluent l'apprentissage supervisé, non supervisé, semi-supervisé et par renforcement.

1.3.1 Les différents types d'apprentissage

L'apprentissage automatique propose divers types adaptés à différents défis, nécessitant des niveaux de guidage variés. Ces types illustrée dans la figure (1.1), se divisent en quatre catégories principales selon le niveau de supervision requis :supervisé, non supervisé, semi-supervisé et par renforcement.

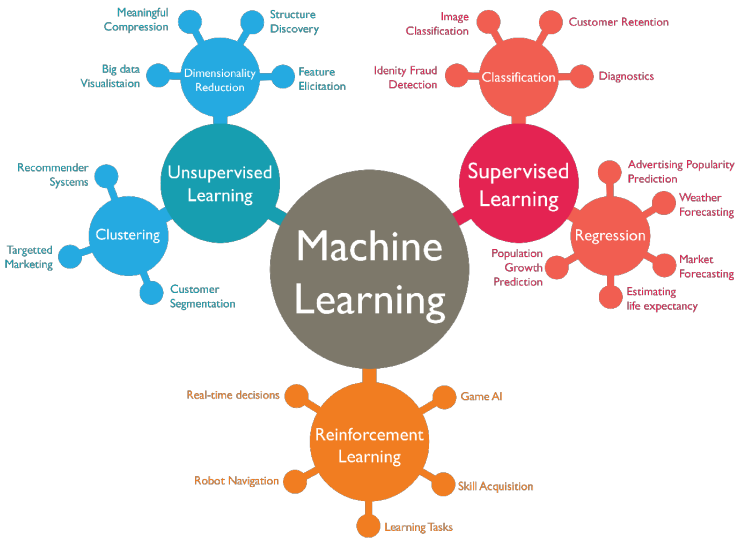

FIGURE 1.1 – Spectre des types de l'apprentissage automatique [4]

1.3.1.1 Apprentissage supervisé (Supervised Learning)

Dans le domaine de l'apprentissage automatique, l'apprentissage supervisé consiste à créer une fonction qui associe des entrées à des sorties, en se basant sur des exemples d'en-

traînement étiquetés. Ces algorithmes nécessitent une intervention externe et utilisent un ensemble de données séparé en deux parties : les données d'entraînement et les données de test. Les données d'entraînement contiennent une variable cible que l'on cherche à prédire ou à classifier. Les algorithmes apprennent à partir des schémas présents dans ces données d'entraînement pour ensuite appliquer ces connaissances aux données de test, dans le but de faire des prédictions ou des classifications [5]. Ce type d'apprentissage est particulièrement utile pour des tâches de classification et de régression. En classification, il permet d'assigner des valeurs d'entrée à différentes catégories [6], tandis qu'en régression, il sert à prédire des valeurs continues [7]. En apprentissage supervisé, on cherche donc à prédire une valeur cible à partir d'observations d'entrée, les entrées du modèle étant appelées "caractéristiques" et les valeurs cibles à prédire étant souvent désignées par le terme "étiquettes" (features) [8]. comme le montre la figure (1.2).

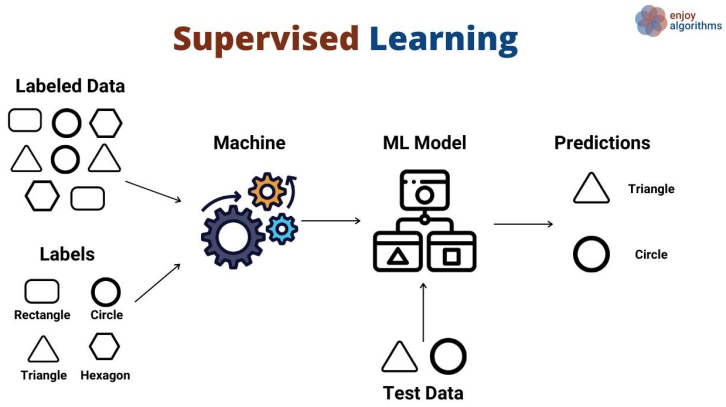

FIGURE 1.2 – Apprentissage supervisé
[9]

1.3.1.2 Apprentissage non supervisé (Unsupervised Learning)

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne repose pas sur des directives explicites données aux algorithmes. Comme illustré dans la figure, les algorithmes doivent eux-mêmes découvrir les structures importantes dans les données brutes pour les classer. Cela permet d'identifier les structures sous-jacentes, simplifiant ainsi les caractéristiques et rendant l'exploration des ensembles de données complexes plus efficace. Parmi les méthodes couramment utilisées, on trouve le clustering, la réduction de la dimensionnalité et l'analyse des composantes principales [5]. Dans cette section, nous nous concentrons sur

le clustering par partitionnement, où les données sont divisées en k groupes. Chaque groupe contient au moins une donnée, et chaque donnée est associée à un seul groupe, comme le montre la classification des légumes dans la (1.3).

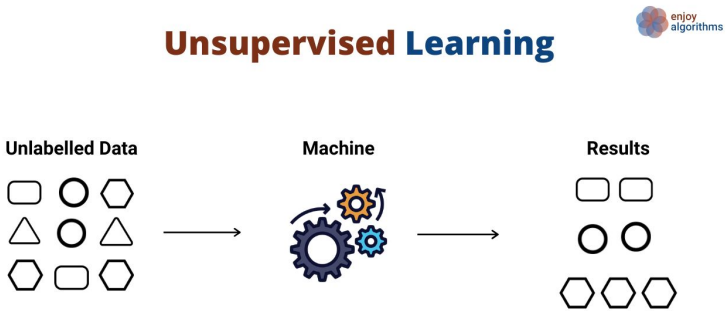

FIGURE 1.3 – Apprentissage non supervisé
[10]

1.3.1.3 Apprentissage semi supervisé (Semi Supervise Learning)

L'apprentissage semi-supervisé combine le meilleur des deux approches de l'apprentissage, supervisé et non supervisé. C'est utile dans les domaines de l'apprentissage automatique et de l'extraction de données, surtout quand il y a beaucoup de données non étiquetées et que l'obtention de données étiquetées est difficile. Contrairement aux méthodes traditionnelles qui se basent uniquement sur des données étiquetées, l'apprentissage semi-supervisé utilise à la fois des données étiquetées et non étiquetées pour améliorer les performances du modèle. Cela offre une solution flexible et efficace pour diverses tâches d'analyse de données [5].

1.3.1.4 Apprentissage par renforcement (Reinforcement Learning)

L'apprentissage par renforcement est une méthode d'apprentissage où un programme informatique, appelé agent, apprend à prendre des décisions en interagissant avec son environnement. L'objectif principal de l'agent est de maximiser les récompenses qu'il reçoit à travers ses actions. Pour y parvenir, l'agent doit trouver un équilibre entre explorer de nouvelles stratégies et exploiter celles qui lui ont déjà permis d'obtenir des récompenses [11]. Cette approche trouve des applications variées dans des domaines comme la robotique, les

recommandations de produits, et les jeux, où il est essentiel de prendre des décisions optimales pour obtenir des résultats positifs à long terme. comme le montre la classification des légumes dans la figure (1.4).

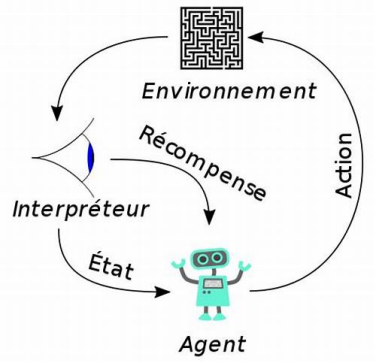

FIGURE 1.4 – Apprentissage par renforcement
[12]

1.3.2 Principales méthodes et algorithmes en apprentissage automatique

Dans le domaine scientifique, les experts utilisent divers algorithmes d'apprentissage automatique pour étudier de vastes quantités de données et en extraire des informations utiles. Bien que ces méthodes ne couvrent qu'une seule partie de l'intelligence artificielle, elles jouent un rôle essentiel dans ce domaine. Dans cette section, nous allons brièvement présenter les principaux modèles de classification couramment utilisés.

1.3.2.1 Machine à vecteurs de support (support vector machine)

Les machines à vecteurs de support ou Support Vector Machine (SVM) sont une technique spécifique d'apprentissage supervisé qui se distingue par sa capacité à maximiser la marge entre les classes. Cette approche, introduite par Vapnik, vise à minimiser les risques structurels [13]. Les SVM fonctionnent en convertissant un problème de classification en la recherche d'un hyperplan dans un espace caractéristique où les données sont séparées par une marge maximale. Cette marge représente la distance entre les classes et elle est optimisée pour garantir la meilleure séparation possible. Les SVM sont également connues sous le nom de séparateurs à marge large pour cette raison. Pour rendre les données linéairement

séparables, elles utilisent des noyaux, qui sont des fonctions mathématiques permettant de projeter et de séparer les données dans un espace vectoriel. Les "vecteurs de support" sont les points les plus proches de la frontière de décision. L'objectif est de trouver la frontière la plus éloignée de tous les points dentraînement, ce qui permet une généralisation optimale du modèle. Comme le montre la figure (1.5).

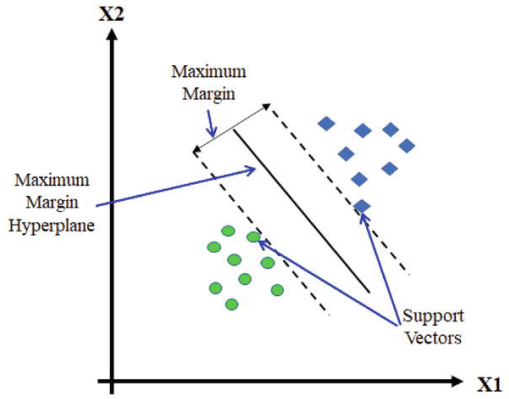

FIGURE 1.5 – Machine à vecteurs de support
[14]

1.3.2.2 K plus proches voisins

L'algorithme K-Nearest Neighbors (KNN) est l'un des concepts les plus simples en apprentissage automatique. Son principe est de classifier une nouvelle observation en se basant sur les votes des voisins les plus proches de cette observation. En d'autres termes, la classe de la nouvelle observation est déterminée en fonction de la classe majoritaire parmi ses k voisins les plus proches. Cette méthode est dite non paramétrique car elle ne nécessite pas de faire des hypothèses sur la relation entre les variables. KNN se base sur le principe du voisinage pour réaliser la classification [15].

1.3.2.3 Arbres de décision (decision trees)

Un arbre de décision est comme une feuille de route pour prendre des décisions. Il est construit automatiquement à partir des données disponibles, en utilisant des caractéristiques pour diviser progressivement ces données en groupes plus précis. Ces arbres, qui sont utilisés pour classer les données, sont organisés comme des arbres avec des noeud, des branches et des feuilles. Chaque noeud de l'arbre représente une règle de classification, déterminée à partir

des données. Pour comprendre comment un arbre de décision prend une décision, il suffit de suivre les branches depuis la racine jusqu'à la feuille correspondant à la réponse. Les arbres de décision peuvent fournir des réponses simples (vrai/faux) Comme le montre la figure (1.6) pour la classification ou des réponses numériques pour la régression. Ils sont appréciés pour leur rapidité et leur facilité d'utilisation, car ils offrent une représentation visuelle complète du processus de décision. Cela est particulièrement utile lorsqu'il est nécessaire d'expliquer comment une conclusion a été obtenue à un public intéressé [16].

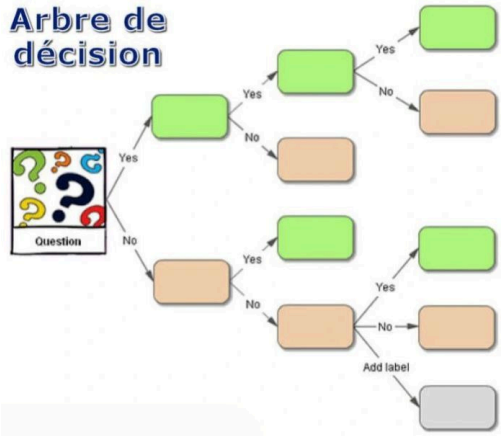

FIGURE 1.6 – Exemple dillustration dun arbre de décision [17]

1.3.2.4 Modèles de markov cachés

Les modèles de markov cachés ou Hidden Markov Model (HMM) sont essentiels pour la reconnaissance de texte, permettant d'identifier les caractères, les mots et les lignes en utilisant des connaissances préalables comme les modèles de langage et les lexiques. Initialement conçus pour la reconnaissance vocale, leur succès les a rendus utiles pour la reconnaissance de l'écriture manuscrite. Leur simplicité, robustesse et adaptabilité ont conduit à leur adoption dans divers domaines tels que la biologie (séquençage ADN), le traitement du langage naturel, la robotique, la climatologie, la cardiologie, l'économie et l'extraction d'informations textuelles.

1.3.2.5 Les réseaux de neurones artificiels (ANN)

Les réseaux de neurones artificiels ou sont des structures informatiques inspirées du fonctionnement du cerveau humain. Ils se composent dune série de neurones disposés en couches successives : une couche dentrée (input layer), une ou plusieurs couches cachées (hidden layers) et une couche de sortie (output layer). Le comportement global de ces neurones est déterminé par les connexions établies entre eux et par les paramètres qui régissent l'architecture du réseau. Chaque connexion entre les neurones des différentes couches est associée à un poids $\boldsymbol{w_{ij}}$ pour la connexion entre la couche dentrée et la couche cachée, et $\boldsymbol{w_{ki}}$ pour celle entre la couche cachée et la couche de sortie, qui représente lintensité de la liaison entre un neurone dune couche et le neurone de la couche suivante. Un réseau neuronal typique comporte donc ces trois types de couches mentionnées. La complexité du système dépend non seulement du nombre de couches cachées mais aussi du nombre de neurones présents dans chacune de ces couches [18]. La figure (1.7) illustre une architecture typique de réseau de neurones artificiels (ANN) avec une seule couche cachée.

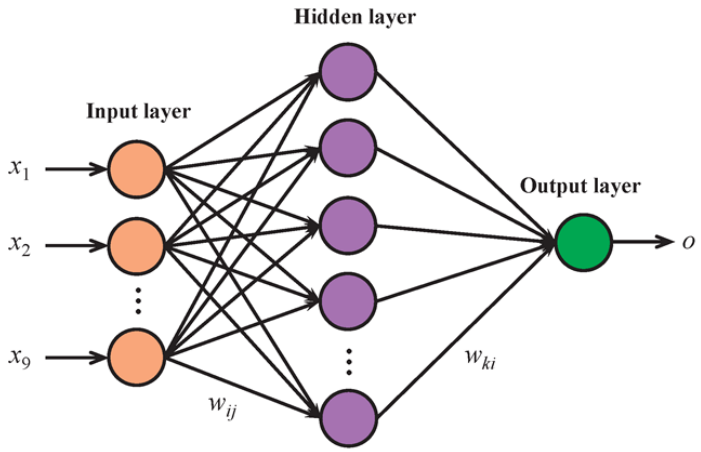

FIGURE 1.7 – La structure générale d'un réseau de neurones artificiels avec une seule couche cachée

[19]

Le perceptron multi-couches (MLP) : Le perceptron multicouche ou Multilayer Perceptron (MLP) est un modèle d'architecture de réseau de neurones artificiels très courant. Il se compose d'une couche d'entrée, de plusieurs couches cachées et d'une couche de sortie. Chaque couche est composée de plusieurs neurones, et chaque neurone est relié aux neurones

de la couche suivante par des poids qui déterminent leur influence réciproque. Le but d'un MLP est de produire une sortie spécifique à partir de certaines entrées. Pour y arriver, le réseau apprend en utilisant des exemples d'entrées et de sorties souhaitées. Il calcule ensuite l'erreur entre la réponse obtenue et la réponse attendue. Pendant l'apprentissage, les poids des connexions sont ajustés pour réduire cette erreur. Le MLP est efficace pour résoudre des problèmes complexes, faire des classifications et approcher des fonctions continues [20]

1.3.2.6 Les fonctions d'activation

Les fonctions d'activation sont essentielles pour les réseaux neuronaux artificiels car elles transforment les signaux d'entrée en sorties qui sont ensuite transmises aux couches suivantes du réseau. Elles commencent par calculer la somme pondérée des entrées en utilisant les poids correspondants, puis appliquent une fonction spécifique à chaque couche pour générer la sortie. Cette sortie devient alors l'entrée pour la couche suivante, ce qui est crucial pour assurer le bon fonctionnement des réseaux neuronaux à plusieurs niveaux. La régularité des poids permet un calcul efficace du gradient de la fonction de perte globale par rapport à ces poids, .facilitant ainsi un entraînement efficace du réseau. Pour garantir un apprentissage optimal avec un gradient approprié, il est recommandé d'utiliser des fonctions d'activation qui ne saturent pas trop rapidement [21].

Il y a plusieurs sortes de fonctions d'activation, voici quelques exemples (1.8) :

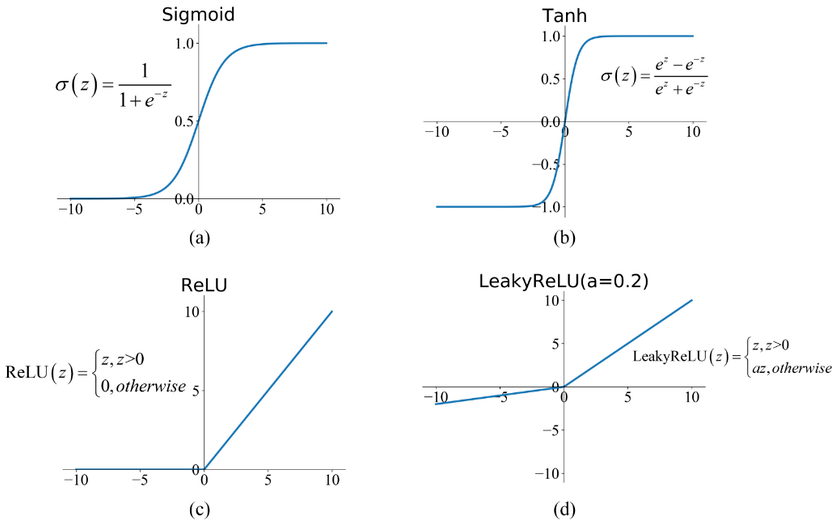

FIGURE 1.8 – Fonctions d'activation couramment utilisées [22]

— **La fonction logistique (sigmoïde) :** Cest l'une des fonctions les plus fréquemment utilisées, souvent appelée fonction logistique ou sigmoïde. Elle produit des valeurs comprises entre 0 et 1, ce qui peut être interprété comme la probabilité que le neurone soit activé de manière aléatoire [23]. Cela se définit comme :

$$\boldsymbol{\sigma(z) = \frac{1}{1 + e^{-z}}} \tag{1.1}$$

— **La fonction tangente hyperbolique (TanH) :** La fonction TanH, abréviation de tangente hyperbolique, est largement utilisée dans les réseaux de neurones en raison de ses propriétés distinctives. Elle génère des valeurs dans une plage comprise entre -1 et 1, définie par :

$$\boldsymbol{\sigma(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}} \tag{1.2}$$

Cette fonction constitue une alternative efficace à la sigmoïde, notamment pour capturer les relations non linéaires entre les données d'entrée et de sortie des réseaux neuronaux. De plus, lors de l'entraînement par rétropropagation du gradient, la dérivée de TanH est préférée car elle maintient les gradients dans une plage appropriée, facilitant ainsi la convergence des algorithmes d'optimisation [23]. La fonction TanH est définie par la formule :

$$\mathrm{TanH}(\boldsymbol{z}) = \boldsymbol{2 \cdot \mathrm{sigm}(2z) - 2} \tag{1.3}$$

— **La fonction ReLU (Rectified Linear Unit) :** Est considérée comme une représentation simplifiée du comportement des neurones dans le cerveau (homologue biologique). Elle fonctionne en supprimant toute activité lorsque l'entrée est négative (retourne 0),et en laissant passer l'entrée telle quelle si elle est positive. Cela est défini par la formule suivante :

$$\mathrm{ReLU}(\boldsymbol{z}) = \boldsymbol{\max(0, z)} \tag{1.4}$$

Cette approche directe et non linéaire est devenue très populaire dans les réseaux de neurones modernes, en particulier pour les applications de vision par ordinateur [24]. Elle permet une convergence rapide et efficace lors de l'apprentissage.

— **L'unité linéaire d'erreur gaussienne (GELU) :** La fonction Gaussian Error Linear Unit (GELU) est couramment employée dans les réseaux de neurones en raison de sa capacité à éviter le problème de disparition des gradients, ce qui la différencie des fonctions comme la sigmoïde et la tangente hyperbolique. Elle assure un gradient bien défini même dans les zones négatives, évitant ainsi le phénomène des neurones morts. Par ailleurs, elle permet d'obtenir des performances optimales dans les modèles de transformers utilisés pour le traitement du langage naturel et la vision par ordinateur [25]. Cela se définit comme suit :

$$\text{GELU}(\boldsymbol{z}) = \mathbf{0.5}\boldsymbol{z}\left(\mathbf{1} + \mathbf{tanh}\left[\sqrt{\mathbf{2}/\boldsymbol{\pi}}\left(\boldsymbol{z} + \mathbf{0.044715}\boldsymbol{z}^{\mathbf{3}}\right)\right]\right) \tag{1.5}$$

— **Softmax :** Softmax est une fonction essentielle dans les réseaux neuronaux traitant des problèmes de classification à plusieurs classes.Elle transforme un vecteur de K valeurs réelles en un nouveau vecteur où chaque valeur est une probabilité, garantissant que la somme totale de ces probabilités est égale à 1. Peu importe les valeurs initiales,positives, négatives, nulles ou supérieures à un, softmax les normalise dans une plage de 0 à 1, les rendant ainsi facilement interprétables comme des probabilités. Les entrées faibles ou négatives sont converties en probabilités réduites, tandis que les valeurs élevées se traduisent par des probabilités plus élevées, tout en maintenant toujours une échelle cohérente de 0 à 1 pour chaque classe [23]. Cela se définit comme :

$$\text{softmax}(\mathbf{z})_i = \frac{\boldsymbol{e}^{z_i}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{e}^{z_j}} \tag{1.6}$$

— **La fonction Leaky-ReLU :** Le Leaky Rectified Linear Unit (Leaky ReLU) est une version améliorée de la fonction ReLU qui vise à résoudre le problème de "Dying ReLU". Contrairement à ReLU qui met à zéro les valeurs négatives, Leaky ReLU permet à une petite fraction des valeurs négatives d'être transmises, assurant ainsi que l'unité ne reste pas bloquée à zéro pour toutes les entrées. Cela favorise un meilleur flux d'information et améliore la stabilité de l'apprentissage dans les réseaux de neurones. Cela se définit comme :

$$\text{LeakyReLU}_{\boldsymbol{a}}(\boldsymbol{z}) = \mathbf{max}(\boldsymbol{az}, \boldsymbol{z}) \tag{1.7}$$

1.3.2.7 Naïve bayes

C'est une méthode de classification qui s'appuie sur le théorème de Bayes et part du principe que les différentes caractéristiques prédictives sont indépendantes les unes des autres. Autrement dit, il considère que la présence d'une caractéristique dans une classe n'influence pas la présence d'une autre caractéristique. Cette technique est largement utilisée dans le domaine de la classification de textes, où elle regroupe et classe les données en se basant sur les probabilités conditionnelles d'occurrence [5]. Le classifieur Naïve Bayes est souvent utilisé pour la détection de spam et l'analyse de sentiment.

1.3.2.8 K-means

Dans l'apprentissage non supervisé, un algorithme basique est utilisé pour regrouper les données. Ce processus consiste à définir k centres, chacun représentant un groupe. Le positionnement de ces centres est essentiel, car il a un impact direct sur les résultats. Une bonne méthode consiste à maximiser la distance entre les centres. Ensuite, chaque point de données est associé au centre le plus proche, formant ainsi un regroupement initial. Pour affiner ce regroupement, les centres sont ajustés pour refléter le centre de gravité de leurs groupes respectifs, et cette étape est répétée jusqu'à ce que le processus converge vers une solution [5].

1.4 Apprentissage profond (Deep Learning)

L'apprentissage profond ou Deep Learning (DL), est une évolution de l'apprentissage automatique, se base sur l'utilisation de plusieurs couches cachées de réseaux neuronaux artificiels (ANNs) pour transformer les données et découvrir des informations de plus en plus complexes. En structurant ces transformations à différents niveaux, cette approche permet aux modèles de détecter automatiquement des caractéristiques cruciales pour diverses applications telles que la vision par ordinateur, le traitement du langage naturel et même la reconnaissance des émotions. Inspiré par le fonctionnement du cerveau humain, le deep learning permet aux réseaux neuronaux artificiels d'apprendre de manière autonome, sans avoir besoin d'une surveillance constante d'experts. Ces avancées permettent d'obtenir d'excellentes performances pour résoudre des problèmes complexes sur de vastes ensembles de

données [26].

Après avoir examiné la définition de l'apprentissage profond, il est crucial de souligner la différence fondamentale avec l'apprentissage automatique traditionnel, souvent désigné sous le terme d'apprentissage classique.

Lorsqu'on parle d'apprentissage automatique, il faut généralement fournir des instructions claires sur ce qu'il faut prédire et quelles données spécifiques utiliser pour ces prédictions. En revanche, avec l'apprentissage profond, les algorithmes peuvent apprendre par eux-mêmes à faire des prédictions précises en analysant les données à travers des réseaux de neurones artificiels. Cela élimine la nécessité d'extraire manuellement les caractéristiques des données.

En termes simples, le machine learning fonctionne bien pour des tâches simples avec moins de données, tandis que le deep learning se distingue dans des tâches complexes qui nécessitent de vastes ensembles de données, même si cela requiert d'avantage de puissance de calcul. Les modèles de machine learning sont souvent plus faciles à comprendre car ils reposent sur des caractéristiques prédéfinies, tandis que les modèles de deep learning apprennent à identifier leurs propres caractéristiques au fil de l'entraînement, ce qui peut les rendre plus complexes à interpréter [27].

1.4.1 Les réseaux de neurones récurrents (RNN)

Les réseaux de neurones récurrents ou Recurrent Neural Network (RNN) sont des outils avancés d'apprentissage automatique conçus spécifiquement pour analyser des séquences temporelles telles que des conversations, des vidéos et des données chronologiques. Leur architecture en boucle leur permet de garder en mémoire des informations précédentes, ce qui est crucial pour faire des prédictions futures [28] comme le montre la figure (1.9). Cependant, bien que très utiles pour analyser des données qui évoluent dans le temps, les RNN standard rencontrent des difficultés à saisir les relations à long terme [29] en raison des défis liés à la gestion des gradients. Pour surmonter ces obstacles, les réseaux de mémoire à long terme Long Short Term Memory (LSTM) ont été développés, offrant des structures plus sophistiquées qui améliorent la capacité des RNN à apprendre et à se souvenir de dépendances complexes sur de longues séquences temporelles [30].

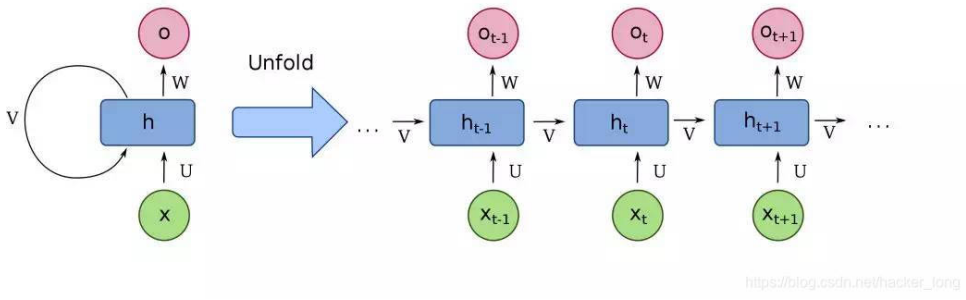

FIGURE 1.9 – Structure du RNN simple et RNN déplié
[31]

1.4.1.1 Architecture du RNN

L'architecture d'un réseau de neurones récurrents RNN comprend plusieurs couches clés qui permettent de traiter efficacement des données séquentielles.

— **Couche d'entrée :** La couche d'entrée est utilisée pour traiter des séquences de données où chaque élément contribue à un moment particulier. Contrairement aux réseaux de neurones convolutifs (CNN) qui sont spécialisés dans la reconnaissance d'images, les RNN sont adaptés pour capturer les dépendances temporelles et contextuelles dans les données.

— **Couche récurrente (Recurrent layer) :** Cette couche permet au RNN de conserver en mémoire les informations passées tout au long d'une séquence, ce qui est essentiel pour comprendre le contexte global et prendre des décisions basées sur l'historique complet des données d'entrée.

— **Couche de sortie :** La couche de sortie prend en compte les informations des étapes précédentes pour effectuer des prédictions ou des classifications. Selon la nature de la tâche, elle peut générer divers types de résultats : une seule valeur pour une classification spécifique ou un ensemble de données pour des applications telles que la traduction automatique.

1.4.1.2 Le long short term memory (LSTM)

Les LSTM sont une sorte de RNN spécialement conçue pour gérer les relations à long terme dans les données séquentielles [32]. Grâce à leurs connexions récurrentes, ils sont particulièrement bons pour classer, traiter et prédire des données dans le temps, car ils

peuvent prendre en compte les décalages entre des événements importants. Les LSTM ont été inventés pour résoudre les problèmes de gradient explosif et de gradient qui disparaît, ce qui les rend plus robustes face aux longues séquences de données.Ils sont constitués d'une cellule principale et de trois portes internes : la porte d'oubli (forget gate), la porte d'entrée (input gate) et la porte de sortie (output gate), qui contrôlent le flux d'informations. Grâce à leur capacité à gérer les dépendances à long terme [33], les LSTM sont souvent préférés aux RNN traditionnels [29] pour de nombreuses applications en apprentissage profond. La figure (1.10) présente un schéma d'une cellule LSTM.

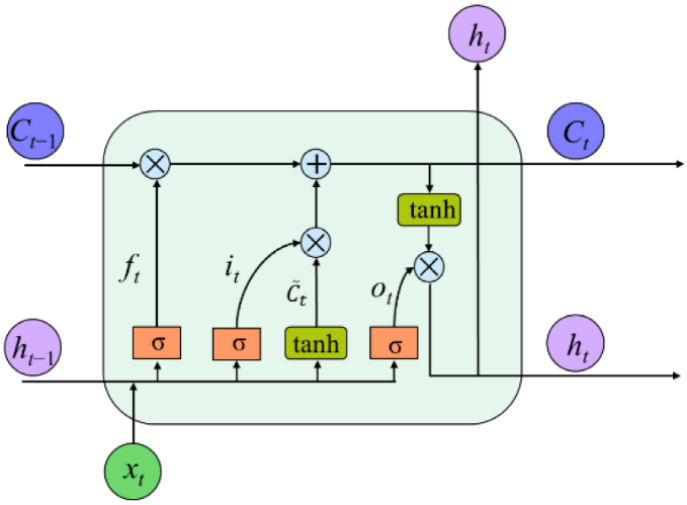

FIGURE 1.10 – Cellule Long Short-Term Memory (LSTM) [34]

La porte dentrée (i) détermine quelles valeurs seront mises à jour dans l'état, la porte doubli (f) décide quelles informations de l'état précédent seront rejetées, et la porte de sortie (o) sélectionne quelles informations seront transmises en sortie. Les LSTM peuvent examiner les séquences dans les deux directions pour saisir les relations temporelles à la fois antérieures et ultérieures. Ainsi que, leur processus d'apprentissage est plus long et leur parallélisation est complexe. De plus, pour les LSTM bidirectionnels, il est nécessaire d'avoir accès à l'intégralité de la séquence afin d'effectuer des prédictions, ce qui peut limiter leur utilisation en temps réel.

1.4.1.3 Les unités récurrentes à portes (gated recurent units)

Les unités récurrentes à portes ou Gated Recurent Units (GRU) sont une version simplifiée et plus rapide des réseaux neuronaux récurrents, offrant une alternative aux LSTM. Elles utilisent deux portes pour contrôler le flux d'informations : la porte de Mise à jour décide quelles informations sont importantes pour la prochaine étape, tandis que la porte de Réinitialisation détermine si les informations passées doivent être conservées ou oubliées. Cette approche optimise l'utilisation de la mémoire et facilite l'apprentissage, tout en résolvant les problèmes de gradient qui disparaît. Les GRU sont largement utilisées dans des applications comme la modélisation des signaux vocaux, la traduction automatique et la reconnaissance d'écriture [35].

1.4.2 Les réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs ou Convolutional Neural Network (CNN) sont reconnus comme des approches puissantes en apprentissage profond, particulièrement efficaces pour la vision par ordinateur. Typiquement, un CNN se compose de trois couches principales : convolution, pooling et entièrement connectée, chacune spécialisée dans l'extraction et la classification des informations visuelles [36]. Le processus d'apprentissage comprend deux phases : transformation de l'image pour prédire une sortie comparée aux étiquettes réelles, suivie de l'ajustement des paramètres via la rétropropagation pour minimiser la perte jusqu'à la convergence du réseau. Les CNN utilisent la convolution au lieu de la multiplication matricielle pour extraire des caractéristiques pertinentes des données corrélées localement, facilitant ainsi l'apprentissage des abstractions grâce à des fonctions d'activation non linéaires [37]. La Figure (1.11) présente une représentation détaillée étape par étape d'un réseau de neurones convolutionnel GRU conçu pour la classification d'images.

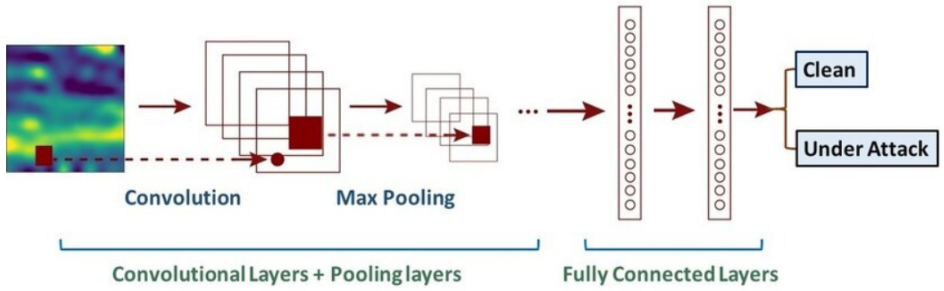

FIGURE 1.11 – Architecture typique des réseaux de neurones convolutifs CNN [38]

1.4.2.1 Architecture du CNN

Le CNN se compose de trois couches principales : convolution (CONV), pooling (POOL), et entièrement connectée (FC). Chaque couche joue un rôle spécifique dans le traitement et la classification des données visuelles.

— **La couche de convolution (Convolutional Layer) :** La couche convolutive est cruciale dans les CNN car elle réalise lextraction de caractéristiques à partir dimages en couleur.Ce processus utilise la convolution, où un produit scalaire est calculé entre les pixels de l'image et les filtres, générant ainsi des cartes de caractéristiques [39] comme démontré dans la Figure (1.12) .

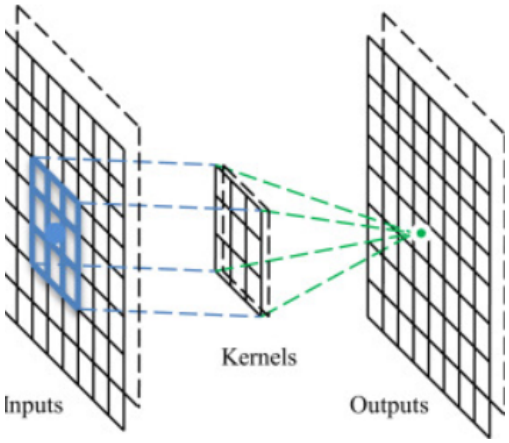

FIGURE 1.12 – L'opération de convolution [40]

Ces cartes sont ensuite traitées par des fonctions non linéaires telles que ReLU pour introduire de la non-linéarité dans le modèle. Les avantages de cette couche incluent le partage de poids et la connectivité locale, qui permettent d'apprendre les relations

spatiales entre pixels voisins et dassurer une certaine invariance à la position des objets dans l'image, tout en réduisant le nombre total de paramètres grâce au partage de poids au sein des cartes de caractéristiques.

— **La couche de pooling (Pooling layer) :** Après les couches de convolution, les couches de pooling diminuent progressivement la taille des cartes de caractéristiques, ce qui aide à optimiser le modèle. Elles ignorent les décalages et regroupent les informations des pixels voisins. Le max-pooling est souvent préféré au pooling moyen car il conserve mieux les informations essentielles tout en réduisant la taille des cartes de caractéristiques, améliorant ainsi la représentation des données. La figure (1.13) illustre l'utilisation d'un max-pooling avec des cartes de caractéristiques de 8 x 8 et des sorties de déserts économiques de 4 x 4, en appliquant un opérateur de 2 x 2 avec un décalage de 2 [39].

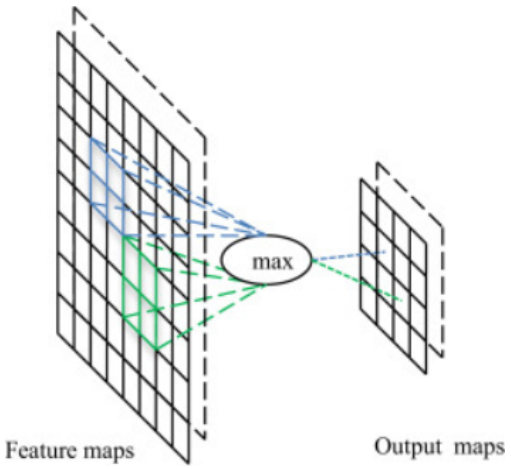

FIGURE 1.13 – L'opération de Max pooling [39]

— **La couche entièrement connectée (Fully connected layer) :** La dernière couche d'un réseau de convolution est cruciale car elle combine les caractéristiques distinguées des étapes précédentes pour effectuer une classification spécifique. Elle prend les informations extraites des données d'entrée et les transforme en un seul vecteur de caractéristiques en les aplatissant. Ensuite, les couches connectées complètes utilisent ces activations pour générer des scores de classe, garantissant une classification précise et distinctive [41].

1.4.2.2 Modèles CNN pré-entrainés

Les modèles CNN pré-entraînés tels que LeNet-5, AlexNet, VGGNet, GoogLeNet/Inception, ResNet, MobileNet et MobileNetV2 ont profondément transformé le domaine de la vision par ordinateur en introduisant des avancées majeures et en réalisant d'excellentes performances dans des compétitions de haut niveau comme l' Imagenet Large Scale Visual Recognition Challenge (ILSVRC) [42].

— **LeNet-5 :** Le LeNet-5, créé par Yann LeCun en 1998, est un modèle de base des réseaux neuronaux convolutifs (CNN).Composé de sept couches (deux de convolution, deux de pooling, trois entièrement connectées), il a joué un rôle essentiel dans la reconnaissance des chiffres manuscrits. Bien adapté aux images de 32x32 pixels sans nécessiter de prétraitement, il était limité pour des tâches plus complexes en raison du manque de données d'entraînement et des capacités de calcul de l'époque [43].

— **AlexNet :** AlexNet est reconnu pour sa profondeur avec huit couches au total : cinq de convolution et trois entièrement connectées. Ses avancées incluent l'introduction de ReLU pour un apprentissage plus rapide et l'utilisation du dropout pour éviter le sur-apprentissage. En ajustant le taux d'apprentissage, AlexNet a accéléré la convergence de l'entraînement. En augmentant le volume de données d'entraînement par la génération de variations aléatoires des images originales, AlexNet a renforcé la robustesse de son apprentissage [44].

— **ZFnet :** Cette nouvelle architecture représente une amélioration par rapport à AlexNet, en affinant les réglages des paramètres importants tels que la taille des couches convolutionnelles et l'échelle focale du noyau utilisé dans la première couche.

— **VGGNet :** VGGNet, conçue par des chercheurs de l'Université d'Oxford en 2014, a obtenu la deuxième place lors de compétitions majeures.Elle est connue pour ses 19 couches profondes, en utilisant des filtres de convolution plus petits (3 * 3) pour mieux extraire les détails des images, ce qui représente une amélioration par rapport aux méthodes antérieures avec des filtres plus grands [45].

— **GoogLeNet/Inception :** En 2014, GoogLeNet a révolutionné les réseaux neuronaux avec ses modules "Inception". Plutôt que de simplement empiler des couches de calcul, ces modules ont adopté une approche innovante en explorant différentes tailles et types

de convolutions simultanément. Cette méthode a permis d'approfondir le réseau tout en utilisant moins de paramètres [46], améliorant ainsi considérablement ses performances et son efficacité globale.

— **ResNet :** En 2015, Microsoft a développé l'architecture du Réseau Résiduel, qui a marqué une avancée majeure dans l'apprentissage profond. En introduisant des connexions résiduelles, cette architecture a permis aux réseaux de devenir extrêmement profonds, atteignant jusqu'à 152 couches, tout en résolvant le problème des gradients qui disparaissent. Grâce à cette innovation, le Réseau Résiduel Profond a surpassé la précision humaine dans le défi ImageNet [47].

— **MobileNet :** Les réseaux neuronaux convolutionnels profonds sont connus pour leur petite taille et leur capacité à surpasser d'autres modèles en termes de performance. Ils optimisent l'efficacité en utilisant des convolutions séparables en profondeur, traitant chaque canal de couleur séparément pour une gestion plus précise des informations [48].

— **MobileNetV2 :** MobileNetV2 améliore MobileNet en ajoutant des blocs inversés avec des caractéristiques de réduction, tout en réduisant considérablement le nombre de paramètres. Il peut traiter des images plus grandes que 32 x 32, ce qui améliore ses performances pour ces tailles d'image [49].

1.4.3 Les transformers

Malgré les avancées en apprentissage automatique et en deep learning, ces techniques ont montré des limites importantes telles que la difficulté à gérer les dépendances à longue distance dans les séquences complexes, la complexité temporelle accrue, ainsi que l'effet de disparition du gradient. Ces défis ont eu un impact significatif sur leur capacité à modéliser efficacement ces relations complexes. C'est à ce moment que les Transformers, introduits pour la première fois par Vaswani et ses collègues en 2017 dans leur article "Attention is All You Need", ont représenté une avancée majeure. Ce modèle séquence à séquence a bouleversé l'architecture des modèles d'apprentissage profond, notamment dans le domaine du traitement du langage naturel (NLP). Il repose sur l'utilisation de l'attention multi-têtes et d'une architecture sans récurrence, permettant un traitement parallèle plus rapide des données et

améliorant la compréhension des relations complexes entre les mots sur de longues distances. Contrairement aux modèles récurrents précédents, le Transformer améliore considérablement les performances dans des tâches telles que la traduction automatique et la génération de texte.

L'architecture encodeur-décodeur du Transformer (Figure (1.14)), comme décrite dans l'article de Vaswani [50], se divise en deux parties distinctes : l'encodeur à gauche et le décodeur à droite. Dans l'encodeur, les différentes couches alternent entre l'auto-attention (ou attention multi-têtes), qui capture les relations entre les mots dans les données d'entrée, et des réseaux neuronaux pour le traitement de ces informations. Chaque couche de l'encodeur utilise des connexions résiduelles pour prévenir les problèmes liés à la réduction du gradient, ainsi qu'une normalisation de couche pour stabiliser les valeurs. De manière similaire, le décodeur utilise une couche supplémentaire d'attention encodeur-décodeur (ou attention multi-têtes) pour se concentrer sur les parties cruciales de l'entrée encodée.

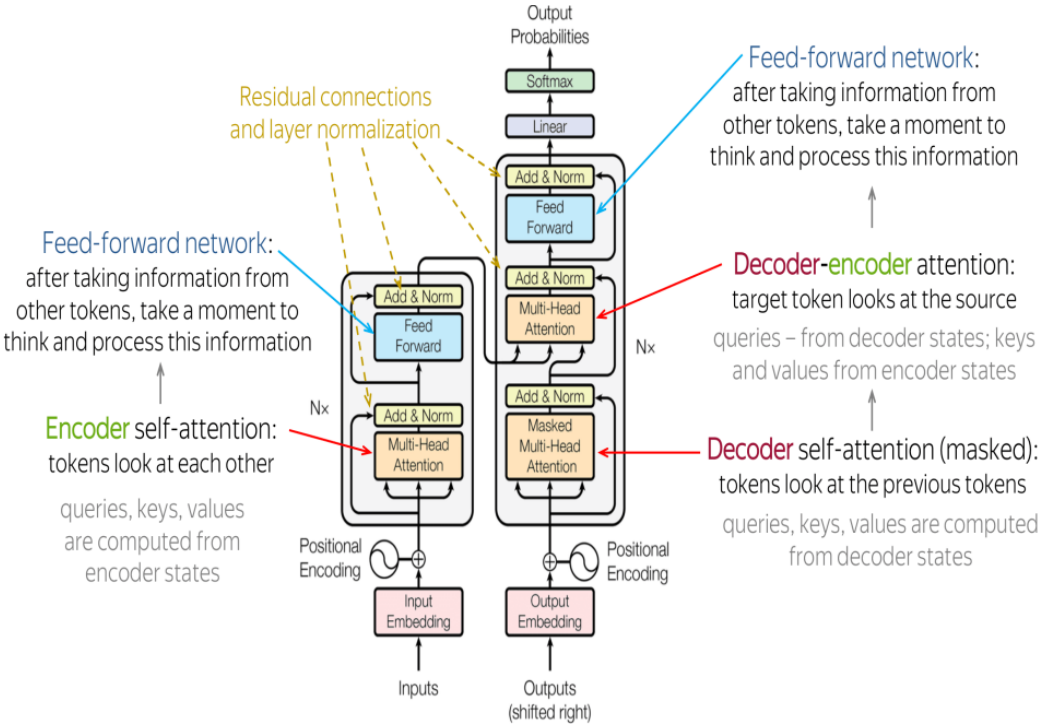

FIGURE 1.14 – L'architecture générale du Transformer [51]

Grâce aux Transformers, à l'apprentissage profond et à l'apprentissage automatique, l'intelligence artificielle a fait d'importants progrès, rendant possible des applications avancées comme les assistants virtuels avancés et les diagnostics médicaux assistés par ordinateur.

Après avoir introduit le modèle Transformer, nous pouvons aborder les étapes essentielles du modèle "vanilla" Transformer tel qu'il a été présenté par Vaswani et ses collaborateurs en 2017 dans leur article "Attention is All You Need" [50].

— **Input embedding (Incorporation d'entrée) :** Les tokens (jetons) d'entrée sont transformés en vecteurs grâce aux embeddings appris dans le modèle, dans la couche d'incorporation. Chaque vecteur a une taille fixe de 512 dimensions, ce qui signifie que chaque mot similaire est représenté par un vecteur qui lui ressemble.

— **Positional encoding (PE) :** Il existe deux formes d'encodage positionnel : statique et appris. Les deux visent à représenter la position des mots dans le texte d'origine sous forme de vecteur. L'encodage positionnel statique utilisé par le transformer utilise des ondes sinusoïdales et cosinusoïdales avec différentes longueurs d'onde, calculées selon les équations suivantes :

$$\mathrm{PE}(\boldsymbol{pos}, \boldsymbol{2i}) = \mathbf{sin}\left(\frac{\boldsymbol{pos}}{\mathbf{10000}^{\boldsymbol{2i}/\mathrm{d}}\ \mathrm{model}}\right) \tag{1.8}$$

$$\mathrm{PE}(\boldsymbol{pos}, \boldsymbol{2i+1}) = \mathbf{cos}\left(\frac{\boldsymbol{pos}}{\mathbf{10000}^{\boldsymbol{2i}/\mathrm{d}}\ \mathrm{model}}\right) \tag{1.9}$$

Les deux formules am fi over leaf Ou : i désigne la taille et pos représente la position dans la séquence temporelle. Le Transformer fusionne les encodages indiquant la position dans le texte avec les embeddings vectoriels, puis il transmet les résultats de cette combinaison à divers encodeurs, suivis de décodeurs.

— **Multi-head attention (Attention multi-têtes) :** L'attention multi-têtes est une composante essentielle des modèles Transformer, permettant au modèle de se focaliser simultanément sur différentes parties de la séquence d'entrée. Pour bien comprendre cette couche (voire la figure 1.15), il est d'abord nécessaire de saisir le mécanisme de l'attention par produit scalaire pondéré. Cette méthode implique une transformation linéaire des vecteurs de requête, clé et de valeur. Le score d'attention est obtenu en effectuant le produit scalaire entre la requête $(\boldsymbol{Q})$ et la transposée des vecteurs clé $(\mathbf{K}^{\intercal})$, puis en divisant le résultat par la racine carrée de la dimension du vecteur clé $(\boldsymbol{d_k})$ pour l'échelle. Une fonction softmax est ensuite utilisée pour normaliser ce

résultat en poids d'attention, qui servent à pondérer et sommer les valeurs afin d'obtenir la nouvelle représentation des embeddings. Les vecteurs clé, requête et valeur sont générés en multipliant les embeddings d'entrée par trois matrices de poids distinctes : $\mathbf{W}_k$, $\mathbf{W}_q$, $\mathbf{W}_v$. L'opération d'attention se formule ainsi par l'équation suivante :

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{\boldsymbol{d}_k}}\right)\boldsymbol{V} \tag{1.10}$$

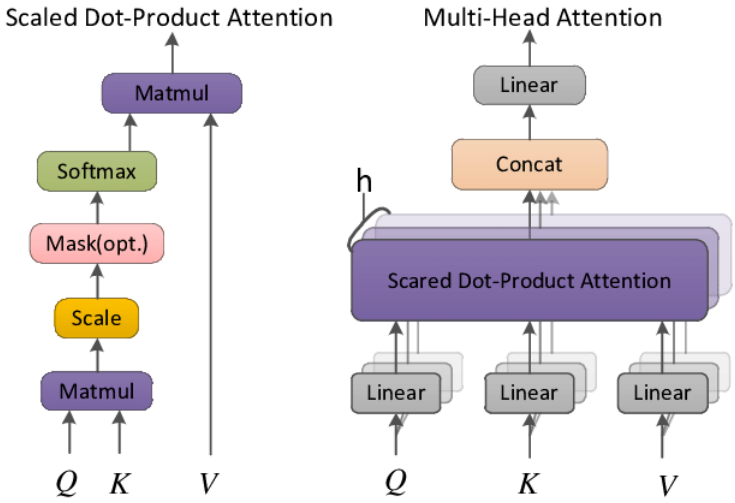

FIGURE 1.15 – Schéma montrant les différentes étapes de l'attention à produit scalaire et de l'attention multi-tête

[52]

La couche d'Attention Multi-Têtes utilise plusieurs mécanismes d'attention. Chaque mécanisme apprend à traiter les relations entre les mots d'entrée de manière différente et simultanée (voir la Figure 1.15). Voici comment est calculée la sortie de l'attention multi-têtes.

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_0, \ldots, \text{head}_h)\,\boldsymbol{W}\boldsymbol{O} \tag{1.11}$$

où

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}}, \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{1.12}$$

— **Réseau de propagation avant :** Un réseau de propagation avant ou Feed Forward Network (FFN) est un élément clé des modèles Transformer. Il est appliqué de

manière indépendante et identique à chaque position de la séquence d'entrée. Cette sous-couche comprend un réseau à deux couches avec une fonction d'activation ReLU. La première couche du réseau est quatre fois plus grande que la taille du modèle (dff égale à 2048), offrant ainsi au Transformer une capacité de représentation adéquate. La seconde couche réduit ensuite la sortie pour correspondre à la taille initiale du modèle (dmodel égale à 512). Le fonctionnement d'un FFN sur un vecteur d'entrée hi est exprimé par l'équation suivante :

$$\mathrm{FFN}(\boldsymbol{x}) = \mathrm{ReLU}(\boldsymbol{h_i W_1} + \boldsymbol{b_1})\boldsymbol{W_2} + \boldsymbol{b_2} \tag{1.13}$$

Cette équation montre comment le vecteur d'entrée hi est transformé d'abord par une couche linéaire, puis passe par une activation ReLU, avant d'être projeté de nouveau dans l'espace de dimension originale par une seconde couche linéaire.

— **Attention multi-tête masquées (Masked multi-head attention) :** Dans les modèles tels que le Transformer, l'attention multi-tête masquée joue un rôle crucial en permettant de gérer les informations de manière sélective tout en évitant l'accès à des valeurs inappropriées ou futures. Ce mécanisme implique l'application d'un masquage sur les sorties futures lors du processus de décodage, garantissant ainsi que chaque résultat dépend exclusivement des résultats antérieurs. L'équation correspondante pour l'attention multi-tête masquée est la suivante :

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\intercal + \boldsymbol{M}}{\sqrt{\boldsymbol{d_k}}}\right)\boldsymbol{V} \tag{1.14}$$

1.4.3.1 Transformers pour NLP

L'utilisation des transformers en traitement du langage naturel a conduit à l'émergence de modèles innovants tels que le Generative Pretrained Transformer (GPT), connu pour générer du texte, et le Bidirectional Encoder Representations from Transformers (BERT), qui crée des représentations bidirectionnelles des encodeurs.

— **Generative Pre-trained Transformer GPT :** GPT, Créé en 2018 par Radford et d'autres chercheurs [53] , est un outil avancé pour le traitement du langage naturel. Il est utilisé pour générer du texte, faire des résumés et traduire automatiquement. Ce

modèle, basé sur des décodeurs Transformer, se concentre sur le positionnement, l'attention multi-tête et les réseaux feedforward. GPT fonctionne en mode non supervisé, s'entraînant sur de vastes ensembles de données en ligne et utilisant des techniques telles que l'addition, la normalisation de couche et l'activation pour améliorer ses performances.

— **Bidirectional encoder representations from tansformers BERT :** un modèle de langage révolutionnaire basé sur l'architecture des réseaux Transformer, a été conçu par Devlin [54] chez Google AI. En analysant le texte de manière bidirectionnelle, BERT améliore significativement la compréhension contextuelle. Il est disponible en deux configurations BERTbase, avec 110 millions de paramètres répartis sur 12 couches et une dimension cachée de 768, et BERTlarge, avec 340 millions de paramètres répartis sur 24 couches et une dimension cachée de 1024. BERT est fondé sur deux phases cruciales d'entraînement la Modélisation de Langage Masquée (MLM), qui cache 15 hna lzm porcent des mots d'une phrase pour anticiper ces termes en fonction du contexte, et la Prédiction de la Phrase Suivante (NSP), qui juge la logique entre deux phrases successives. Ces procédés permettent à BERT de saisir les relations contextuelles complexes, simplifiant ainsi le transfert d'apprentissage pour de nombreuses applications avancées de Traitement Automatique du Langage Naturel (TALN) [55]. La structure globale du modèle BERT (Comparaison des architectures BERT Base et BERT Large) est présentée dans la figure (1.16).

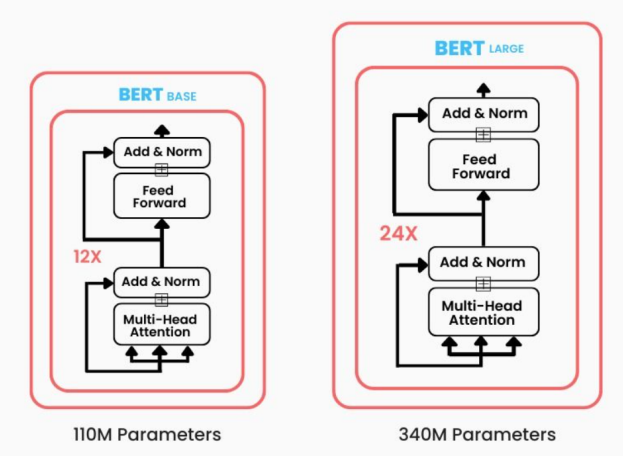

FIGURE 1.16 – Comparaison des architectures des modèles BERTbase et BERTlarge [56]

1.4.3.2 Transformers pour la vision Vision Transformer (ViT)

Les transformers, qui ont rencontré un succès énorme dans le domaine du traitement du langage naturel, sont maintenant utilisés dans la vision par ordinateur. Parmi les développements récents, on trouve Image GPT (iGPT) et les ViT.

— **Image GPT :** spécialisé dans la génération d'images [57], mais son utilisation est limitée par des exigences élevées en puissance de calcul et une qualité dimage inférieure [58].

— **Vision Transformer :** Le vision transformer est un nouveau type de réseau neuronal conçu pour la classification d'images. Il transforme une image en une série de vecteurs afin de saisir les relations à long terme entre ses différentes parties. Cette méthode novatrice a démontré d'excellentes performances en vision par ordinateur. Grâce à son importance croissante, le vision transformer est au coeur de notre prochaine étape de discussion.

1.5 Vision Transformers (ViT)

An Image is Worth 16*16 Words : Transformers for Image Recognition at Scale [59] a été réalisée par Neil Houlsby, Alexey Dosovitskiy et d'autres membres de l'équipe Google Brain ont présenté un nouveau modèle appelé ViT. Ce modèle fonctionne en découpant une image en morceaux de taille fixe, en les combinant de manière efficace, et en intégrant la notion de position dans le processus de codage du transformateur. La figure (1.17) fournit une illustration du fonctionnement de ce modèle, connu sous le nom de vision transformer.

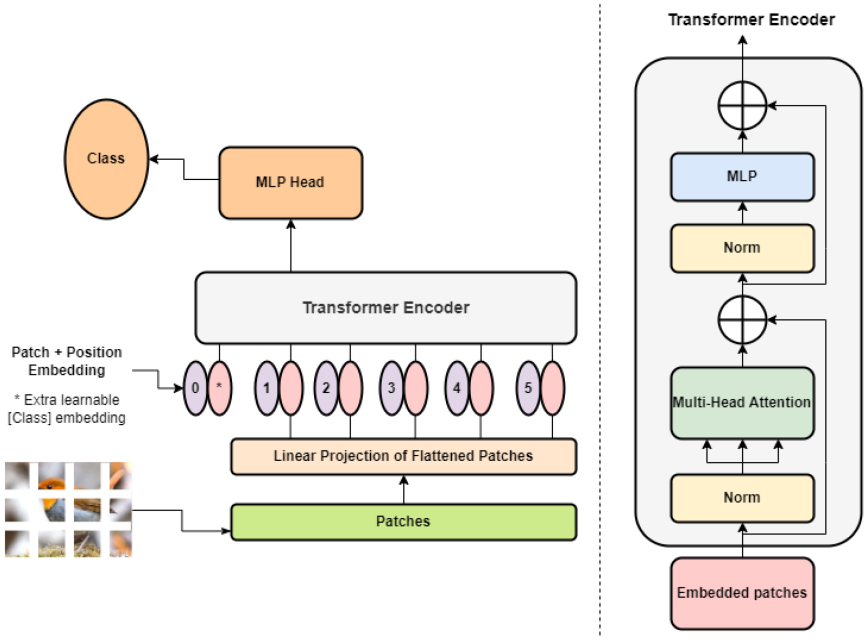

FIGURE 1.17 – Présentation du modèle Vision Transformer (ViT) [60]

1.5.1 Structure de vision transformers (ViT)

La structure de l'architecture Transformer s'adapte bien aux tâches de vision par ordinateur, ce qui a conduit à la création de l'architecture ViT. En pratique, cette méthode découpe les images d'entrée en patchs non chevauchants, qui sont ensuite intégrés comme des jetons. L'un des principaux avantages de l'utilisation de l'attention pour les tâches de vision par ordinateur est sa capacité à évaluer l'importance relative des différentes parties d'une image, offrant ainsi une approche globale plutôt que locale, contrairement aux convolutions traditionnelles. Cette méthode n'exige pas une structure encodeur-décodeur distincte [59]. Ils utilisent simplement la même structure pour l'encodeur, où les entrées sont d'abord intégrées, suivies d'une série de blocs comprenant un bloc d'auto-attention multi-têtes (MSA) et un bloc de perceptron multicouche MLP. Enfin, le "MLP Head" se trouve à la fin du dernier bloc.

— **Couche Multi head Self Attention (MSA) :** Cette couche combine linéairement toutes les sorties d'attention pour les ajuster aux dimensions appropriées. Les mécanismes d'attention multi-têtes facilitent l'apprentissage des dépendances locales et globales dans une image.

— **Couche de multi-layer perceptron MLP :** Cette couche se compose de deux couches qui utilisent l'unité linéaire d'erreur gaussienne (GELU).

— **MLP Head :** Un MLP utilise un jeton spécial appelé "jeton [classe]" ([class] token) en entrée, issu du dernier bloc d'attention, pour prédire la classe de l'entrée.

Tokenisation d'entrée et encodage de position (Input tokenization and positional encoding) : Traiter chaque pixel comme un jeton et calculer l'attention entre eux serait irréalisable, car cette opération demanderait trop de mémoire et de temps d'exécution en raison de sa complexité $\boldsymbol{O(n^2)}$. C'est pourquoi [58], les images sont découpées en patchs d'entrée non chevauchants. Ces patchs sont ensuite transformés en jetons pour réduire le nombre total d'entrées dans l'opération d'attention. Ces jetons sont incorporés en étant projetés linéairement en vecteurs de dimension Rd model. En pratique, cette projection est réalisée efficacement avec une couche de convolution dont le pas est égal à la taille du patch. Une autre différence notable dans l'architecture proposée par Dosovitskiy [58] par rapport au Transformer original de Vaswani [50] réside dans les codages de position ici, ils sont apprenables, contrairement à leur nature fixe et prédéterminée dans le modèle original.

Class token (Le jeton de classe) Après avoir créé les jetons d'entrée, Un vecteur est inséré au début de la séquence de jetons et est ensuite manipulé conjointement avec les autres jetons. Ce vecteur, désigné sous le nom de jeton [classe], est ensuite extrait des résultats du dernier bloc d'attention et utilisé pour prédire la classe de l'entrée. L'état initial du vecteur (celui ajouté avant le traitement) peut être modifié par l'apprentissage du modèle. Le jeton de classe est utilisé pour mettre l'accent sur les parties les plus importantes de l'image. Différents chercheurs ont proposé des variantes améliorées du Vision Transformer (ViT), dont le modèle SwinT . De plus, il existe les Transformers d'image à entraînement de données efficace (DeiT), une autre proposition de modèle amélioré [61]. SwinT utilise des fenêtres décalées pour améliorer la représentation des informations spatiales dans les images. De son côté, DeiT se concentre sur l'amélioration de l'entraînement des modèles ViT en utilisant des techniques de transfert d'apprentissage et de distillation de connaissances. Grâce à cette méthode, des performances satisfaisantes sont obtenues même avec peu de données d'entraînement.

1.5.2 Exemples d'applications des vision transformers

Les Vision Transformers (ViT) sont devenus une alternative compétitive aux réseaux de neurones convolutifs (CNN), qui sont actuellement en tête de la technologie en vision par ordinateur et largement utilisés pour différentes tâches de reconnaissance d'images. Avec la croissance continue des ensembles de données et le développement de méthodes non supervisées et semi-supervisées, il devient crucial de concevoir de nouvelles architectures de vision capables de s'entraîner plus efficacement sur ces ensembles de données. Le Vision Transformer (ViT) représente une première étape vers des architectures polyvalentes et évolutives qui peuvent résoudre une variété de tâches de vision. Par conséquent, le ViT devient de plus en plus important dans les domaines de la recherche en raison de sa polyvalence et de sa capacité d'adaptation [62]. Nous énumérons ici quelques-unes des applications les plus importantes de ViT :

— **Classification des images :** Les réseaux de neurones convolutionnels (CNN) sont considérés comme les plus avancés en classification d'images, car ils excellent dans la capture des détails locaux grâce à des champs récepteurs restreints. En revanche, les Transformateurs visionnaires (ViT) surpassent les CNN sur de grands ensembles de données en prenant en compte les informations à une échelle globale.

— **Description d'image :** Les Vits ont la capacité de générer des descriptions détaillées du contenu des images, ce qui améliore la qualité de leur catégorisation. Contrairement à l'utilisation de mots-clés simples, ils offrent une représentation globale des données, simplifiant ainsi la création de textes descriptifs pour chaque image.

— **Détection dune anomalie :** La détection et la localisation des anomalies dans les images, qui s'appuient sur les transformers, adoptent une méthode de reconstruction et de codage par patch. Les transformers maintiennent la disposition spatiale des patchs intégrés, ce qui simplifie l'identification des zones anormales à l'aide d'un réseau de mélange gaussien.

— **Segmentation des images :** Dense Prediction Transformers (DPT), développé par Intel, représente une avancée majeure dans le domaine de la segmentation d'images. En utilisant l'architecture des vision transformers, ce modèle propose une méthode sophistiquée et précise pour segmenter sémantiquement les images, avec un score im-

pressionnant de 49,02% en termes de mIoU pour la segmentation sémantique sur le jeu de données ADE20K. De plus, il améliore les performances relatives jusqu'à 28% par rapport à un réseau entièrement convolutionnel pour l'estimation de profondeur monoculaire.

1.6 Apprentissage par transfert (Transfer learning)

L'apprentissage par transfert, ou le Transfer Learning (TL), est une technique (approche) d' apprentissage automatique conçue pour surmonter les défis liés à l'entraînement de modèles avec des ensembles de données limités. Cette technique permet aux modèles de réutiliser les connaissances acquises en résolvant des tâches similaires sur de vastes ensembles de données,ce qui améliore leurs performances tout en nécessitant moins de données et de temps d'entraînement. Elle est largement employée dans des domaines tels que la vision par ordinateur, le traitement du langage naturel et le traitement audio, et montre son efficacité même avec un nombre limité de données étiquetées [63]. La figure (1.17) illustre le concept de l'apprentissage par transfert.

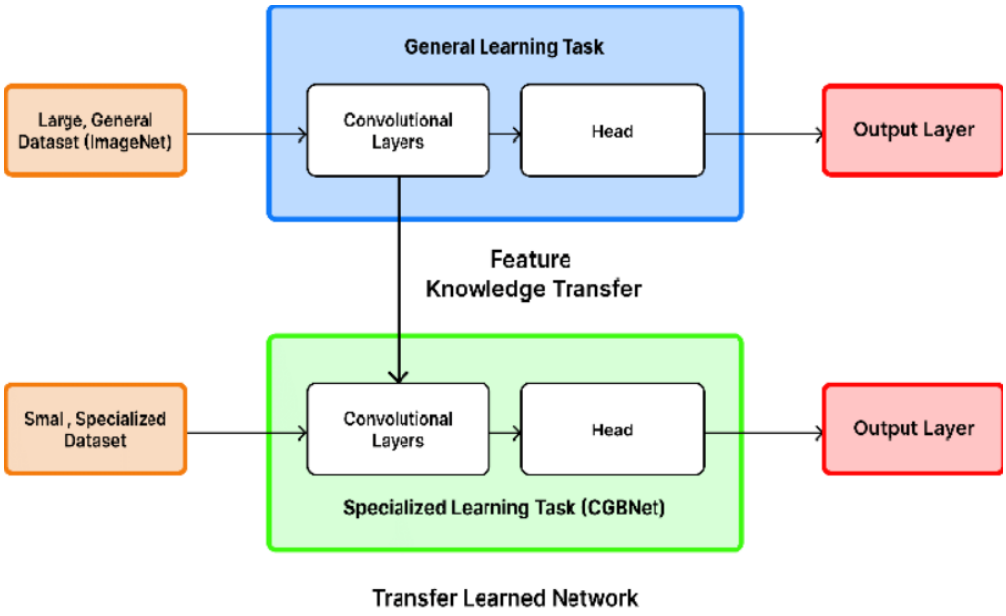

FIGURE 1.18 – Le concept de l'apprentissage par transfert. [64]

Après avoir expliqué le transfer learning, voyons maintenant son principe fondamental. L'apprentissage par transfert repose sur l'idée d'utiliser des modèles pré-entraînés sur

de vastes ensembles de données pour améliorer la performance sur des tâches spécifiques nécessitant moins de données. Cette méthode comprend deux étapes principales : le préentraînement et le fine-tuning.

1.6.1 Pré-entrainement (pretraining)

Il consiste à former un modèle sur une vaste base de données généraliste, comme ImageNet pour la vision par ordinateur. Au cours de cette phase, le modèle apprend à reconnaître les caractéristiques et les structures essentielles des données, telles que les contours, les textures et les formes de base.

1.6.2 Fine-tuning

Le modèle pré-entraîné est affiné (ré-entrainé) sur un ensemble de données plus petit et spécifique à la tâche ciblée. Cette étape permet d'ajuster le modèle pour qu'il capte les particularités de la nouvelle tâche tout en conservant les connaissances générales acquises lors du préentraînement. Ce processus optimise les performances du modèle en combinant les connaissances générales avec des ajustements spécifiques.

Après avoir exploré le principe du transfert d'apprentissage, nous allons maintenant aborder ses avantages :

— Optimisation de l'utilisation des données : Cette approche permet de capitaliser sur les connaissances acquises lors de l'apprentissage d'une tâche afin d'améliorer les performances sur des tâches similaires, même lorsque les ensembles de données disponibles sont limités pour la nouvelle tâche.

— Amélioration des performances : Transférer des connaissances d'une tâche à une autre permet souvent d'obtenir des performances nettement supérieures par rapport à des modèles formés sans recourir au transfert d'apprentissage.

— Adaptabilité à diverses tâches : Le transfert d'apprentissage peut être appliqué à une multitude de domaines tels que la vision par ordinateur, le traitement du langage naturel et la reconnaissance vocale, ce qui en fait une approche extrêmement polyvalente et adaptable.

— Réduction du surapprentissage : En transférant des connaissances d'une tâche à une autre, il est souvent possible de limiter le risque de surapprentissage, ce qui est particulièrement avantageux lorsque les ensembles de données d'entraînement sont restreints.

1.7 Conclusion

Dans ce chapitre, nous avons exploré l'évolution remarquable du domaine de l'intelligence artificielle en intégrant des techniques avancées telles que l'apprentissage automatique, le deep learning et les Transformers. Nous avons présenté un aperçu de ces techniques ainsi que de leurs différents types, puis nous avons examiné en détail leur impact sur le développement de l'intelligence artificielle. Nous avons commencé par l'apprentissage automatique, où les algorithmes sont entraînés à partir de données pour accomplir des tâches spécifiques. Ensuite, nous avons abordé l'émergence du deep learning en introduisant des architectures de réseaux de neurones profonds telles que les RNN et les CNN. C'est à ce stade que nous avons exploré les Transformers, qui ont introduit une nouvelle approche grâce à leur mécanisme d'attention. Enfin, nous avons discuté de l'utilisation des Vision Transformers (ViT) comme méthode de classification pour notre système.

Dans le chapitre suivant, nous explorerons à travers une revue de la littérature comment les techniques de l'intelligence artificielle peuvent être utilisées pour reconnaître les émotions à partir des expressions faciales.

CHAPITRE 2

LA RECONNAISSANCE DES ÉMOTIONS À PARTIR DES EXPRESSIONS FACIALES : ETAT DE L'ART

2.1 Introduction

Aujourdhui, la reconnaissance des émotions attire l'attention en tant que domaine d'étude intéressant, centré sur la compréhension fine des émotions humaines, que ce soit par les expressions du visage ou les variations de voix. Cette discipline vise à explorer profondément la nature de nos sentiments, distinguant ainsi notre compréhension des interactions sociales. Identifier avec précision les émotions est crucial pour faciliter une communication authentique entre les individus. En observant attentivement les visages, nous pouvons percevoir les nuances de la joie, de la tristesse, de la colère ou de la peur, ce qui peut éclairer nos choix et nos interactions.

Les avancées technologiques, comme les systèmes de reconnaissance faciale avancés, enri-

chissent cette exploration des émotions humaines, ouvrant la voie à des applications pratiques comme la détection des mensonges et l'analyse des sentiments sur les réseaux sociaux. La reconnaissance des émotions à travers les expressions faciales offre une perspective fascinante sur la nature humaine, favorisant des interactions sociales plus authentiques et gratifiantes.

Dans ce chapitre, nous présentons une exploration théorique de la reconnaissance des émotions, couvrant différents concepts, de la définition de l'émotion à la manière dont elles sont reconnues à partir de diverses sources d'information.

2.2 Fondement théorique

Dans cette section, nous allons discuter de deux sujets principaux : l'émotion en philosophie et en psychologie, et la modélisation des émotions. Ces domaines nous aideront à mieux comprendre comment les émotions fonctionnent et comment elles peuvent être représentées.

2.2.1 Emotion

Une émotion est une réaction profonde qui fusionnant les réponses physiologiques et psychologiques, déclenchée par la perception consciente ou inconsciente d'un objet ou d'un événement. Elle est souvent liée à l'humour, au tempérament, à la personnalité, à la disposition et à la motivation. Les émotions jouent un rôle essentiel dans la communication humaine, la prise de décision, les interactions et les processus cognitifs. Elles sont basées sur des expériences subjectives que les individus expriment à travers une diversité de termes sémantiques tels que : la Jalousie, la Peur, la Joie, la Colère, la Tristesse [65].

2.2.1.1 L'émotion en philosophie

En philosophie, une émotion peut être décrite comme une manifestation de notre vie émotionnelle, souvent accompagnée d'une sensation de bien-être ou de malaise. Elle représente une perturbation à court terme et un déséquilibre momentané. Parfois, elle peut être intense et se traduire par une augmentation de l'activité (enthousiasme, colère), tandis que d'autres fois, elle peut se manifester par une inhibition de l'action (peur ou "coup de foudre" dans l'amour) [66].

2.2.1.2 L'émotion en psychologie

En psychologie, Piéron caractérise l'émotion comme une réponse émotionnelle de force moyenne, influencée par les centres cérébraux associés au diencéphale, et qui se traduit souvent par des symptômes végétatifs. Bien que l'émotion soit perçue consciemment, son intensité semble fluctuer [67].

2.2.2 Modélisation des émotions

Ici, nous allons voir deux façons de modéliser les émotions : la modélisation dimensionnelle (continue) et la modélisation catégorielle (discrète). Nous expliquerons comment chaque méthode décrit et montre les émotions.

2.2.2.1 Modélisation dimensionnelle (Continue)

L'approche dimensionnelle suggère de représenter les émotions dans un espace multi-dimensionnel en considérant qu'elles sont engendrées par un ensemble défini de concepts, comme illustré dans la figure (2.1) Ces dimensions varient en fonction des exigences du modèle, mais le modèle de Russell, avec ses dimensions de valence et d'activation, est le plus répandu. Les émotions positives telles que la joie et les émotions négatives comme la colère peuvent être différenciées par leur valence. L'activation quant à elle représente le niveau d'excitation corporelle. Le principal avantage de cette approche est de représenter les émotions sans recourir à des étiquettes, et il est possible d'associer certaines zones à des étiquettes émotionnelles [68].

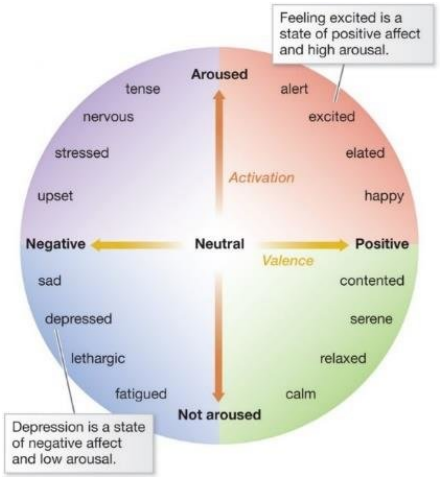

FIGURE 2.1 – Modélisation dimensionnelle
[69]

2.2.2.2 Modélisation catégorielle (Discrète)

Les approches discrètes considèrent les émotions comme des entités universelles et distinctes, chaque émotion étant représentée par une étiquette spécifique. Cette théorie repose sur l'idée que certaines émotions sont fondamentales, universelles, irréductibles et innées. Par exemple, des émotions de base comme la peur et la tristesse se retrouvent chez les individus de toutes nationalités et cultures, illustrant ainsi la nature universelle des émotions. Cependant, il existe un débat sur le nombre exact et la nature de ces émotions fondamentales. L'un des principaux avantages de cette approche est que, une fois les émotions clairement identifiées, elles deviennent plus faciles à analyser et à manipuler [68].

Les modèles et théories des émotions fournissent des cadres pour comprendre et catégoriser les émotions humaines. Ces modèles et théories varient en termes de leur niveau de détail et de complexité. Chaque théorie offre une perspective unique sur la manière dont les émotions sont générées, vécues et exprimées, et elles continuent de contribuer à l'exploration des expériences émotionnelles humaines. Dans ce qui suit, nous présentons quelques exemples parmi les nombreux modèles et théories des émotions que les chercheurs ont développés au fil des ans tels qu'ils ont été présentés dans [70].

— **Théorie des émotions de base :** Paul Ekman et son collègue Wallace V. Friesen [71] ont proposé la théorie des émotions de base, qui suggère qu'il existe un petit ensemble d'émotions universelles qui sont biologiquement préprogrammées et reconnaissables

à travers différentes cultures. Ils ont initialement identifié six émotions de base : la joie, la tristesse, la colère, la peur, le dégoût et la surprise. Plus tard, Ekman a élargi ses idées sur les émotions de base et leur universalité dans diverses publications [72] [73] [74] [75]. La théorie dEkman met en avant que ces émotions sont associées à des expressions faciales distinctes qui sont innées et interculturelles. Il a mené des recherches approfondies, notamment en étudiant. Son travail a été influent dans des domaines tels que la psychologie, l'anthropologie et les neurosciences, ainsi que dans des applications telles que la détection du mensonge et la technologie de reconnaissance des émotions. La figure (2.2) fournit une illustration de la théorie de paul Ekman, connu sous le nom des émotions de base.

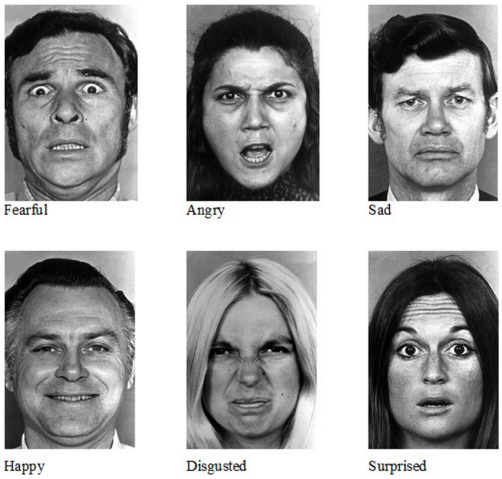

FIGURE 2.2 – La théorie des émotions de base
[76]

— **Le modèle circumplex des émotions :** James A.[77] a développé le modèle circumplex des émotions, qui se concentre sur la manière dont les émotions sont liées les unes aux autres et peuvent être cartographiées dans un cadre circulaire. Le modèle catégorise les émotions en fonction de deux dimensions : la valence (à quel point une émotion est positive ou négative) et l'activation (à quel point une émotion est intense ou calme). Le modèle circumplexe dispose les émotions le long de ces dimensions, créant un agencement circulaire où les émotions avec des niveaux de valence et d'activation similaires sont adjacentes les unes aux autres. Par exemple, les émotions comme la joie et l'excitation sont élevées à la fois en valence et en activation et sont positionnées près l'une de l'autre sur le cercle. Ce modèle offre un moyen de visualiser et de comprendre

les relations entre différentes émotions, et il a été utilisé pour explorer les expériences émotionnelles dans divers contextes, tels que les troubles de l'humeur, les relations interpersonnelles et le comportement des consommateurs. La figure (2.3) fournit une illustration de la théorie de James A. Russell , connu sous le nom de modèle circumplex des émotions.

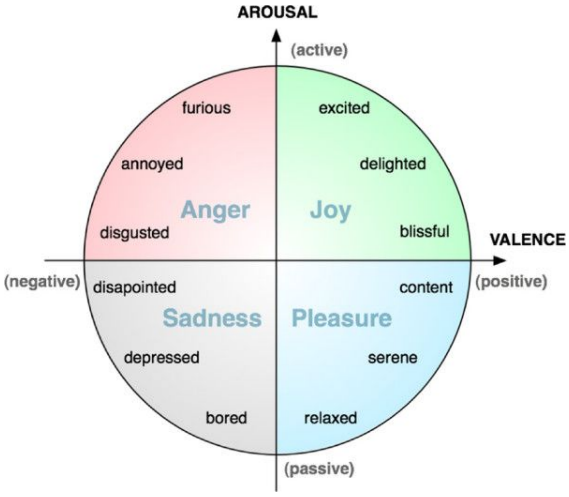

FIGURE 2.3 – Le modèle circumplex des émotions
[78]

— **La roue des émotions :** Robert [79] a proposé la "Roue des émotions". Il s'agit d'un autre modèle qui organise les émotions selon un cadre circulaire. Plutchik a proposé huit émotions primaires : la joie, la confiance, la peur, la surprise, la tristesse, le dégoût, la colère et l'anticipation. Il a également suggéré que ces émotions primaires peuvent être combinées pour former des émotions plus complexes. La roue représente les relations entre ces émotions, avec des émotions opposées situées en face les unes des autres (par exemple, la joie et la tristesse). La figure (2.4) fournit une illustration de la théorie de Robert Plutchik , connu sous le nom de La roue des émotions.

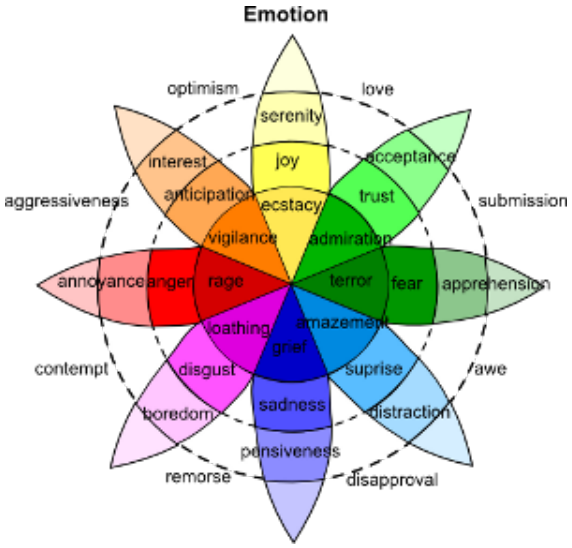

FIGURE 2.4 – La roue des émotions
[80]

— **Le modèle PAD (Plaisir-Activation-Dominance) :** PAD [81] est un modèle psychologique qui décrit et catégorise les expériences émotionnelles selon trois dimensions principales : plaisir, activation et dominance. Ce modèle a été développé comme une méthode pour offrir une approche complète et structurée pour comprendre et analyser les émotions. Chacune de ces dimensions capture un aspect différent des expériences émotionnelles. Ce modèle est souvent utilisé dans des domaines tels que la psychologie, la recherche sur le comportement des consommateurs et l'interaction humain-machine pour étudier les réponses émotionnelles à divers stimuli, produits et expériences. La figure (2.5) fournit une illustration de modèle de Mehrabian et Russell , connu sous le nom de modèle PAD (Plaisir-Activation-Dominance).

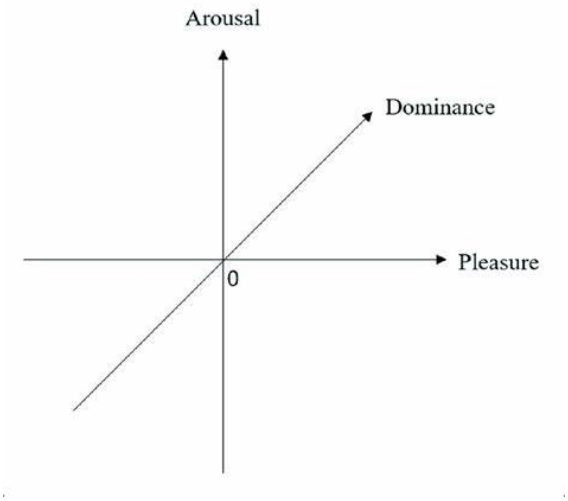

FIGURE 2.5 – Le modèle PAD (Plaisir-Activation-Dominance) [82]

- **Plaisir :** Cette dimension représente à quel point une émotion est positive ou négative. Les émotions positives, telles que le bonheur et la joie, sont associées à des sentiments de plaisir, tandis que les émotions négatives, telles que la tristesse et la colère, sont associées à des sentiments de mécontentement. La dimension du Plaisir varie de plaisant à déplaisant.
- **Activation :** Cette dimension mesure le niveau d'activation physiologique et psychologique causé par une émotion. Les émotions peuvent varier de faible activation (calme) à haute activation (excitation intense ou anxiété). Par exemple, le calme et la relaxation sont des états de faible activation, tandis que la peur et l'excitation sont des états de haute activation.
- **Dominance :** La dimension de Dominance reflète la perception du contrôle ou de la dynamique de pouvoir dans une expérience émotionnelle donnée. Les émotions peuvent être classées comme soumises ou dominantes. Les émotions soumises sont celles où les individus se sentent moins en contrôle, tandis que les émotions dominantes sont celles où les individus se sentent plus en contrôle. Par exemple, les sentiments d'impuissance peuvent être associés à des émotions soumises, tandis que la confiance peut être liée à des émotions dominantes.

— **La théorie de l'évaluation cognitive :** Richard [83] a proposé la théorie de l'évaluation cognitive, qui met l'accent sur le rôle des évaluations cognitives dans l'expérience

émotionnelle. Selon cette théorie, les émotions découlent de l'évaluation qu'un individu fait d'une situation, en tenant compte de facteurs tels que la pertinence, les objectifs personnels et les ressources d'adaptation. Lazarus a souligné que ces évaluations cognitives déterminent si une situation est perçue comme positive ou négative, conduisant ainsi à la réponse émotionnelle correspondante. La figure (2.6) fournit une illustration de la théorie de Richard Lazarus , connu sous le nom de l'évaluation cognitive.

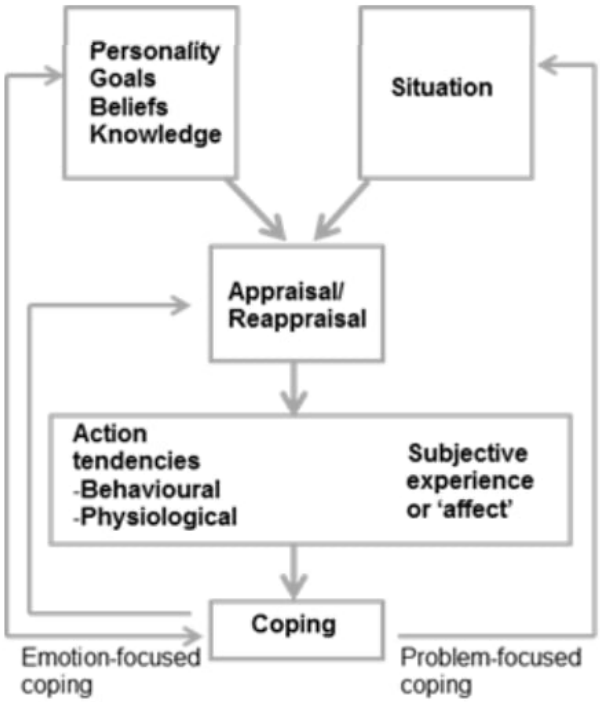

FIGURE 2.6 – La théorie de l'évaluation cognitive
[84]

— **Le modèle des émotions d'Arnold :** Magda [85] a proposé un modèle suggérant que les émotions résultent de l'interaction entre l'activation physiologique et l'interprétation cognitive. Elle a avancé que l'activation physiologique est un état général qui peut être interprété différemment en fonction de l'évaluation cognitive individuelle de la situation, conduisant à des émotions spécifiques.

— **Le modèle constructiviste des émotions :** Ce modèle suggère que les émotions sont construites sur la base de facteurs internes et externes. Les facteurs internes incluent les processus cognitifs individuels, tels que l'attention portée à certains aspects d'une situation, l'interprétation des événements et la formation des attentes. Les fac-

teurs externes incluent le contexte social et culturel, tels que les normes sociales, les croyances culturelles et les comportements expressifs des autres. Ce modèle suggère que les émotions ne sont pas des catégories fixes, mais plutôt des constructions flexibles qui sont façonnées par l'expérience individuelle et le contexte dans lequel elles se produisent [86] [87]. La figure (2.7) fournit une illustration du modèle constructiviste des émotions.

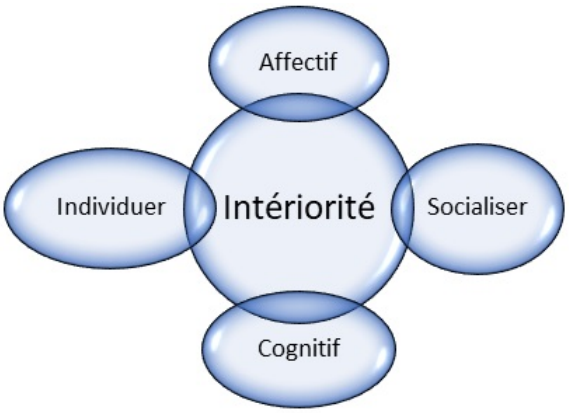

FIGURE 2.7 – Le modèle constructiviste des émotions
[88]

— **Le modèle des processus composant (Component Process Model, CPM) :** Ce modèle suggère que les émotions impliquent une combinaison de différents composants, tels que les changements physiologiques, les évaluations cognitives et les comportements expressifs. Les changements physiologiques peuvent inclure des variations du rythme cardiaque, de la respiration et de la conductance cutanée. Les évaluations cognitives se réfèrent à la perception et à l'interprétation individuelle de la situation, tandis que les comportements expressifs impliquent les expressions faciales, les vocalisations et le langage corporel. Ce modèle suggère que les émotions sont le produit de l'interaction entre ces différents composants, et que différentes émotions peuvent être distinguées par des différences dans la combinaison et l'intensité de ces composants [89]. La figure (2.8) fournit une illustration du modèle des processus composant (CPM).

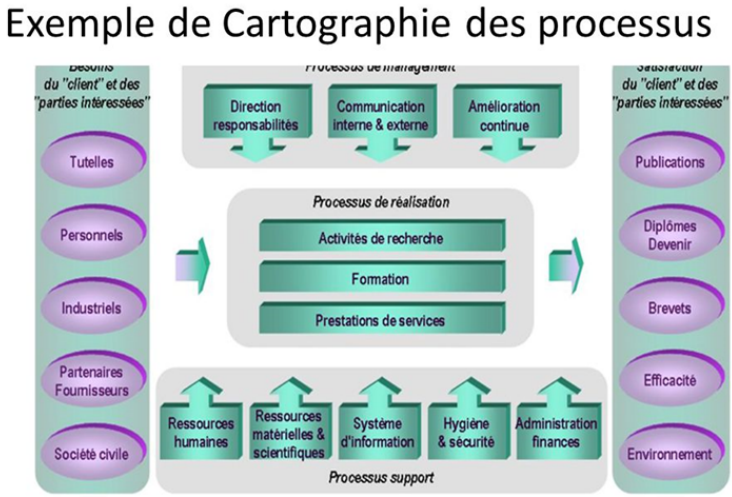

FIGURE 2.8 – Le modèle des processus composant CPM [90]

— **La théorie différentielle des émotions :** Carroll [91] suggère qu'il existe émotions de base présentes dès la naissance : la joie, l'intérêt-excitation, la surprise, la tristesse, la colère, le dégoût, le mépris, la peur, la honte et la culpabilité. Il a souligné l'importance d'étudier le développement émotionnel chez les enfants et le rôle de ces émotions de base dans l'expérience humaine. Contrairement à la recherche et à la théorie d'Ekman sur l'explication des variations universelles et culturelles dans les expressions faciales de l'émotion, Izard concentre son attention sur l'exploration des fonctions des émotions et leur rôle en tant que composante motivante du comportement humain. Les modèles émotionnels peuvent être classés en deux catégories : les modèles discrets/catégoriels comme le modèle d'Ekman et les modèles continus/dimensionnels comme Plaisir Activation dominance (PAD) et le modèle Circumplex. La représentation des émotions influence significativement la conception des systèmes de reconnaissance émotionnelle (MER). Les modèles catégoriels discrets qui classent les émotions en classes simples comme "heureux" et "triste" sont couramment utilisés pour leur facilité de mise en oeuvre. Cependant, ils manquent de nuance et de dynamique temporelle. Les modèles dimensionnels continus qui cartographient les émotions sur des points dans un espace de dimensions comme l'activation, la valence et la dominance peuvent mieux capturer la complexité et la fluidité de l'affect. Cependant, ils posent des défis de modélisation plus importants et nécessitent des algorithmes différents axés sur la régression plutôt

que sur la classification. Les modèles hybrides qui combinent des représentations catégorielles et dimensionnelles offrent un compromis. Enfin, le choix de la représentation des émotions doit être aligné sur les demandes spécifiques de l'application et la nature du problème de reconnaissance. Les systèmes catégoriels simples permettent des mises en oeuvre simples tandis que les modèles dimensionnels permettent une compréhension plus nuancée des états émotionnels et des changements au fil du temps. La figure (2.9) fournit une illustration de la théorie de Carroll Izard, connue sous le nom de théorie différentielle des émotions.

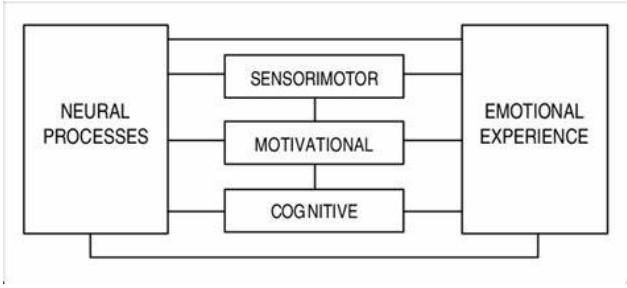

FIGURE 2.9 – La théorie différentielle des émotions
[92]

2.3 Principaux sources des données émotionnelles

Les émotions peuvent être détectées à partir de diverses sources, telles que les expressions faciales, les messages écrits, les tonalités vocales ou même les réponses physiologiques, comme illustré dans la figure (2.10) Cependant, chaque source a ses propres caractéristiques de capture des émotions. Dans les paragraphes suivants, nous décrirons ces différentes sources en détail.

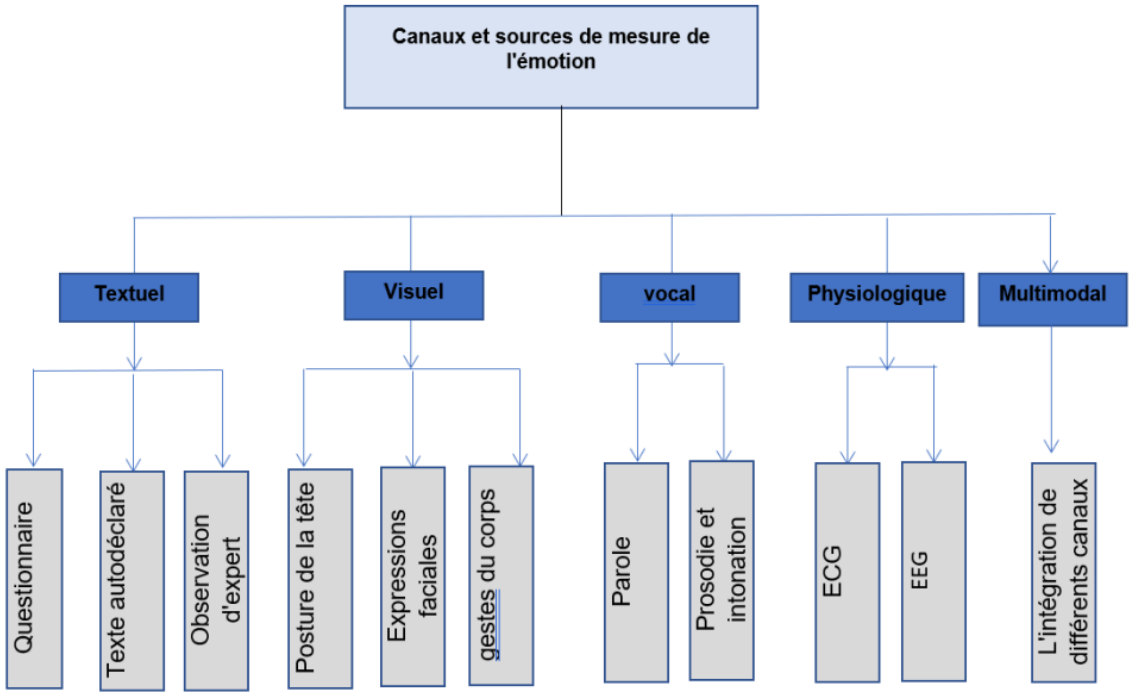

FIGURE 2.10 – Sources d'informations émotionnelles [93]

2.3.1 Source textuel

Une approche courante pour reconnaître les émotions à partir de textes consiste à utiliser des questionnaires où les participants répondent à une série de questions avec des évaluations émotionnelles. Les réponses peuvent être classées selon différents degrés d'émotion, et les participants peuvent utiliser des échelles numériques ou des réponses prédéfinies. Alternativement, ils peuvent fournir des réponses ouvertes pour évaluer divers aspects émotionnels [94]. Les émotions peuvent également être détectées à travers l'analyse du langage naturel, en examinant les mots, les phrases et la structure des phrases, bien que les émotions puissent parfois être implicites dans le contexte du texte, ce qui influe sur le sens des mots et des expressions [95]. Cette analyse du langage naturel peut également inclure la technique de plongement lexical. Plongement lexical : Cette technique consiste à représenter les mots sous forme de vecteurs numériques, avec des caractéristiques telles que densité et longueur constantes. Ces vecteurs peuvent être créés en se basant soit sur la prédiction, soit sur la fréquence des mots [96].

2.3.2 Source visuel

Les émotions sont souvent détectées en grande partie par les signaux visuels, car ces signaux sont directement liés aux interactions externes. La position de la tête, par exemple, peut véhiculer de nombreuses significations, et différents mouvements peuvent influencer l'expression faciale [97]. En outre, la communication humaine ne se limite pas aux mots, mais englobe également les mouvements corporels. Ces gestes non verbaux sont significatifs pour comprendre l'état émotionnel d'une personne. Par exemple, lorsque quelqu'un est joyeux, son corps adopte divers mouvements, notamment des gestes des mains et de la tête [98].

FIGURE 2.11 – Les expressions faciales
[99]

Une des parties les plus importantes de la communication est la capacité à lire les expressions faciales, représentées dans la figure (2.11) Ces expressions fournissent une mine d'informations sur les émotions et les signaux non verbaux lors d'interactions sociales. En effet, certaines expressions, comme les sourires ou les froncements de sourcils, sont déclenchées par nos émotions. Les recherches d'Ekman se concentrent désormais sur six expressions faciales émotionnelles primaires : La joie, la surprise, la peur, la tristesse, le dégoût et la colère [100]. En plus de la reconnaissance des émotions par les mouvements oculaires, d'autres sources, comme la dynamique de frappe et les mouvements de la souris, nécessitent des outils techniques pour être suivies. En ce qui concerne les caractéristiques, on peut mentionner les aspects géométriques et d'apparence.

Les émotions détectées à partir des signaux visuels, notamment les expressions faciales, fournissent une mine d'informations sur l'état émotionnel d'une personne. Cette capacité à lire les expressions faciales est enrichie par l'analyse des caractéristiques géométriques et d'apparence du visage.

2.3.2.1 Caractéristiques géométriques

Se réfèrent aux mesures des mouvements de traits spécifiques du visage, tels que les coins des lèvres ou des sourcils, comme illustré dans la figure (2.12) Pour créer un vecteur caractéristique qui représente la géométrie faciale, des composantes faciales ou des points de caractéristique faciale sont extraits. En surveillant le mouvement de ces points, il est possible de vérifier l'expression faciale sous-jacente. Selon cette approche géométrique, les émotions influencent la position relative et la taille de nombreuses caractéristiques faciales. L'analyse des caractéristiques géométriques implique souvent l'examen de la région de la taille, en particulier la localisation et le suivi des points clés dans cette région [101]. Différentes méthodes, telles que le modèle d'apparence active (AAM) et le modèle de forme active (ASM), peuvent être utilisées pour extraire ce type de caractéristiques.

2.3.2.2 Caractéristiques d'apparence

Se réfèrent aux aspects physiques et aux attributs visibles de l'apparence extérieure d'une personne. Lorsqu'une action est effectuée, ces caractéristiques, telles que les rides, les bosses, le front et les zones autour des lèvres et des yeux, modifient la texture du visage. Pour extraire un vecteur caractéristique à partir d'une image faciale, des filtres d'image sont utilisés et appliqués à l'ensemble du visage ou à des parties spécifiques [101]. Différentes méthodes, comme les Gabor wavelets et le Local Binary Pattern (LBP), peuvent être employées pour extraire ce type de caractéristiques. La figure (2.12) illustre les points dintérets du visage.

FIGURE 2.12 – Les points dintérets du visage
[102]

2.3.3 La source vocale

Le langage parlé est un moyen complexe de communication qui transmet une variété d'informations sur le locuteur, le message, la langue et même l'humeur. En effet, une grande partie de l'intention du locuteur est souvent communiquée par des signaux non verbaux. La manière dont les mots sont prononcés peut transmettre des informations non linguistiques tout aussi importantes que le texte lui-même. Ainsi, le même message peut être interprété de différentes manières en fonction de l'expression émotionnelle qui l'accompagne [103]. Pour analyser ces caractéristiques, des techniques telles que les coefficients cepstraux de fréquence Mel, les coefficients de prédiction linéaire et la prédiction linéaire perceptuelle sont utilisées. La manière dont les mots sont prononcés peut transmettre des informations non linguistiques tout aussi importantes que le texte lui-même, ce qui est crucial dans l'analyse des caractéristiques acoustiques pour la reconnaissance émotionnelle.

Les caractéristiques acoustiques : sont essentielles dans la reconnaissance émotionnelle car elles permettent d'extraire et d'analyser les propriétés de la parole telles que le son, le ton, le rythme et la fréquence vocale. Elles peuvent être divisées en deux domaines principaux : le domaine temporel, qui représente les changements de signal dans le temps, et le domaine de fréquence, qui se concentre sur le contenu en fréquence du signal sonore [104].

2.3.4 La source Physiologique

Les émotions ont leur origine dans le cerveau et se manifestent à travers des activités et des stimuli électriques, qui peuvent être enregistrés par des dispositifs tels que l'électroencéphalogramme (EEG), traduisant ainsi les réponses émotionnelles en signaux électriques [105]. Parallèlement, l'électrocardiographie (ECG) est une méthode classique pour surveiller l'activité électrique du cur en temps réel, et elle peut également être utilisée pour détecter les émotions en observant les variations du rythme cardiaque ou de la pression artérielle [106]. Ces signaux cérébraux et cardiaques sont capturés et enregistrés à l'aide d'outils techniques pour être ensuite interprétés sous forme de signaux électriques.

2.4 Processus de la reconnaissance des émotions

Le processus de la reconnaissance des émotions comprend plusieurs étapes, de la collecte des données à l'intégration des modèles dans des applications, visant à comprendre et interpréter les états émotionnels des individus à partir de leurs expressions faciales. Le module de la sélection des caractéristiques sélectionne la caractéristique la plus fondamentale. Le module de classification classe la caractéristique sélectionnée dans plusieurs classes d'objets.

2.4.1 Prétraitement des données

Le prétraitement des données est une étape cruciale dans le traitement des images, surtout pour des applications complexes comme la reconnaissance faciale. Ce processus transforme les données brutes en un format plus propre et structuré, optimisé pour les algorithmes d'apprentissage automatique et de deep learning. Il comprend plusieurs sous-étapes importantes qui améliorent la qualité des images et réduisent les variations inutiles, garantissant ainsi une meilleure performance des modèles.

2.4.1.1 Détection des visages

La détection des visages est une étape cruciale dans le prétraitement des données pour la reconnaissance des visages. Elle consiste à identifier et extraire la région du visage de chaque image. Divers algorithmes sont utilisés pour cette tâche, chacun offrant des avantages uniques en termes de précision et de performance.
Elle utilise des algorithmes de détection de visages pour localiser et extraire les régions faciales dans les images ou les vidéos tels que les cascades de Haar, le HOG+SVM et les CNN. Chacun de ces algorithmes offre ses propres avantages en termes de précision et de performance, permettant ainsi d'extraire efficacement la région du visage de chaque image pour une analyse ultérieure.

— **Haar cascades :** Créé par Paul Viola et Michael Jones, cet algorithme utilise des caractéristiques de Haar, des motifs simples capables de détecter les bords, les lignes et d'autres formes basiques. Il fonctionne grâce à une cascade de classificateurs, ce qui permet de filtrer rapidement les zones non pertinentes de l'image. Cet algorithme est rapide et efficace, bien qu'il puisse être sensible aux variations de lumière et de pose.

— **HOG et SVM :** Une méthode puissante pour détecter les visages est l'association de l'approche Histogram of Oriented Gradients (HOG) avec un Support Vector Machine (SVM). Le HOG décrit l'apparence et la forme des objets en analysant les gradients d'intensité dans les images et en créant des histogrammes basés sur ces gradients. Ensuite, ces caractéristiques sont utilisées avec un SVM, une technique d'apprentissage supervisé réputée pour sa capacité à résoudre des problèmes de classification efficacement.

— **CNN :** Les Réseaux de Neurones Convolutifs (CNN) sont la pointe de la technologie en termes de détection faciale. Ces architectures neurales spéciales peuvent extraire des caractéristiques complexes directement à partir des images en utilisant des couches de traitement spécifiques. Par exemple, des modèles comme MTCNN (Multi-task Cascaded Convolutional Networks) sont capables de repérer les visages avec une précision remarquable, même dans des situations difficiles où les angles de vue varient et l'éclairage change constamment.

2.4.1.2 Alignement et normalisation

Pour que notre modèle de reconnaissance faciale fonctionne correctement, il est crucial de bien aligner et normaliser les visages. Cela nous assure que les variations causées par des angles de vue différents ou des conditions d'éclairage variables n'affectent pas la capacité du modèle à reconnaître les visages avec précision.

— **Alignement des visages :** Dans cette étape, on utilise des repères clés sur le visage pour ajuster sa position et son orientation. En repérant des points comme les coins des yeux, le nez et la bouche, on peut appliquer des transformations géométriques pour aligner chaque visage selon une configuration standard de référence. Cela atténue les effets des différentes poses et assure que chaque visage est présenté au modèle de manière uniforme.

— **Normalisation des images :** Pour rendre les conditions d'éclairage et de contraste plus cohérentes, on applique différentes techniques de normalisation. Cela peut impliquer l'utilisation de l'égalisation d'histogramme, qui redistribue les niveaux de gris pour améliorer le contraste, ou des méthodes de normalisation d'intensité, qui ajustent les valeurs des pixels pour obtenir une intensité moyenne de zéro et une variance de

un. Ces approches aident à réduire les variations causées par l'éclairage et à présenter des images plus uniformes au modèle d'apprentissage.

— **Redimensionnement des images :** Les images doivent d'abord être redimensionnées uniformément à une taille standard, généralement 224x224 pixels. Ce redimensionnement est crucial car il garantit que toutes les images d'entrée ont des dimensions cohérentes, ce qui facilite le traitement par le modèle. La taille 224x224 est choisie en fonction des normes établies par des architectures de modèles populaires telles que ResNet et ViT, équilibrant la résolution des images et les contraintes computationnelles.

2.4.2 Extraction des caractéristiques faciales

L'extraction des caractéristiques faciales est une étape clé dans les systèmes de reconnaissance faciale. Elle peut être réalisée soit manuellement, avec des experts en vision par ordinateur identifiant les traits distinctifs des visages tels que les contours et les textures, soit automatiquement, à l'aide d'algorithmes d'apprentissage automatique comme les CNN et les RNN. Les modèles de deep learning sont capables d'apprendre des représentations significatives des visages à partir des données brutes, ce qui les rend adaptés à la reconnaissance faciale. Des exemples d'approches automatisées incluent l'utilisation de CNN comme MTCNN pour la détection et l'extraction de caractéristiques facials . Ces méthodes offrent une efficacité accrue mais nécessitent des ensembles de données étiquetées volumineux et une expertise en apprentissage automatique pour leur mise en uvre.

2.4.3 Entraînement du modèle

L'entraînement du modèle est une étape cruciale dans le processus de reconnaissance des émotions à partir des expressions faciales. Cette phase implique l'utilisation de données d' entraînement étiquetées pour ajuster les paramètres internes du modèle, lui permettant ainsi d'apprendre à reconnaître les patterns et les caractéristiques associés à différentes émotions. Selon le type de modèle choisi, qu'il s'agisse de méthodes basées sur des règles, d'algorithmes d'apprentissage automatique traditionnels ou de réseaux de neurones profonds, l'entraînement vise à minimiser l'erreur de prédiction et à optimiser les performances du modèle. Cette phase peut nécessiter plusieurs itérations, avec des ajustements continus des paramètres du

modèle, afin d'atteindre des niveaux de précision et de généralisation satisfaisants. En fin de compte, un modèle correctement entraîné sera capable de reconnaître efficacement les émotions à partir des expressions faciales dans une variété de situations et de conditions.

2.4.4 Classification des émotions

La classification des émotions repose sur l'analyse des expressions faciales pour reconnaître les émotions affichées. Cela commence par le prétraitement des images afin d'améliorer la qualité des données. Ensuite, on extrait les caractéristiques faciales pertinentes qui permettent d'associer chaque expression à une émotion spécifique, telle que la joie, la tristesse, la colère, la surprise, le dégoût ou la peur.

2.4.5 Test ou évaluation des performances

Cette phase consiste à évaluer les performances du modèle entraîné sur des données de test distinctes de celles utilisées pour l'entraînement et la validation. Les données de test permettent de mesurer la capacité du modèle à généraliser à de nouvelles situations et à des données réelles. En effectuant des prédictions sur ces données de test et en comparant les résultats avec les étiquettes réelles, on peut évaluer la précision et la fiabilité du modèle dans la reconnaissance des émotions. Le test permet également de détecter d'éventuels problèmes ou biais du modèle qui n'auraient pas été identifiés lors de l'entraînement ou de la validation. En fin de compte, le test est crucial pour s'assurer que le modèle est prêt à être déployé dans des applications réelles.

Des métriques telles que la précision, le rappel, le F-score et la matrice de confusion sont souvent utilisées pour évaluer les performances des modèles de classification des émotions.

2.4.5.1 Précision

La précision nous dit combien de nos prédictions sont correctes parmi toutes celles que notre modèle a faites. On la calcule en regardant combien de fois notre modèle a correctement prédit quelque chose par rapport au nombre total de fois où il a dit que c'était le cas. Une précision élevée signifie que notre modèle ne se trompe pas souvent.

2.4.5.2 Rappel

Le rappel nous dit combien de vrais cas positifs notre modèle a réussi à identifier parmi tous ceux qui existent réellement. On le calcule en regardant combien de fois notre modèle a bien identifié quelque chose par rapport au nombre total de fois où c'était vraiment le cas. Un rappel élevé signifie que notre modèle est bon pour repérer les situations positives.

2.4.5.3 F-score

Le F-score combine la précision et le rappel pour donner une seule mesure de performance du modèle. Il équilibre les erreurs de type faux positifs et faux négatifs. Cette mesure est particulièrement pratique lorsque les classes ne sont pas équitablement représentées dans les données.

2.5 Les expressions faciales

Les expressions du visage sont le résultat des mouvements des muscles sous la peau du visage, agissant comme une manière de communiquer les émotions sans mots. Ces expressions, générées par la contraction musculaire, façonnent les caractéristiques du visage pour aider à la communication et à l'expression émotionnelle [107] [108] les principales caractéristiques du visage qui influent sur les expressions sont la bouche, les yeux et les sourcils, tandis que d'autres éléments comme les rides, les couleurs ou les cheveux jouent un rôle moins important [109]. L'expression faciale est cruciale pour comprendre les émotions humaines, bien que la reconnaissance précise dans des contextes non contrôlés reste un défi important malgré les recherches en cours [110]. Les expressions faciales sont un indicateur clé pour identifier diverses émotions humaines, il est possible de détecter une gamme variée d'émotions, y compris les émotions de base couramment reconnues tells que [111] :

— **Joie :** Elle se manifeste par l'épanouissement d'une personne dans un état de satisfaction profonde, résultant de l'accomplissement de ses désirs, de ses succès, de son bien-être et de ses réalisations, ainsi que de son approche positive de la vie. Les signes distinctifs sur son visage incluent des mains levées, une bouche ouverte révélant les dents, des sourcils levés, et l'apparition de ridules sous les yeux plutôt qu'aux coins externes.

— **Tristesse :** C'est un état émotionnel marqué par une souffrance morale résultant dune insatisfaction ou de préoccupations, souvent déclenchées par une perte, un deuil ou un obstacle, entraînant souvent un repli sur soi. Sur le visage, cela se traduit par une baisse des sourcils vers l'intérieur et une courbure descendante des coins des lèvres.

— **Dégoût :** C'est quand quelqu'un n'aime pas certains aliments ou n'a pas envie de manger, souvent à cause d'un rejet ou d'un conflit avec quelqu'un. En général, les gens ont tendance à se retirer dans ces situations. Sur le visage, on peut voir la lèvre supérieure se relever, montrant plus de dents selon le niveau de dégoût, et des rides peuvent apparaître sur le nez.

— **Peur :** C'est un état où une personne se sent menacée par un danger réel ou perçu, souvent causé par une menace, un danger imminent ou la présence d'inconnus. Dans ces moments, la plupart des individus ont tendance à privilégier la fuite. Sur le visage, les caractéristiques typiques incluent des yeux grands ouverts avec des pupilles dilatées, des sourcils relevés et rapprochés, des paupières supérieures levées, exposant ainsi le blanc des yeux, et une tension des muscles autour de la bouche, souvent entraînant un abaissement de la lèvre inférieure.

— **Surprise :** C'est un état où une personne est prise de surprise par quelque chose d'inattendu, souvent en raison d'un danger imminent, d'une situation imprévue ou de la présence d'inconnus. Dans ces circonstances, la plupart des individus ont tendance à réagir en fuyant ou en sursautant. Les caractéristiques typiques sur le visage comprennent une élévation à la fois des parties interne et externe des sourcils, des rides horizontales pouvant se former sur le front, une ouverture de la bouche et des yeux, et des lèvres soit fermées hermétiquement soit légèrement entrouvertes (prêtes à crier). Ces signes sont souvent des préparatifs du corps à une éventuelle attaque, qu'elle soit physique ou verbale.

— **Colère :** C'est un état où une personne réagit violemment et de manière agressive face à une contrariété, souvent provoquée par une injustice ou une atteinte à ses valeurs. Dans ces situations, la plupart des gens ont tendance à adopter une attitude offensive. Sur le visage, les sourcils se froncent et se contractent, créant des rides verticales entre eux.

— **Mépris :** Est un sentiment intense de dédain, de désapprobation ou de dévalorisation envers quelque chose ou quelqu'un. C'est une attitude souvent marquée par un sentiment de supériorité ou de condescendance, où l'individu considère l'objet de son mépris comme insignifiant, médiocre ou indigne d'attention. Ce sentiment peut se manifester par des expressions faciales, des gestes ou des paroles méprisantes.

— **Intérét :** Est le degré d'importance ou de valeur qu'une personne accorde à quelque chose ou à quelqu'un. Cela peut inclure un intérêt personnel, financier, émotionnel, intellectuel ou autre. Il reflète généralement l'attraction ou l'attention qu'une personne porte à un sujet, une activité ou une personne en raison de son utilité, de son plaisir, ou de son importance perçue.

2.6 Etat de l'art sur la reconnaissance des émotions à partir des expressions faciales

Dans cette section, nous examinons l'état actuel de la recherche sur la reconnaissance des émotions à partir des expressions faciales, en mettant en évidence les bases de données essentielles et les avancées récentes.

2.6.1 Bases de données

Une base de données ou dataset est un ensemble structuré de données regroupées pour être utilisées dans diverses applications, notamment l'apprentissage automatique. Il peut contenir divers types de données telles que des textes, des images, des vidéos ou des enregistrements sonores, et peut être collecté dans des environnements contrôlés ou dans des conditions réelles. Les jeux de données servent de référentiel pour entraîner, tester et évaluer les performances des algorithmes et modèles. Dans les sections suivantes nous présenterons une variété de bases de données couramment utilisées dans le domaine de la reconnaissance des émotions faciales.

2.6.1.1 Fer2013 dataset

L'un des ensembles de données les plus largement utilisés est appelé FER qui signifie reconnaissance d'expression faciale. Composé de 35887 images, les photos sont mises à l'échelle à la configuration de taille 48x48. FER2013 a 28709 photos pour la formation, 3589 validations et 3589 images de test. FER2013 a trois colonnes qui définissent chaque photo. Ces colonnes sont nommées Type d'émotion en format numérique 0-6 qui décrit individuellement la colère, le dégoût, la peur, le bonheur, la tristesse, la surprise et la neutralité. La deuxième colonne est un tableau de valeurs numériques représentant les photos. La dernière colonne indique le type de photo, qu'il s'agisse d'un entraînement ou d'un test [112].

2.6.1.2 Cohn-Kanade dataset (CK +)

L'ensemble de données étendu CK+ comprend jusqu'à 920 images provenant des ensembles de données originaux CK+. Ces images ont été redimensionnées en 48x48 pixels, converties en niveaux de gris et recadrées avec haarcascade "frontalface" default. Pour assurer une identification claire, les images bruitées en raison de la lumière ambiante, du style de cheveux et de la couleur de la peau ont été ajustées à l'aide du classificateur Haar. Les données sont organisées en trois colonnes : émotion, pixels et usage. Les émotions sont classées comme suit : Anger (45 échantillons), Disgust (59 échantillons), Fear (25 échantillons), Happiness (69 échantillons), Sadness (28 échantillons), Surprise (83 échantillons), Neutral (593 échantillons), Contempt (18 échantillons). Chaque ligne de la colonne contient 2304 pixels (48x48). Cet ensemble de données a été développé initialement pour une comparaison avec le jeu de données ROHIT VERMA - FER2013 [113].

2.6.1.3 Facial Expression Research Group database (FERG)

Facial Expression Research Group (FERG) est une base de données spécialisée contenant des images de personnages de dessin animé, chacune annotée avec des expressions faciales spécifiques. Cette base de données comprend un total de 55769 images annotées, représentant six personnages distincts. Les images de chaque personnage sont regroupées en sept types d'expressions cardinales : colère, dégoût, peur, joie, neutre, tristesse et surprise [114].

2.6.1.4 Japanese female facial expression database (JAFFE)

La base de données JAFFE (Japanese Female Facial Expression) est une collection accessible au public contenant 213 images d'expressions faciales de 10 femmes japonaises. Chaque sujet a exprimé six émotions de base ainsi qu'une expression neutre, avec une répartition comme suit : 30 images de colère, 29 de dégoût, 33 de peur, 30 de bonheur, 31 de tristesse, 30 de surprise et 30 de neutre. Chaque émotion comprend entre 3 et 4 images par sujet. Les images sont en niveaux de gris et ont une résolution de 256 x 256 pixels. Toutes les photos ont été prises dans des conditions strictement contrôlées avec un éclairage similaire et sans occlusion comme les cheveux ou les lunettes, offrant une vue frontale claire des expressions faciales [115].

2.6.1.5 Karolinska Directed Emotional Faces (KDEF)

La base de données KDEF est un ensemble de données accessible au public, comprenant 4900 images d'expressions faciales. Ce jeu de données contient des images de 70 individus, chacun affichant sept expressions émotionnelles différentes. Chaque expression est photographiée deux fois sous cinq angles différents [115].

2.6.1.6 AffectNet

AffectNet est le plus grand ensemble de données fournissant à la fois des annotations catégoriques et des annotations Valence-Arousal. Il contient plus d'un million d'images provenant d'internet, obtenues en recherchant des mots-clés liés à l'expression faciale sur trois moteurs de recherche. Parmi ces images, 450 000 sont annotées manuellement avec huit étiquettes d'expression de base, similaires à celles utilisées dans FERPlus [116].

2.6.1.7 OAHEGA

OAHEGA est une base de données qui propose une variété d'images qui représentent six émotions différentes : le bonheur, la colère, la tristesse, le neutre, la surprise et Ahegao. Ces images, en format RVB, montrent des visages créés de manière artificielle, chacun lié à une émotion particulière. La création et la publication de ce jeu de données sont réalisées par Volodymyr Kovenko et Vitalii Shevchuk. Ces images ont été collectées en extraisant

du contenu des plateformes de médias sociaux comme Facebook et Instagram, ainsi que des vidéos YouTube, en plus des ensembles de données déjà existants tels que IMMDB et AffectNet respectivement [117].

2.6.2 Synthèse des travaux récents

Dans cette partie, quelques travaux récents dans le domaine de la reconnaissance des émotions à partir des expressions faciales sont présentés, chronologiquement dans le tableau (2.1).

TABLE 2.1 : Tableau des travaux récents

Référence	**Technique**	**Émotions**	**Métriques d'évaluation**	**Dataset**	**Résultats**
Minaee et Abdolrashidi (2019) [118]	Attentional Convolutional Network	Colère, Dégoût, Peur, Heureux, Triste, Surprise, Neutre	Accuracy	FER-2013, CK+, FERG, JAFFE	70.02%, 98.0%, 99.3%, 92.8%
Qin et al. (2020) [119]	Gabor Wavelet Transform, CNN	/	Accuracy	CK+	96.81%
Mehendale (2020) [120]	CNN	/	Accuracy	CK+	96%
Tomar et al. (2021) [121]	CNN, SVM	/	Accuracy	RAF-DB, FER2013	92.80%, 90.85%

Référence	**Technique**	**Émotions**	**Métriques d'évaluation**	**Dataset**	**Résultats**
Chen et Khaireddin (2021) [122]	CNN (VGGNet)	Colère, Neutre, Dégoût, Peur, Joie, Triste, Surprise	Accuracy	FER2013	73.28%
Akhand et al. (2021) [123]	VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, DenseNet-161	Peur, Colère, Dégoût, Triste, Heureux, Surprise, Neutre	Accuracy	KDEF, JAFFE	96.51%, 99.52%
Aouayeb et al. (2021) [124]	Vision Transformer (ViT)	Colère, Neutre, Dégoût, Peur, Joie, Triste, Surprise	Accuracy	CK+, SFEW, JAFFE, RAF-DB	99.80%, 54.29%, 92.92%, 87.22%
Sati et al. (2021) [125]	Haar Cascades, ResNet-34	Colère, Dégoût, Peur, Heureux, Triste, Surprise, Neutre	Accuracy	/	/

Référence	**Technique**	**Émotions**	**Métriques d'évaluation**	**Dataset**	**Résultats**
Rahul et al. (2022) [126]	CNN, RNN	Surprise, Tristesse, Neutre, Joie, Colère, Dégoût, Peur	Accuracy	EMOTIC, FER-13, FERG	72.64%, 94.08%, 68.10%
Sarma et al. (2022) [127]	CNN	Colère, Dégoût, Peur, Heureux, Triste, Surprise, Neutre	Accuracy	Fer2013	58%
Chaudhari et al. (2022) [128]	ResNet-18, ViT-B/16/S, ViT-B/16/SG, ViT-B/16/SAM	/	Accuracy	FER-2013, AffectNet, CK+48	50.05%, 52.25%, 52.42%, 53.10%
Agrawal et al. (2023) [129]	AlexNet, VGG16, VGG19, ResNet50, InceptionV3	Heureux, Triste, Colère, Neutre, Surprise, Ahegao	Accuracy	OAHEGA	Train :92% Test :54%

Référence	**Technique**	**Émotions**	**Métriques d'évaluation**	**Dataset**	**Résultats**
Helaly et al. (2023) [130]	ResNet-18	Colère, Dégoût, Peur, Heureux, Triste, Surprise, Neutre	Accuracy	FER2013, CK+	83%, 98%
Alonazi et al. (2023) [131]	AFER-POADCNN	Colère, Mépris, Dégoût, Peur, Heureux, Neutre, Triste, Surprise	Accuracy	FER	99.05%
Chowdary et al. (2023) [132]	Resnet50, VGG19, InceptionV3, Mobile Net	/	Accuracy	CK+	96%
Mukhiddinov et al. (2023) [133]	CNN	/	Accuracy	AffectNet	69.3%

Référence	**Technique**	**Émotions**	**Métriques d'évaluation**	**Dataset**	**Résultats**
Bialek et al. (2023) [134]	RESNET50, VGG16	Colère, Dégoût, Peur, Heureux, Neutre, Triste, Surprise	Accuracy, Precision, Recall, F1-Score	FER2013	RESNET50 : 72.72%, 73.21%, 72.27%, 72.73%, VGG16 : 70.22%, 74.19%, 65.30%, 69.26%
Prasad et Chandana (2023) [135]	Efficient Net, YOLOv4, DenseNet	Colère, Joie, Tristesse, Dégoût, Surprise, Peur, Neutre	Accuracy	RGB-D-T	95.97%
Assiri et Hossain (2023) [136]	CNN	/	Accuracy	/	96.87%
JS et al. (2024) [137]	CNN, SVM	/	Accuracy	JAFFE, FERC-2013	/

Référence	**Technique**	**Émotions**	**Métriques d'évaluation**	**Dataset**	**Résultats**
Talaat et al. (2024) [138]	Xception, ResNet, MobileNet	Colère, Peur, Joie, Naturel, Tristesse, Surprise	Accuracy, Sensitivity, Specificity, AUC	/	95.23%, 93.2%, 94.21%, 91.34%
Ahmad et al. (2024) [139]	ShuffleNet V2	Heureux, Peur, Surprise, Neutre, Tristesse, Colère, Dégoût	Accuracy	CK+, FER_2013, TFEID, KMU_FED, KDEF	98%, 97%, 97%, 99%, 99%
Bellamkonda et Settipalli (2024) [140]	VGG16, EFLLCNN	Colère, Dégoût, Peur, Heureux, Neutre, Triste, Surprise	Accuracy	JAFFE, CK+, MUG, KDEF	VGG16 : 97.24%, 97.94%, 95.97%, 91.27%, EFLLCNN : 98.50%, 99.20%, 97.23%, 92.53%

Référence	Technique	Émotions	Métriques d'évaluation	Dataset	Résultats
Bakariya et al. (2024) [141]	CNN	Colère, Peur, Joie, Neutre, Triste, Surprise	Accuracy	FER-2013, OAHEGA	73.02%

2.7 Conclusion

Dans ce chapitre, nous avons abordé les fondements théoriques de la reconnaissance des émotions à partir des expressions faciales, en présentant les concepts clés liés aux émotions ainsi que les principales sources de données émotionnelles. Nous avons également examiné le processus de reconnaissance des émotions et exploré l'importance de l'expression faciale dans ce contexte. Dans la deuxième partie de notre étude, nous nous sommes concentrés sur l'état de l'art en matière de reconnaissance des émotions à partir des expressions faciales, en mettant en lumière les bases de données utilisées et en synthétisant les travaux récents dans ce domaine.

Dans le chapitre suivant, nous mettrons en uvre des Vision Transformers pour développer un système de classification des émotions basé sur les expressions faciales.

CHAPITRE 3

UN SYSTÈME DE CLASSIFICATION DES ÉMOTIONS UTILISANT VIT POUR L'ANALYSE DES EXPRESSIONS FACIALES

3.1 Introduction

Les expressions faciales analytiques sont utiles pour reconnaître les émotions. Cependant, en raison de similitudes dans diverses situations, ces expressions peuvent entraîner des erreurs. Il est donc primordial de se concentrer sur l'amélioration des résultats en augmentant la précision de classification. Notre objectif principal est de créer un modèle capable de classer efficacement différentes émotions à partir des expressions faciales. Ce modèle pourrait être utilisé dans divers domaines tels que : la sécurité, la santé, l'automobile , la finance, l'éducation, le marketing et l'analyse des sentiments.

Dans ce chapitre, nous présentons notre système basé sur les Vision Transformers (ViT). Le choix de ViT s'explique par leur capacité à capturer des informations contextuelles à

long terme et à surmonter les limitations des modèles de convolution traditionnels grâce à leur mécanisme d'attention. Ce chapitre est structuré en deux sections. La première section décrit les étapes de conception du modèle ViT, notamment la préparation des données et l'architecture du modèle. La seconde section présente les expérimentations et les résultats obtenus, démontrant l'efficacité de notre modèle pour la reconnaissance des émotions à partir des expressions faciales.

3.2 Architecture générale du système

Dans cette section, nous allons décrire notre système de classification des émotions basé sur les expressions faciales. Avant d'entrer dans les détails, nous présentons tout d'abord son architecture générale. Cette structure est illustrée par le schéma en blocs présenté dans la figure (3.1).

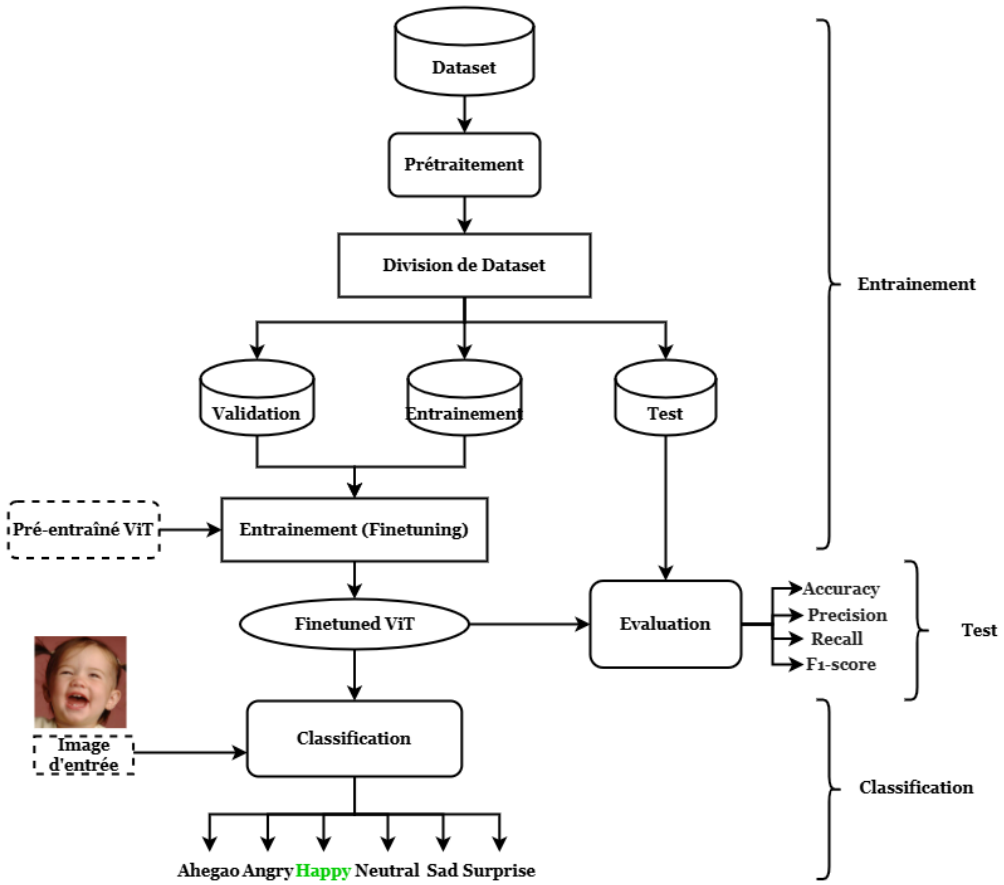

FIGURE 3.1 – Architecture de notre Système

En principe, notre processus commence par l'utilisation d'une base de données regroupant les images nécessaires pour l'entraînement du modèle ViT. Ensuite, nous passons à l'étape de prétraitement des données, où nous améliorons la qualité des images et les préparons pour l'entraînement. À l'aide de la méthode de fine-tuning, nous entraînons le modèle sur une partie des données d'apprentissage traitées, tandis que l'autre partie est réservée à la validation. Lors de la phase de test ou de classification, ce modèle sera utilisé pour classer de nouvelles images en différentes catégories d'émotions telles que la joie, la tristesse, la colère, la surprise et la neutralité.

Après avoir présenté l'architecture générale de notre système, nous allons maintenant examiner en détail, à travers les trois phases suivantes, les différentes étapes nécessaires à sa mise en oeuvre.

3.2.1 Phase d'apprentissage

L'objectif de l'apprentissage est de réentrainer le modèle pré-entrainé ViT, qui sera ensuite employé pour classer les émotions en se basant sur les expressions faciales présentes dans les images.

3.2.1.1 Préparation des données

Pour réaliser notre système de reconnaissance des émotions avec le ViT, il est essentiel de collecter des ensembles de données d'images faciales étiquetées avec des annotations émotionnelles précises et d'inclure des expressions faciales des émotions de base. Cette diversité permet au modèle ViT de reconnaître les émotions dans diverses situations réelles.

— **Collecte des données (ensemble de données) :** La première étape dans la réalisation de notre système de reconnaissance des émotions à partir des expressions faciales est la collecte des données. Cette collecte peut se faire de deux façons : en récueillant des données spécifiques au domaine visé, par exemple en collectant des données d'une entreprise pour un système personnalisé, ou en utilisant des ensembles de données existants disponibles pour la recherche (benchmarks standards). Dans notre cas, nous avons choisi d'utiliser deux ensembles de données stadards pour le ré-entrainement et l'évaluation des performances du modèle ViT, à savoir, CK+ et OAHEGA. Cela nous

permet d'évaluer notre modèle de façon objective en comparant nos résultats avec ceux d'autres chercheurs sur les mêmes ensembles de données.

— **Prétraitement des données (data preprocessing) :** Le prétraitement des données est une étape cruciale dans le développement de tout modèle d'apprentissage automatique, car la qualité des données influe directement les performances du modèle. Avant d'être fournies au modèle, les données subissent plusieurs opérations de transformation. Par exemple, dans le cadre de la détection des émotions faciales à l'aide du modèle ViT, les images sont d'abord converties en niveaux de gris puis transformées en images RGB (red, green, blue) pour s'adapter au modèle. Ensuite, les images sont redimensionnées pour assurer une taille uniforme, ce qui est essentiel pour les modèles basés sur les Transformers. De plus, des techniques d'augmentation de données telles que la rotation aléatoire, l'ajustement de la netteté, et la symétrie horizontale sont appliquées pour enrichir la diversité des données d'entraînement. Ces opérations augmentent la robustesse du modèle en le rendant capable de généraliser à des données non vues auparavant.

  Une opération essentielle de prétraitement est la normalisation des données pour assurer la qualité et la cohérence des données d'entrée dans le développement des modèles d'apprentissage automatique. Elle consiste à ajuster les valeurs des pixels des images pour les rendre compatibles avec une plage spécifique, entre 0 et 1 en l'occurnce. Cela est crucial pour stabiliser et accélérer le processus d'apprentissage du modèle. La normalisation permet également d'éviter les problèmes liés aux différences d'échelle entre les différentes caractéristiques des données. Dans le cadre de la préparation des données pour un modèle ViT, la normalisation est appliquée en utilisant la moyenne (mean) et l'écart-type (std) des pixels, ajustés pour correspondre aux exigences des images en niveaux de gris converties en RGB. Cette étape assure que les données d'entrée sont cohérentes, ce qui facilite une meilleure performance du modèle de reconnaissance des émotions.

— **Division des données (Data splitting) :** Nous avons séparé cette base de données en trois parties distinctes : un ensemble pour l'entraînement, un autre pour la validation, et enfin un dernier pour les tests (train/val/test).

- **L'ensemble d'entraînement (train) :** L'ensemble d'entraînement est essentiel pour permettre au modèle d'apprendre à partir des données, d'ajuster ses paramètres pour améliorer les prédictions et de développer la capacité de généraliser à de nouvelles données.
- **L'ensemble de validation (validation) :** L'ensemble de validation est utilisé pour évaluer les performances du modèle pendant son entraînement de manière à ce qu'il généralise bien aux nouvelles données, en évitant les pièges du surapprentissage et du sous-apprentissage.
- **L'ensemble de test (test) :** L'ensemble de test garantit que le modèle n'est pas seulement performant sur les données d'entraînement, mais aussi sur de nouvelles données, offrant ainsi une évaluation impartiale et fiable de ses capacités de généralisation.

3.2.1.2 Entrainement du modèle

Le processus d'entraînement commence par l'initialisation d'un modèle ViT pré-entraîné sur ImageNet, suivi d'une adaptation spécifique aux émotions à partir d'images faciales étiquetées.

— **Initialisation du modèle ViT :** Nous avons commencé par initialiser un modèle ViT pré-entraîné sur l'ensemble de données ImageNet. Ensuite, nous avons réentraîné les dernières couches du modèle sur un ensemble de données plus spécifique, constitué d'images faciales annotées pour les émotions. Cette approche, appelée fine-tuning, permet au modèle de s'adapter à la tâche spécifique tout en conservant les connaissances générales acquises lors du pré-entraînement sur un ensemble de données plus large. Ainsi, le modèle ViT peut généraliser efficacement à de nouvelles tâches de reconnaissance des émotions basées sur les expressions faciales, sans nécessiter un entraînement coûteux sur de vastes ensembles de données.

— **Extraction des caractéristiques (Feature extraction) :** Dans notre architecture ViT, l'extraction des caractéristiques est une étape clé. Nous convertissons les images faciales en vecteurs de caractéristiques, ou chaque vecteur capturant les éléments essentiels des expressions faciales, comme les mouvements des muscles et les variations de texture. Grâce à son architecture axée sur l'attention, le ViT peut détecter et se

concentrer sur les zones clés du visage, telles que les yeux, la bouche et les sourcils, afin d'obtenir des informations pertinentes. Ces vecteurs de caractéristiques constituent ensuite le fondement de la classification des émotions, permettant au modèle d'analyser efficacement les expressions faciales et de déterminer avec précision l'émotion correspondante. L'architecture générale de ViT se compose de deux parties principales : l'encodeur de patches (patch encoder) et l'encodeur Transformer (Transformer encoder). Ces composants jouent un rôle crucial dans l'extraction des caractéristiques à partir des images, garantissant ainsi une analyse approfondie et précise des expressions faciales.

- **L'encodeur de patches (patch encoder) :** L'encodeur de patches divise chaque image faciale en petits segments appelés patches, puis transforme chaque patch en un vecteur d'embedding. Cela capture les informations locales et permet de traiter les images comme une séquence de mots, facilitant l'analyse des expressions faciales.

  La couche patch embedding : Les images sont divisées en 256 patches de taille 16x16 (voir la figure 3.2), les patches sont non superposés qui sont ensuite incorporés en tant que tokens (jetons).

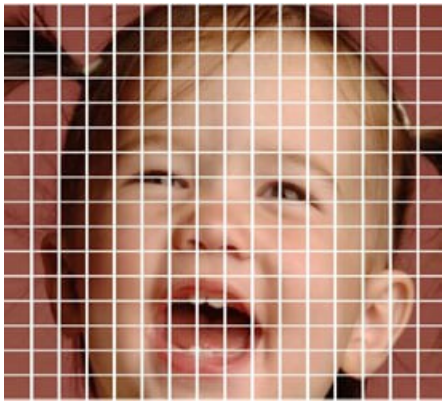

FIGURE 3.2 – Image divisées en 256 patches de taille 16x16

Projection linéaire : Les patches sont transformés à travers une projection linéaire dans un espace vectoriel de dimension réduite, donnant lieu à une séquence de vecteurs, chacun représentant un patch. Dans notre système de reconnaissance des émotions à partir des expressions faciales dans les images, nous utilisons le ViT. Étant donné que le ViT requiert des images en 2D en entrée, il est nécessaire

de les convertir en une séquence 1D avant d'appliquer la projection linéaire sur chaque séquence de patches. Cette méthode permet de simplifier le traitement de chaque patch par les couches de l'encodeur Transformer en réduisant le nombre de dimensions et de paramètres en entrée. Après cette étape, les vecteurs représentant les patches sont ensuite analysés par les différentes couches du Transformer. Ces couches apprennent les relations complexes entre les patches ainsi que les caractéristiques globales et les relations spatiales entre eux. La représentation finale obtenue est alors utilisée pour classer les images, en identifiant les émotions à partir des expressions faciales.

Ajout de la position encodée : Dans notre système de reconnaissance des émotions à partir des expressions faciales, l'ajout de la position encodée est une étape fondamentale. Lors de l'utilisation du ViT, après avoir divisé les images faciales en patches et appliqué un embedding linéaire, nous ajoutons des encodages de position à ces vecteurs de caractéristiques. Ces encodages de position fournissent des informations contextuelles sur la localisation spatiale des patches au sein de l'image, permettant au modèle de préserver la structure et la disposition des traits du visage. En combinant ces encodages positionnels avec les vecteurs de caractéristiques, le ViT est capable de mieux comprendre les relations spatiales entre différents éléments du visage, comme les yeux, la bouche et les sourcils. Cette compréhension améliorée permet une analyse plus précise des expressions faciales, et donc une reconnaissance des émotions plus fiable et efficace. Dans notre système, l'embedding linéaire et l'ajout de la position encodée jouent des rôles cruciaux. Au sein du ViT, les images faciales sont d'abord divisées en patches, puis chaque patch est transformé en un vecteur de caractéristiques via un embedding linéaire. Cette étape permet de convertir les informations visuelles brutes en représentations numériques compactes et significatives, facilitant ainsi l'interprétation et le traitement des données par l'encodeur Transformer.

Embedding linéaire : Dans le cadre de notre système, le concept d'embedding linéaire joue un rôle important. Au sein du ViT, les images faciales sont d'abord divisées en patches, puis chaque patch est transformé en un vecteur de caractéristiques via un embedding linéaire. Cette étape permet de convertir les

informations visuelles brutes en représentations numériques compactes et significatives. L'embedding linéaire facilite l'interprétation et le traitement des données par l'encodeur Transformer, permettant ainsi au modèle de se concentrer sur les éléments clés des expressions faciales, comme les mouvements des muscles et les variations de texture. Grâce à cette transformation, notre système peut analyser avec précision les émotions en exploitant les détails subtils des expressions faciales, assurant ainsi une reconnaissance émotionnelle fiable et efficace. Après cette transformation initiale, nous ajoutons des encodages de position à ces vecteurs de caractéristiques. Ces encodages fournissent des informations contextuelles sur la localisation spatiale des patches au sein de l'image, permettant au modèle de préserver la structure et la disposition des traits du visage. En combinant ces encodages positionnels avec les vecteurs de caractéristiques, le ViT est capable de mieux comprendre les relations spatiales entre différents éléments du visage, comme les yeux, la bouche et les sourcils. Cette compréhension améliorée permet une analyse plus précise des expressions faciales, et donc une reconnaissance des émotions plus fiable et efficace.

- **L'encodeur Transformer (Transformer encoder) :** L'encodeur du Transformer traite les vecteurs des patchs d'image en les passant à travers plusieurs couches successives. Ces couches, organisées en pile, apprennent à comprendre les relations entre les différents patchs de l'image. Chaque couche du transformer comporte deux sous-couches principales : une self-attention multi-tête et une couche feedforward. La self-attention multi-tête permet d'apprendre les relations entre les patchs à diverses échelles, tandis que la couche feedforward introduit des non-linéarités dans chaque représentation de patch, enrichissant ainsi les caractéristiques apprises (figure (3.3)).

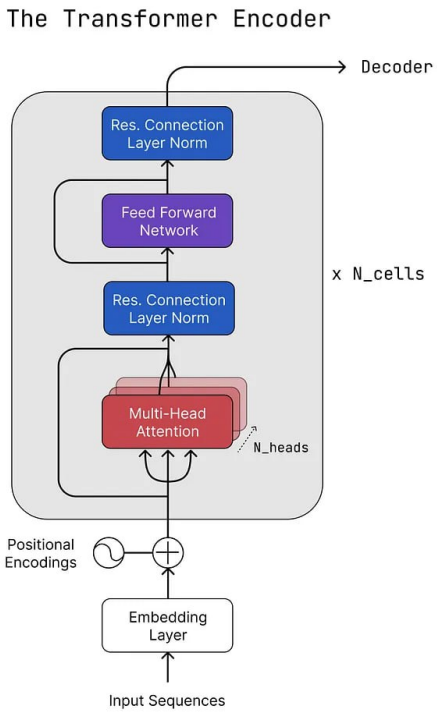

FIGURE 3.3 – L'architecture de l'encodeur du Transformer [142]

Pour optimiser les paramètres de notre modèle et améliorer la vitesse de l'entraînement dans notre système de reconnaissance des émotions à partir des expressions faciales en utilisant ViT, nous procédons de la manière suivante :

Tout d'abord, la sortie de l'encodeur Transformer est transmise à une couche de normalisation de lot (batch normalization). Ensuite, une couche de classification linéaire convertit la forme des données en un vecteur unidimensionnel. Nous utilisons le *weight decay* après cette transformation pour éviter le sur-apprentissage et assurer que notre modèle ne devienne pas trop dépendant de certaines caractéristiques spécifiques des données.

Dans la partie classification, nous avons intégré un bloc MLP (Multilayer Perceptron) comprenant une couche dense avec 100 neurones. Enfin, la sortie finale de notre modèle est générée par une couche dense avec un nombre de neurones correspondant aux différentes catégories d'émotions que nous voulons reconnaître. Les sorties de cette couche sont des probabilités qui indiquent notre confiance dans la classification de chaque émotion. Ainsi, plus la probabilité associée à une

catégorie est élevée, plus nous sommes sûrs que cette émotion est présente sur l'image analysée (figure (3.4)).

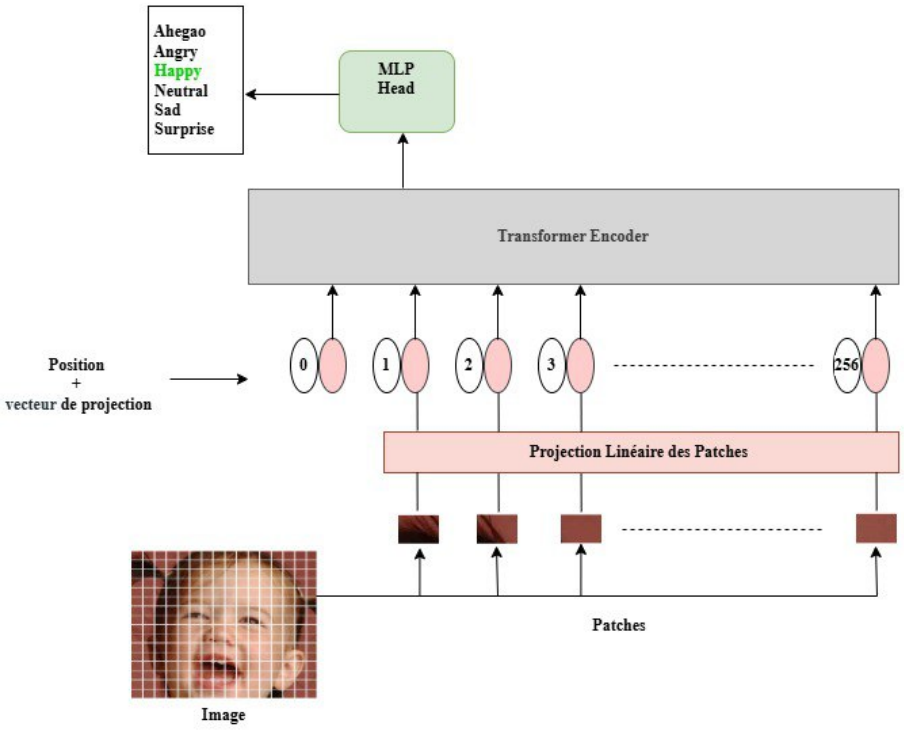

FIGURE 3.4 – L'architecture globale de ViT

Le fine-tuning : Nous avons adopté une approche de transfer learning pour notre système, en utilisant le modèle pré-entraîné ViT. Ce modèle a été initialement formé sur des ensembles de données volumineux comme ImageNet, couvrant une grande diversité d'objets du monde réel. En exploitant les représentations générales apprises par le ViT, nous les avons adaptées à notre propre ensemble de données axé sur les émotions. Ce processus de fine-tuning consiste à ajuster les poids du modèle pré-entraîné pour mieux correspondre aux caractéristiques spécifiques de notre tâche. Ainsi, nous avons configuré le ViT pour la reconnaissance des émotions, en tirant parti de ses connaissances existantes sans avoir à recommencer l'entraînement depuis le début.

Ajustement des paramètres : Lors de l'entraînement de notre modèle, nous utilisons une boucle qui parcourt les données d'entraînement sur plusieurs époques. À chaque lot (batch) de données, nous réalisons les étapes suivantes :

Initialisation des hyperparamètres : Nous définissons les valeurs initiales de tous

les hyperparamètres, tels que le taux d'apprentissage (Learning rate), la taille du lot (batch size), le nombre dépoques.

Boucle d'entraînement :

Calcul de la perte : En comparant ces prédictions avec les étiquettes réelles (par exemple, la joie, la tristesse, la colère, la surprise, la neutralité et l' ahegao, nous calculons la perte en utilisant une fonction de perte appropriée (la cross-entropy).

Évaluation sur l'ensemble de validation : Après chaque époque (ou après un certain nombre d'itérations), nous évaluons les performances du modèle sur un ensemble de données de validation pour surveiller son comportement et détecter d'éventuels problèmes d'overfitting.

Ajustement des hyperparamètres : Sur la base des performances sur l'ensemble de validation, nous ajustons les valeurs des hyperparamètres (comme le taux d'apprentissage) pour optimiser les performances du modèle.

Arrêt prématuré (Early Stopping) : Nous surveillons également les performances sur l'ensemble de validation à travers le calcul de l'accuracy pour décider quand arrêter l'entraînement afin d'éviter le surapprentissage.

Sauvegarde du modèle : Nous sauvegardons périodiquement les poids du modèle afin de pouvoir reprendre l'entraînement à partir d'un point précédent si nécessaire.

3.2.2 Phase de test

Une fois que notre modèle de reconnaissance des émotions est entraîné, nous l'examinons sur l'ensemble de test pour évaluer ses performances sur des données inconnues. En évaluant avec diverses mesures de performance telles que l'exactitude (accuracy), la précision, le rappel et le score F1, nous jugeons l'efficacité de notre modèle. Cette évaluation approfondie, comparant les prédictions du modèle aux émotions réelles, garantit sa fiabilité et son adaptation à de nouvelles données, assurant des performances optimales dans la reconnaissance des émotions à partir des expressions faciales.

3.2.3 Classification

Dans la phase de classification, nous utilisons notre modèle pour prédire la classe d'émotion d'une nouvelle image. Cette prédiction est donnée sous forme de probabilité pour chaque émotion. La classe prédite est celle ayant la probabilité la plus forte.

3.3 Résultats expérimentaux et discussion

Dans cette section, nous présentons d'abord les deux ensembles de données OAHEGA et CK+, utilisés pour l'apprentissage et l'évaluation de notre modèle ViT. Ensuite, nous réalisons des expérimentations pour trouver la meilleure combinaison d'hyperparamètres. Enfin, nous comparons les meilleurs résultats obtenus avec ceux de quelques travaux de la littérature de la reconnaissance des émotions à partir des expressions faciales.

3.3.1 Base de données OAHEGA

Nous avons choisi une base de données provenant de l'Université d'Oulu en Finlande pour garantir l'équité dans notre recherche scientifique. Cette base de données, appelée OAHEGA, est disponible en open source sur Kaggle. Elle contient 15453 images RVB de visages humains artificiels, chacun associé à une émotion spécifique. Ces images ont été capturées dans des conditions contrôlées pour garantir leur qualité et une diversité d'expressions. Les émotions couvertes incluent la joie, la tristesse, la colère, la surprise, la peur et le dégoût. Chaque image est annotée avec des étiquettes émotionnelles précises, ce qui facilite leur utilisation pour l'apprentissage automatique et la reconnaissance des émotions. La répartition des émotions dans cet ensemble de données est montrée sur la figure (3.5).

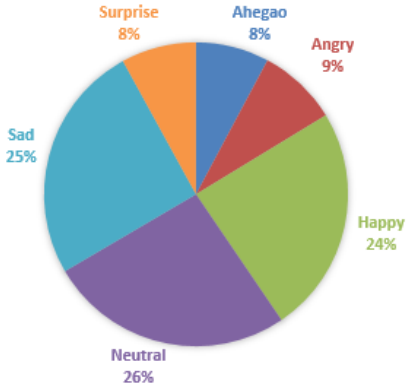

FIGURE 3.5 – Répartition des émotions dans la base de données OAHEGA

3.3.2 Base de données CK+

L'ensemble de données étendu CK+ est une extension de l'ensemble de données CK+ original, contenant 981 images supplémentaires. Ces images ont été prétraitées pour une meilleure cohérence et une identification plus facile des visages. Elles ont été redimensionnées à 48x48 pixels, converties en niveaux de gris et recadrées à l'aide du classificateur Haar. De plus, les images ont été ajustées pour réduire le bruit causé par la lumière ambiante, les coiffures et la couleur de la peau. Les données sont organisées en trois colonnes : émotion, pixels et usage. Les images sont étiquetées avec l'une des sept émotions suivantes : anger 135 échantillons, disgust 177 échantillons, fear 75 échantillons, happy 207 échantillons, sad 84 échantillons, surprise 249 échantillons, contempt 54 échantillons (voir la figure 3.6).

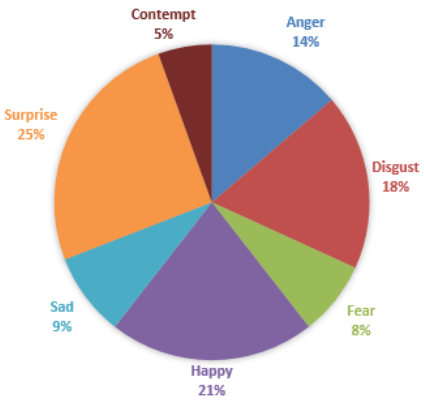

FIGURE 3.6 – Répartition des émotions dans la base de données CK+

3.3.3 Ajustement des hyperparamètres

Dans cette section, nous montrons les résultats expérimentaux du modèle ViT pour la classification des émotions exprimées par les visages dans les images. Nous réalisons des expériences pour évaluer les performances de l'architecture et d'identifier les hyperparamètres qui offrent les meilleurs résultats. Les hyperparamètres que nous étudions comprennent la taille des lots (batch size), la décroissance des poids (weight decay), le pas d'apprentissage (learning rate) et le nombre d'époques (number of epochs).

3.3.3.1 Taille de batch

Dans cette section, nous examinons l'impact de la taille de batch sur les performances du modèle. Ce paramètre représente le nombre d'échantillons d'images regroupés et traités ensemble à chaque itération pendant l'entraînement. Nous explorons différentes tailles de batch pour comprendre leur influence sur l'accuracy du modèle. Le tableau (3.1) présente les résultats des expérimentations réalisées pour analyser cet effet. Notre objectif est de déterminer la taille de batch optimale qui garantit les meilleures performances de classification des émotions à partir des expressions faciales.

TABLE 3.1 – Impact de la taille du batch sur la précision d'entraînement et de validation

Taille Batch	Train Accuracy	validation Accuracy	F1 score
4	**100%**	**100%**	**100%**
8	99.42%	99.43%	99.43%
16	90.28%	90.29%	90.23%
32	96.57%	96.57%	96.55%
64	98.85%	98.86%	98.84%

Les résultats du tableau (3.1) montrent que la taille de batch influence significativement les performances du modèle de classification des émotions. Des tailles de batch de 4, 8 et 64 offrent des précisions élevées, avec des accuracy d'entraînement et de validation de 100%, 99.43%, et 98.86% respectivement. En revanche, une taille de batch de 16 donne des résultats moins optimaux avec une accuracy de 90.29%.

3.3.3.2 La décroissance des poids (weight decay)

Dans notre exploration pour réaliser un modèle ViT performant dans la reconnaissance des émotions à partir des expressions faciales dans les images, nous étudions la décroissance des poids, également connue sous le nom de *weight decay*. L'idée principale derrière est de régulariser les poids du modèle en ajoutant une pénalité à la fonction de perte, ce qui aide à réduire le surapprentissage en limitant la magnitude des poids. Cette régularisation est importante pour notre objectif principal, qui est d'avoir un modèle ViT capable de généraliser efficacement à la fois aux données d'apprentissage et aux nouvelles données de test, assurant ainsi des performances fiables et cohérentes dans divers contextes. Cette expérimentation vise à déterminer le taux optimal de régularisation du *weight decay*, permettant ainsi de réduire l'overfitting et d'améliorer la capacité de généralisation de notre modèle ViT. Nous avons examiné les résultats de notre expérimentation pour divers taux de *weight decay* dans le tableau suivant.

TABLE 3.2 – Impact du taux de weight decay sur la précision du modèle.

Weight decay	Train Accuracy	validation Accuracy	F1 score
0.5	99.42%	99.43%	99.43%
0.05	87.42%	87.43%	86.90%
0.005	**100%**	**100%**	**100%**
0.0005	98.85%	98.86%	98.84%
0.00005	45.71%	85.71%	85.75%

Les résultats du tableau (3.2) montrent que le taux de weight decay influence significativement les performances du modèle ViT. Un taux de 0.005 offre les meilleures performances avec une accuracy de 100% pour l'entraînement et le validation, ainsi qu'un F1 score parfait, indiquant une excellente capacité de généralisation. Les taux de 0.5 et 0.0005 montrent également de bonnes performances, avec des accuracies proches de 99.43% et 98.86% respectivement. En revanche, un taux de 0.05 conduit à une diminution notable des performances avec une accuracy de 87.43%, et un taux de 0.00005 entraîne une surapprentissage marqué, avec une accuracy d'entraînement faible de 45.71% mais une validation accuracy de 85.71%. Ces observations soulignent l'importance de choisir un taux de weight decay optimal pour équilibrer l'apprentissage et la généralisation du modèle.

3.3.3.3 Le pas d'apprentissage (Learning rate)

Ce paramètre joue un rôle important dans l'entraînement des modèles d'apprentissage automatique, car il détermine la taille des pas effectués lors de la mise à jour des poids du modèle. Un learning rate trop élevé peut entraîner une divergence du modèle, tandis qu'un learning rate trop faible peut entraîner une convergence lente ou bloquée dans un minimum local. Nous explorons différentes valeurs de learning rate pour trouver celui qui permet une convergence rapide et stable du modèle tout en évitant les problèmes de divergence ou de surapprentissage. Les performances mesurées en termes d'accuracy pour différentes valeurs de pas d'apprentissage sont résumées dans le tableau (3.3).

TABLE 3.3 – L'effet des expériences sur la définition du taux d'apprentissage.

Pas d'apprentissage	Train Accuracy	validation Accuracy	F1 score
0.01	45.71%	85.71%	85.75%
0.001	82.28%	82.29%	82.30%
0.0001	98.85%	98.86%	98.84%
0.00001	**100%**	**100%**	**100%**

D'après le tableau (3.3) les résultats indiquent que le learning rate a un impact majeur sur l'accuracy du modèle, que ce soit sur les données d'entraînement ou de validation. Un learning rate de 0.01 donne une accuracy d'entraînement faible 45.71% mais une accuracy de validation élevée 85.71%. En revanche, avec un learning rate de 0.00001, l'accuracy atteint 100% sur les deux types de données, ce qui pourrait indiquer un surapprentissage. Le learning rate de 0.0001 semble être un choix équilibré, offrant une convergence rapide et une bonne capacité de généralisation 98.85% d'accuracy sur les données de validation.

3.3.3.4 Les étapes de warm-up (warm-up steps)

Les étapes de warmup en apprentissage profond se réfèrent à une phase initiale où le taux d'apprentissage est progressivement augmenté d'une petite valeur jusqu'au taux cible pour stabiliser le processus d'optimisation et prévenir la divergence. Cette technique est particulièrement utile lorsqu'on entraîne un modèle de réseau de neurones à partir de zéro. En effet, commencer avec un taux d'apprentissage élevé peut entraîner une instabilité ou une divergence du modèle. Le warmup permet d'augmenter progressivement le taux d'apprentissage sur un nombre spécifié d'époques, assurant ainsi une convergence plus stable et efficace

[143].

TABLE 3.4 – Impact des étapes de warm-up sur les performances du modèle

warm-up steps	Train Accuracy	validation Accuracy	F1 score
100	**100%**	**100%**	**100%**
200	87.42%	87.43%	86.90%
500	99.42%	99.43%	99.43%
1000	45.71%	85.71%	85.75%

Les résultats du tableau (3.4) montrent que le nombre d'étapes de warm-up a un impact significatif sur les performances du modèle ViT. Avec 100 étapes de warm-up, le modèle atteint une précision parfaite de 100% pour l'entraînement et le validation, ainsi qu'un F1 score parfait, indiquant une convergence optimale. À 500 étapes, les performances restent élevées avec environ 99.43% de précision, démontrant également une bonne stabilité. En revanche, 200 et 1000 étapes de warm-up entraînent une diminution notable des performances, avec des précisions de 87.43% et 85.71% respectivement

3.3.3.5 Le nombre d'époques d'apprentissage (number of epochs)

Pour examiner le comportement de notre modèle pendant la phase d'apprentissage, nous examinons de près le paramètre du nombre d'époques d'apprentissage, qui indique combien de fois l'ensemble des données est parcouru par le modèle pendant l'entraînement. Ce choix a un impact direct sur la capacité du modèle à assimiler les données d'entraînement et à généraliser à de nouvelles données. Un nombre d'époques trop faible peut entraîner un sous-apprentissage, où le modèle ne comprend pas suffisamment les schémas des données, tandis qu'un nombre d'époques trop élevé peut conduire à un surapprentissage, où le modèle s'ajuste trop aux données d'entraînement et ne se généralise pas bien aux données de validation. Ainsi, nous menons des expériences avec différentes valeurs pour le nombre d'époques afin de trouver celui qui permet un équilibre optimal entre l'apprentissage des données d'entraînement et la capacité de généralisation aux nouvelles données. La représentation graphique des résultats est affichée sous forme de courbe dans la figure (3.7).

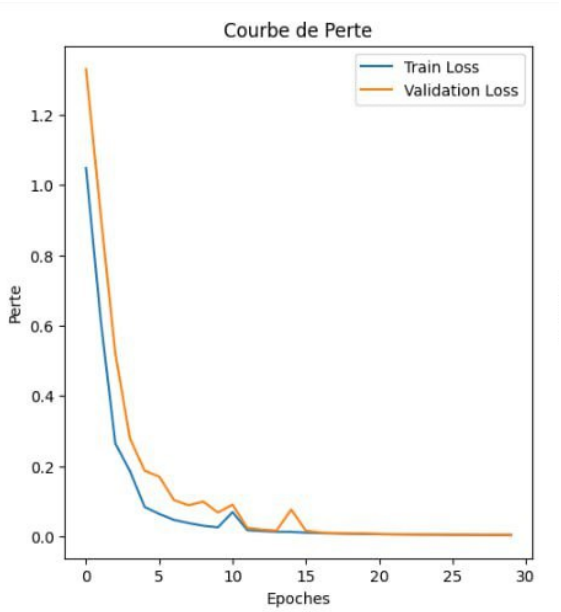

FIGURE 3.7 – Courbes de perte

Dans la figure (3.7), on peut observer les courbes de perte pour l'entraînement et la validation de notre modèle en fonction du nombre d'époques. La courbe de perte montre une diminution rapide de la perte, que ce soit pour l'entraînement ou la validation. Après environ 15 époques, les valeurs atteignent des niveaux proches de zéro, indiquant une convergence rapide et efficace du modèle. Ces courbes indiquent que le modèle apprend bien les schémas des données d'entraînement et généralise efficacement aux données de validation.

3.3.4 Evaluation des performances

Le tableau (3.5) présente les performances en termes d'accuracy de notre modèle sur la base de données CK+. Les résultats du tableau confirment bien les constats des expérimentations précédentes en ce qui concerne la bonne généralisation sur les données de test jamais vues.

TABLE 3.5 – Accuracy de notre modèle sur la base de données CK+.

Images des visages	Train Accuracy	Test Accuracy	Validation Acuracy
Taille de la base	80%	10%	10%
Accuracy	**100%**	**100%**	**100%**

Le tableau (3.5) et la matrice de confusion (3.8) illustrent la performance exceptionnelle de notre modèle sur la base de données CK+. Avec une accuracy de 100% sur les données

d'entraînement, de test et de validation, le modèle montre une parfaite généralisation. La matrice de confusion confirme cette performance en démontrant une classification sans erreur pour toutes les catégories d'émotions, chaque case hors diagonale étant à zéro. Chaque étiquette de vérité (true label) correspond parfaitement à l'étiquette prédite, soulignant une précision et un F1 score parfaits de 1.0000. Ces résultats confirment que notre modèle est capable de distinguer les émotions faciales avec une fiabilité absolue sur des données jamais vues, attestant de son efficacité et de sa robustesse.

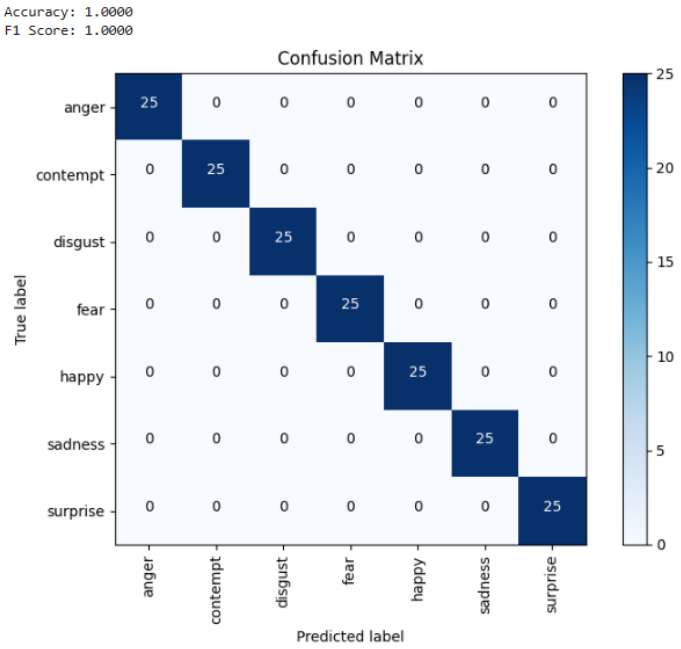

FIGURE 3.8 – Matrice de confusion

Dans le but de généraliser nos conclusions, en plus de la base de données CK+, nous avons utilisé l'ensemble de données OAHEGA dans notre étude. Les résultats obtenus sont repportés dans le tableau (3.6).

TABLE 3.6 – Performances de notre modèle sur OAHEGA et CK+.

Base de données	Train Accuracy	validation Accuracy	F1-score	Loss
OAHEGA	**91.43%**	**90.82%**	**90.81%**	**29.39%**
CK+	**100%**	**100%**	**100%**	**0.42%**

Les résultats du tableau (3.6) montrent que, bien que le modèle atteigne une précision parfaite de 100% sur CK+, ses performances sur OAHEGA sont légèrement inférieures, avec une accuracy d'entraînement de 91,43% et une accura de validation de 90.82%. Le F1-score

et la perte suivent la même tendance, avec des valeurs de 90.81% et 29.39% respectivement pour OAHEGA, contre 100% et 0.42% pour CK+. Ces différences indiquent que le modèle généralise mieux sur CK+ que sur OAHEGA.

3.3.5 Comparaison des résultats avec des travaux de la littérature

Dans ce qui suit, nous examinerons les performances de notre système, en termes d'accuracy, en le comparant avec des travaux existants dans le domaine de la reconnaissance des émotions sur les deux bases de données CK+ et OAHEGA.

Dans le cadre de l'évaluation de notre système par rapport à d'autres études, nous avons comparé son accuracy avec celui de travaux récents effectués sur les deux bases de données CK+ et OAHEGA. Les résultats de ces comparaison sont présentés dans les deux tableaux (3.7) et (3.8).

3.3.5.1 Sur la base de données CK+

Pour évaluer les performances de notre système par rapport à d'autres études existantes, nous avons évalué son efficacité avec certains travaux récents effectués sur la base de données CK+. Les résultats de cette comparaison sont rapportés dans le tableau (3.7).

TABLE 3.7 – Comparaison des performances avec l'état de l'art (CK+).

Système	Technique	Accuracy
Qin et al. 2020 [119]	CNN	96.81%
Mehendale .2020 [120]	CNN	96%
Aouayeb et al. 2021 [124]	Vision Transformer	99.80%
Helaly et al. 2023 [130]	ResNet-18	98%
Chowdary et al. 2023 [132]	Resnet50	96%
Ahmad et al. 2024 [139]	ShuffleNet V2	98%
Bellamkonda et Settipalli. 2024 [140]	EFL-LCNN	99.20%
Notre Système, 2024	**ViT**	**100%**

Pour faciliter la comparaison, le tableau (3.7) est transcrit en graphique illustré sur la figure Comme le montre la figure (3.9).

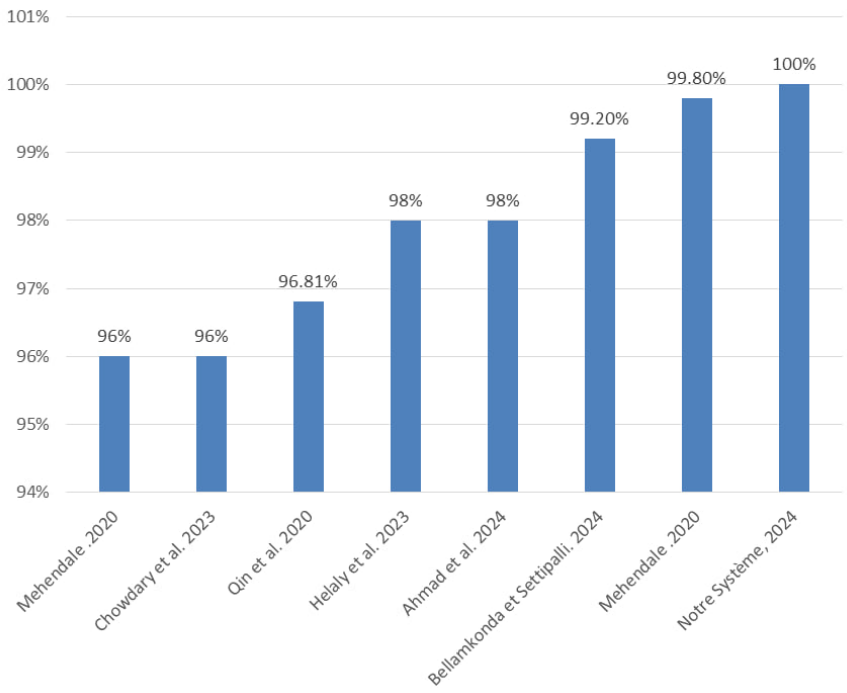

FIGURE 3.9 – Comparison en termes d'accuracy avec l'état de l'art sur CK+

3.3.5.2 Sur la base de données OAHEGA

Après avoir comparé notre système avec des autres études existantes qui utilisaient la base de données Ck +, nous le comparons maintenant avec certain travaux récents qui utilisaient la base de données OAHEGA.

TABLE 3.8 – Comparaison des performances avec l'état de l'art (OAHEGA).

Système	Technique	Accuracy
Agrawal et al. 2023 [129]	AlexNet	54%
Bakariya et al. 2024 [141]	CNN	73.02%
Notre Système, 2024	**ViT**	**90.82%**

Nos résultats montrent une nette amélioration grâce à notre utilisation du modèle ViT par rapport aux études précédentes qui utilisaient AlexNet et CNN sur la base de données OAHEGA. Agrawal et al. (2023) ont obtenu 54% d'accuracy avec AlexNet, Bakariya et al. (2024) ont atteint 73.02% avec CNN, alors que notre système atteint une accuracy de 90.82%. Il est également pertinent de noter que la base de données OAHEGA est moins utilisée dans la littérature par rapport à d'autres jeux de données plus couramment exploités, malgré les performances prometteuses que nous avons obtenues avec ViT . Cette constatation souligne

le potentiel inexploité d'OAHEGA et encourage à explorer d'avantage cette base de données pour de futures études et applications dans le domaine de la reconnaissance et de l'analyse d'émotions.

3.4 Conclusion

Dans ce chapitre, nous avons montré comment nous avons réaliser notre modèle Vision Transformer (ViT) pour classifier les expressions faciales dans les images. Nous avons également étudié comment différents paramètres affectent l'accuracy du modèle. Pour ce faire, nous avons réalisé plusieurs expériences pour trouver les meilleurs réglages pour notre modèle ViT. Nous l'avons testé sur les bases de données CK+ et OAHEGA, et l'avons comparé à des travaux récents. Les résultats de nos expériences ont prouvé que notre système, basé sur les Transformers, est efficace pour classifier les images d'expressions faciales, ouvrant ainsi la voie à une détection plus précise et fiable dans ce domaine.

CONCLUSION GÉNÉRALE

Bilan

Les systèmes de reconnaissance des émotions à partir des expressions faciales dans les images jouent un rôle important dans divers domaines tels que la santé mentale, l'interaction homme-machine et la sécurité. Leur utilisation permet de détecter et d'interpréter les émotions humaines avec précision, ce qui peut aider à améliorer les interactions sociales, à fournir un soutien émotionnel personnalisé et à renforcer la sécurité dans les environnements publics. En fournissant des informations précises sur les états émotionnels, ces systèmes facilitent la compréhension des besoins des individus et la prise de décisions appropriées pour leur bien-être.

Notre travail se concentre sur la classification des expressions faciales dans les images en utilisant la base de données CK+ et OAHEGA. Nous cherchons à améliorer les performances des systèmes de classification des expressions faciales en exploitant les avancées dans le domaine de l'apprentissage automatique. Bien que de nombreux travaux aient été entrepris pour aborder ce problème, la recherche actuelle n'a pas encore atteint le niveau de précision comparable à celui de la capacité humaine dans ce domaine. Notre objectif est donc d'explorer de nouvelles approches et techniques pour améliorer la reconnaissance des émotions faciales et d'atteindre des performances plus proches des capacités humaines.

Dans ce contexte, notre attention s'est portée sur la classification des émotions à partir des expressions faciales en utilisant les ViTs (Vision Transformers), une technique relativement nouvelle en intelligence artificielle qui a apporté des avancées significatives dans le domaine

de la vision par ordinateur.

Pour évaluer les performances du modèle ViT dans notre étude, nous avons réalisé des essais pour ajuster les hyperparamétre afin d'obtenir la meilleure précision (accuracy) possible. Les résultats ont montré que nous avons pu mettre en place un système hautement performant, a atteint un taux de reconnaissance de 100 % pour la base de données CK+ et 90% pour OAHEGA, démontrant ainsi l'efficacité du modèle ViT dans la détection des émotions à partir des expressions faciales sur les images.

Pour évaluer le système basé sur ViT que nous avons proposé, nous l'avons comparé en termes d'accuracy avec des recherches récentes utilisant également CK+ et OAHEGA. Cette comparaison nous a permis de constater que notre système se classe parmi les meilleurs, ce qui confirme l'efficacité de ViT dans la classification des émotions à partir des expressions faciales.

Perspectives

En ce qui concerne nos perspectives d'avenir, plusieurs axes d'amélioration sont envisageables pour poursuivre nos travaux. Les plus importants sont les suivants :

— Nous prévoyons d'améliorer la généralisation du modèle en utilisant des techniques de data augmentation spécialement conçues pour les expressions faciales. Cela devrait aider le modèle à mieux capturer la diversité des expressions et à renforcer ses performances sur des données nouvelles ou moins représentées.

— Non seulement avec ViT, mais il existe d'autres modèles que nous pourrions explorer, tels que ImageGPT et CLIP, qui sont basés sur les Transformers. Ces modèles offrent de nouvelles perspectives en termes de représentations et de capacités de classification, ouvrant ainsi la voie à des améliorations significatives de la précision du modèle dans la reconnaissance des émotions.

— Concevoir des modèles qui peuvent identifier avec précision les différentes nuances des émotions faciales. En distinguant les traits spécifiques de chaque type démotion, nous pourrions améliorer considérablement la précision des classifications. Cela aurait des implications significatives dans des domaines comme l'évaluation de la santé mentale et l'amélioration des interactions entre l'homme et la machine.

BIBLIOGRAPHIE

[1] D. Pastre, "L'intelligence artificielle definition-generalites-historique-domaines," 2000. Intelligence artificielle, Université Paris, 5.

[2] A. L. Samuel, "Machine learning," *The Technology Review*, vol. 62, no. 1, pp. 42–45, 1959.

[3] J. Hu *et al.*, "Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14413–14423, 2020.

[4] Clevy, "Le machine learning décrypté 3/3 : Types dapprentissage et limites," *Clevy Blog*, 2019.

[5] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR) [Internet]*, vol. 9, pp. 381–386, 2020.

[6] S. I. Weiss and C. A. Kulikowski, *Computer Systems That Learn : Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*. San Francisco : Morgan Kaufmann, 1991.

[7] E. M. Zemmouri, "Titre de l'article," *Nom de la revue*, vol. volume, no. numéro, p. pages, année.

[8] F. Emmert-Streib, O. Yli-Harja, and M. Dehmer, "A clarification of misconceptions, myths and desired status of artificial intelligence," *arXiv preprint arXiv :2008.05607*, 2020.

[9] N. Gupta, "A guide to supervised learning," *Medium*, 2024.

[10] M. Kozan, "Supervised and unsupervised learning (an intuitive approach)," *Medium*, 2021.

[11] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. Cambridge : MIT Press, 1998.

[12] Projeduc, "Introduction à l'apprentissage automatique," *GitHub*, 2024.

[13] M. M. Adankon and M. Cheriet, "Model selection for the lssvm. application to handwriting recognition," *Pattern Recognition*, vol. 42, no. 12, pp. 3264–3270, 2009.

[14] J. Chauvin, R. Duran, S. Ng, T. Burke, K. Barton, N. MacKinnon, K. Tavakolian, A. Akhbardeh, and F. Vasefi, *Advanced Optical Technologies in Food Quality and Waste Management*. IntechOpen, 2021.

[15] E. Mathieu-Dupas, "Algorithme des k plus proches voisins pondérés et application en diagnostic," in *42èmes Journées de Statistique*, 2010.

[16] W.-B. A. B. Zoungrana, *Application des algorithmes dapprentissage automatique pour la détection de défauts de roulements sur les machines tournantes dans le cadre de lIndustrie 4.0*. PhD thesis, Université du Québec à Chicoutimi, 2020.

[17] AquaPortail, "Arbre de décision," *AquaPortail*, 2009.

[18] A. Moghaddamnia *et al.*, "Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques," *Advances in Water Resources*, vol. 32, no. 1, pp. 88–97, 2009.

[19] Y. Zong, Y. Wu, Y. Luo, H. Xu, W. Hu, and Y. Yu, "Reips : A secure cloud-based reputation evaluation system for iot-enabled pumped storage power stations," *Sensors*, vol. 23, p. 5620, 2023.

[20] B. M. Franco Ortellado *et al.*, "Applications of artificial neural networks in three agro-environmental systems : microalgae production, nutritional characterization of soils and meteorological variables management," 2019.

[21] S. Sharma, S. Sharma, and A. Anidhya, "Fonctions d'activation dans les rÉseaux de neurones," *Revue internationale des sciences appliquées et de la technologie de l'ingénierie*, 2020.

[22] ResearchGate, "Most common activation functions," *ResearchGate*, 2023.

[23] T. Keldenich, "Fonction dactivation, comment ça marche ? une explication simple," in *Titre du livre où l'article est publié* (N. des éditeurs (s'il y a lieu), ed.), ch. Numéro du chapitre ou pages où l'article est trouvé, Ville de l'éditeur : Nom de l'éditeur, 2021. Consulté le 24 juin 2024.

[24] N. Buduma and N. Locascio, *Fundamentals of Deep Learning : Designing Next-Generation Machine Intelligence Algorithms*. Sebastopol, CA, USA : O'Reilly Media, 2017.

[25] J. Buss, "Activation function gelu in bert," 2023. Accessed : 2024-06-26.

[26] R. Vargas, A. Mosavi, and R. Ruiz, "Deep learning : A review." `https://doi.org/10.20944/preprints201810.0218.v1`, 2018. DOI : 10.20944/preprints201810.0218.v1.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[28] P. Hairy, *Les réseaux de neurones récurrents pour les séries temporelles*. Ville de l'éditeur (si disponible) : Nom de l'éditeur, 2021.

[29] J. Brownlee, "A gentle introduction to long short-term memory networks by the experts," *Machine Learning Mastery*, vol. 19, 2017.

[30] ResearchGate, "Structure of simple recurrent neural network (rnn) and unfolded rnn." `https://www.researchgate.net/figure/Structure-of-simple-recurrent-neural-network-RNN-and-unfolded-RNN_fig8_337268343`. Online ; accessed 27-juin-2024.

[31] Y. Morere, *Les Réseaux de Neurones Récurrents*. Ville de l'éditeur (si disponible) : Nom de l'éditeur, 2021.

[32] P. Srivastava, *Essentials of Deep Learning : Introduction to Long Short Term Memory*. Dec. 2017.

[33] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] F. Kibrete, T. Trzepieciski, H. S. Gebremedhen, and D. E. Woldemichael, "Artificial intelligence in predicting mechanical properties of composite materials," *J. Compos. Sci.*, vol. 7, no. 9, p. 364, 2023.

[35] V. Lendave, "LSTM Vs GRU in Recurrent Neural Network : A Comparative Study," 2021.

[36] W. Ouyang, P. Luo, X. Zeng, and et al., "Deepid-net : Multi-stage and deformable deep convolutional neural networks for object detection," in *Proceedings of the CVPR*, 2015.

[37] A. Khan *et al.*, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, pp. 5455–5516, 2020.

[38] Y. S. Obam, "Comprendre les réseaux de neurones convolutifs (cnn)," *Medium*, 2019.

[39] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding : A review," *Neurocomputing : An International Journal*, vol. 187, pp. 27–48, 2016.

[40] M. D. Zeiler, *Hierarchical Convolutional Deep Learning in Computer Vision.* PhD thesis, New York University, 2013.

[41] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence*, pp. 1237–1242, June 2011.

[42] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 12, pp. 211–252, 2015.

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv :1409.1556*, 2014.

[46] C. Szegedy and et al., "Going deeper with convolutions," *s.l.*, vol. s.n., pp. 1–9, 2015.

[47] K. He and et al., "Deep residual learning for image recognition," *CVPR*, 2016. arXiv preprint arXiv :1512.03385.

[48] V. Perlibakas, "Face recognition using principal component analysis and loggabor filters," *Expert Systems with Applications*, vol. 89, pp. 129–137, 2006.

[49] D. Nguyen, "Multimodal emotion recognition using deep learning techniques," Master's thesis, Queensland University of Technology, 2020.

[50] A. Vaswani and et al., "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000–6010, 2017.

[52] P. Li, J. Zheng, P. Li, H. Long, M. Li, and L. Gao, "Tomato maturity detection and counting model based on mhsa-yolov8," *Sensors*, vol. 23, no. 15, p. 6701, 2023.

[53] A. Radford and et al., "Improving language understanding by generative pre-training," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[54] Auteur(s), "Attention mechanism : Transformers, bert, and gpt tutorial and survey," année.

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT : Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv :1810.04805*, 2018.

[56] M. Chen and et al., "Generative pretraining from pixels," in *International Conference on Machine Learning*, pp. 1691–1703, PMLR, 2020.

[57] A. si disponible, "Bert vs gpt : Differences real-life examples." `https://vitalflux.com/bert-vs-gpt-differences-real-life-examples/`, année de publication si disponible.

[58] M. Usman, T. Zia, and A. Tariq, "Analyzing transfer learning of vision transformers for interpreting chest radiography," *Journal of Digital Imaging*, vol. 35, no. 6, pp. 1445–1462, 2022.

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and N. Houlsby, "An image is

worth 16x16 words : Transformers for image recognition at scale," *arXiv preprint arXiv :2010.11929*, 2020.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words : Transformers for image recognition at scale," *arXiv preprint arXiv :2010.11929*, 2020.

[61] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, pp. 10347–10357, 2021.

[62] F. Shah, J. Smith, J. Doe, *et al.*, "On reading," *Journal of Education*, vol. 12, no. 1, pp. 45–67, 2022.

[63] D. Nelson, "Qu'est-ce que l'apprentissage par transfert ?," *Unite.AI*, October 2020. Consulté le 25 juin 2024.

[64] ResearchGate, "Transfer learning flowchart." `https://www.researchgate.net/figure/Transfer-learning-flowchart_fig3_362881858`. Online ; accessed 27-juin-2024.

[65] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, pp. 2887–2905, 2021.

[66] La Philosophie, "Quest-ce quune émotion ?." `https://la-philosophie.com/emotion-philosophie#:~:text=Quest%2Dce%20qu,%2C%20une%20rupture%20d%C3%A9quilibre`. Accessed : Février 20, 2023.

[67] P. Claudon and M. Weber, "Lémotion : contribution à létude psychodynamique du développement de la pensée de lenfant sans langage en interaction," *Devenir*, vol. 21, no. 1, pp. 61–99, 2009.

[68] I. T. T. Meftah, *Modélisation, détection et annotation des états émotionnels à laide dun espace vectoriel multidimensionnel*. PhD thesis, Université Nice Sophia Antipolis, 2013.

[69] "Psychological science chapter 10 flashcards." Quizlet.

[70] S. Hazmoune and F. Bougamouza, "Using transformers for multimodal emotion recognition : Taxonomies and state of the art review," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108339, 2024.

[71] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[72] P. Ekman, "Facial expressions of emotion : New findings (new questions)," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[73] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[74] P. Ekman, "Basic emotions," in *Handbook of cognition and emotion*, pp. 45–60, John Wiley & Sons, 1999.

[75] P. Ekman and R. Davidson, *The Nature of Emotion : Fundamental Questions*. Oxford University Press, 1994.

[76] Pinterest, "Pin on *Interiors*." `https://www.pinterest.com/pin/164170348899449076/`. Online ; accessed 27-juin-2024.

[77] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[78] M. M. Bradley and P. J. Lang, "Measuring emotion : The self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[79] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Emotion : Theory, Research, and Experience*, pp. 3–33, Academic Press, 1980.

[80] Dino's English, "La rueda de las emociones de robert plutchik." `https://dinosenglish.edu.vn/la-rueda-de-las-emociones-de-robert-plutchik-169040189804305` Online ; accessed 27-juin-2024.

[81] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. The MIT Press, 1974.

[82] ResearchGate, "Le modèle démotion plaisir-excitation-dominance (pad).." `https://www.researchgate.net/figure/`

Le-modele-demotion-plaisir-excitation-dominance-PAD_fig1_334097316. Online ; accessed 27-juin-2024.

[83] R. S. Lazarus, *Emotion and Adaptation*. Oxford University Press, 1991.

[84] H. Kim, *The Moderating Role of Service Design Attributes in Females Fear of Crime in the Underground*. PhD thesis, Brunel University, 2015.

[85] M. B. Arnold, *Emotion and Personality*. 1960.

[86] L. F. Barrett, *How Emotions Are Made : the Secret Life of the Brain*. Pan Macmillan, 2017.

[87] L. F. Barrett, "The theory of constructed emotion : an active inference account of interoception and categorization," *Social Cognitive and Affective Neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[88] "Définir la vie spirituelle des élèves hors des institutions religieuses : un enjeu du service danimation spirituelle et dengagement communautaire (québec)." lien_vers_l_article_sur_openedition.org, année. Article disponible sur OpenEdition.org.

[89] K. R. Scherer, "On the nature and function of emotion : a component process approach," in *Approaches to Emotion*, vol. 2293, p. 31, 1984.

[90] Auteur(s), "Processus procédure mode opératoire instruction formulaire." lien_vers_la_présentation_sur_slideplayer.fr, année. Présentation disponible sur SlidePlayer.fr.

[91] C. E. Izard, "Differential emotions theory," in *Human Emotions*, pp. 43–66, 1977.

[92] C. E. Izard, "Les quatre systèmes dactivation des émotions." Télécharger le diagramme scientifique sur ResearchGate, 1993. URL : https://www.researchgate.net/your-link-here.

[93] M. BELHIRECHE, "Nouvelle approche de reconnaissance multimodale des émotions," Master's thesis, Université 8 mai 1945 - Guelma, Département de l'Informatique, Faculté de mathématiques et de l'informatique et des sciences de la matière, 2023.

[94] D. F. K. Gom-os and K. Y. Yong, "An empirical study on the use of a facial emotion recognition system in guidance counseling utilizing the technology acceptance model and the general comfort questionnaire," *Applied Computing and Informatics*, 2022. Ahead-of-print.

[95] C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin, "Emotion recognition from text using semantic labels and separable mixture models," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, no. 2, pp. 165–183, 2006.

[96] F. Almeida and G. Xexéo, "Word embeddings : A survey," *arXiv preprint arXiv :1901.09069*, 2019.

[97] P. P. Tech, "Challenges faced by facial recognition systems." `https://www.pathpartnertech.com/challenges-faced-by-facial-recognition-system/`, 2023. Accessed on February 20.

[98] F. Noroozi, C. A. Corneanu, D. Kamiska, T. Sapiski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 505–523, 2018.

[99] Pinterest, "Pin on *Interiors*." `https://www.pinterest.com/pin/164170348899449076/`. Online ; accessed 27-juin-2024.

[100] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech : a review," *International Journal of Speech Technology*, vol. 15, pp. 99–117, 2012.

[101] N. N. Khatri, Z. H. Shah, and S. A. Patel, "Facial expression recognition : A survey," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 5, no. 1, pp. 149–152, 2014.

[102] Advanced Source Code, "Face recognition age estimation." `http://www.advancedsourcecode.com/facerecage.asp`. Online ; accessed 27-juin-2024.

[103] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech : a review," *International Journal of Speech Technology*, vol. 15, pp. 99–117, 2012.

[104] D. Mitrovi, M. Zeppelzauer, and C. Breiteneder, "Features for content-based audio retrieval," in *Advances in Computers*, vol. 78, pp. 71–150, Elsevier, 2010.

[105] D. Oude Bos *et al.*, "Eeg-based emotion recognition : The influence of visual and auditory stimuli," *Journal Name*, vol. 56, no. 3, pp. 1–17, 2006.

[106] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition : Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.

[107] C. Padgett and G. W. Cottrell, "Titre de l'article," *Nom du journal*, 1997.

[108] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Titre de l'article," *Nom du journal*, 1997.

[109] P. Monjaux, *Modélisation et animation interactive de visages virtuels de dessins animés*. PhD thesis, Nom de l'université, 2007.

[110] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 435–442, ACM, November 2015.

[111] P. Naga, S. D. Marri, and R. Borreo, "Facial emotion recognition methods, datasets and technologies : A literature survey," *Materials Today : Proceedings*, vol. 80, pp. 2824–2828, 2023.

[112] Davilsena, "Ck+ dataset." `https://www.kaggle.com/datasets/davilsena/ckdataset/data`, 2023. Consulté le 26 juin 2024.

[113] Papers with Code, "Ferg (facial expression research group database)." `https://paperswithcode.com/dataset/ferg#:~:text=FERG%20%28Facial%20Expression%20Research%20Group%20Database%29%20Introduced%20by,containing%2055%2C769%20annotated%20face%20images%20of%20six%20characters`. Consulté le 26 juin 2024.

[114] S. K. Eng, H. Ali, A. Y. Cheah, and Y. F. Chong, "Facial expression recognition in jaffe and kdef datasets using histogram of oriented gradients and support vector machine," in *IOP Conference Series : Materials Science and Engineering*, vol. 705, p. 012031, IOP Publishing, November 2019.

[115] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[116] V. Kovenko and V. Shevchuk, "OAHEGA : Emotion Recognition Dataset," 2021. Consulté le 26 juin 2024.

[117] V. Kovenko and V. Shevchuk, "OAHEGA : Emotion Recognition Dataset," 2021. Consulté le 26 juin 2024.

[118] S. Minaee and A. Abdolrashidi, "Deep-emotion : Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, 2019.

[119] S. Qin, Z. Zhu, Y. Zou, and X. Wang, "Facial expression recognition based on gabor wavelet transform and 2-channel cnn," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 18, no. 02, p. 2050003, 2020.

[120] N. Mehendale, "Facial emotion recognition using convolutional neural networks (ferc)," *SN Applied Sciences*, vol. 2, no. 3, p. 446, 2020.

[121] A. Tomar, S. Gera, S. Kumar, and B. Pant, "HHFER : A hybrid framework for human facial expression recognition," in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 537–541, 2021.

[122] Z. Chen and Y. Khaireddin, "Facial emotion recognition : State of the art performance on fer2013," *arXiv preprint arXiv :2105.03588*, 2021.

[123] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep cnn," *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[124] M. Aouayeb, W. Hamidouche, C. Soladié, K. Kpalma, and R. Séguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," *arXiv*, 2021.

[125] V. Sati, S. M. Sánchez, N. Shoeibi, A. Arora, and J. M. Corchado, "Face detection and recognition, face emotion recognition through nvidia jetson nano," in *Ambient Intelligence–Software and Applications : 11th International Symposium on Ambient Intelligence*, pp. 177–185, Springer International Publishing, 2021.

[126] M. Rahul, N. Tiwari, R. Shukla, D. Tyagi, and V. Yadav, "A new hybrid approach for efficient emotion recognition using deep learning," *International Journal of Electrical and Electronics Research*, 2022.

[127] P. Sarma, T. U. Laskar, D. G. V, and R. M., "Human emotion recognition using deep learning with special emphasis on infants face," *International Journal of Electrical and Electronics Research*, 2022.

[128] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "Vitfer : Facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, 2022.

[129] G. Agrawal, U. Jha, and R. Bidwe, "Automatic facial expression recognition using advanced transfer learning," in *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, pp. 450–458, August 2023.

[130] R. Helaly, S. Messaoud, S. Bouaafia, M. Hajjaji, and A. Mtibaa, "Dtl-i-resnet18 : Facial emotion recognition based on deep transfer learning and improved resnet18," *Signal, Image and Video Processing*, vol. 17, pp. 2731–2744, 2023.

[131] M. Alonazi, H. J. Alshahrani, F. A. Alotaibi, M. Maray, M. Alghamdi, and A. Sayed, "Automated facial emotion recognition using the pelican optimization algorithm with a deep convolutional neural network," *Electronics*, vol. 12, no. 22, p. 4608, 2023.

[132] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Computing and Applications*, vol. 35, no. 32, pp. 23311–23328, 2023.

[133] M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people," *Sensors*, vol. 23, no. 3, p. 1080, 2023.

[134] C. Biaek, A. Matiolaski, and M. Grega, "An efficient approach to face emotion recognition with convolutional neural networks," *Electronics*, vol. 12, no. 12, p. 2707, 2023.

[135] B. R. Prasad and B. S. Chandana, "Human face emotions recognition from thermal images using densenet," *International Journal of Electrical and Computer Engineering Systems*, vol. 14, no. 2, pp. 155–167, 2023.

[136] B. Assiri and M. A. Hossain, "Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism," *Mathematical Biosciences and Engineering*, vol. 20, no. 1, pp. 913–929, 2023.

[137] G. JS, D. Gleran Lobo, and M. Aman, "Reading faces, recommending choices : A systematic review of facial emotion recognition and recommendation systems," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1–12, 2024.

[138] F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, "Modèle de reconnaissance des émotions faciales en temps réel basé sur l'auto-encodeur du noyau et le réseau neuronal convolutif pour les enfants autistes," *Informatique douce*, pp. 1–14, 2024.

[139] F. Ahmad, U. Hariharan, N. Muthukumaran, A. Ali, and S. Sharma, "Emotion recognition of the driver based on klt algorithm and shufflenet v2," *Signal, Image and Video Processing*, pp. 1–18, 2024.

[140] S. Bellamkonda and L. Settipalli, "Efl-lcnn : Enhanced face localization augmented light convolutional neural network for human emotion recognition," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 12089–12110, 2024.

[141] B. Bakariya, A. Singh, H. Singh, P. Raju, R. Rajpoot, and K. K. Mohbey, "Facial emotion recognition and music recommendation system using cnn-based deep learning techniques," *Evolving Systems*, vol. 15, no. 2, pp. 641–658, 2024.

[142] "Implementing a transformer encoder from scratch with jax and haiku," *Towards Data Science*, 2024.

[143] GeeksforGeeks, "In the context of deep learning, what is training warmup steps ?." `https://www.geeksforgeeks.org/in-the-context-of-deep-learning-what-is-training-warmup-steps/`, n.d. Accessed : 2024-06-29.

ANNEXE A

OUTILS D'IMPLÉMENTATIONS

A.1 Introduction

Dans cette annexe, nous commençons par présenter les outils de développement utilisés, puis nous expliquons les différentes étapes de la réalisation du système de classification des émotions.

A.2 Environnement de développement

Dans le but d'obtenir des résultats concrets et adaptables dans notre recherche, nous avons choisi d'utiliser Google Colab, également connu sous le nom de Colab. C'est une plateforme cloud fournie par Google, qui permet de créer un environnement de développement similaire à celui de Jupyter Notebook dans le cloud. Grâce à Colab, de nombreux utilisateurs ont accès à des processeurs graphiques (GPUs) puissants, ce qui est particulièrement avantageux pour ceux qui ne disposent pas des ressources nécessaires pour exécuter des projets de machine learning sur leur propre matériel. En utilisant ce service, nous avons pu développer et exécuter nos modèles de manière efficace, tout en tirant parti des avantages offerts par les GPUs pour accélérer les calculs et améliorer les performances de nos experimentations.

A.3 Langage python

Python est un langage de programmation dynamique, ce qui signifie qu'il est possible d'ajuster le type d'une variable. Le langage interprété est un langage open source. Il permet aux programmeurs de se concentrer sur les actions qu'ils effectuent plutôt que sur les différentes techniques pour les realizer, ce qui leur permet de gagner un temps considérable par rapport aux langages compilés. Cet outil est largement employé, très apprécié et très recherché dans le domaine de l'informatique. Les commentaires servent à préciser les diverses catégories d'arguments ainsi que les valeurs de retour des fonctions. Python est fréquemment employé dans les domaines du développement web, de l'analyse de données, de l'intelligence artificielle, des réseaux de neurones, du calcul scientifique et d'autres domaines de l'informatique avancée. Ce langage est certainement le plus utilisé dans le domaine de la programmation, adapté aussi bien aux projets simples qu'aux projets complexes.

A.4 Les bibliothèques utilisées

Dans notre travail, nous avons employé diverses bibliothèques pour la mise en oeuvre des modèles de classification des émotions à partir des expressions facials, basés sur les Transformers. Ces bibliothèques ont joué un rôle essentiel dans la manipulation des données, le développement des modèles, et l'évaluation des résultats. Elles ont fourni des outils et des fonctionnalités permettant de traiter efficacement les images, d'entraîner les modèles, et d'analyser les performances de nos systèmes. En combinant ces bibliothèques avec nos programme , nous avons pu créer des solutions robustes et précises pour la reconnaissance des émotions dans les images facials.

A.4.1 Pandas

Pandas est une extension Python spécialement créée pour l'analyse de données organisées sous forme de tableaux, appelés DataFrames. Cette bibliothèque a été développée pour répondre au besoin d'un outil puissant et flexible pour l'analyse quantitative, ce qui en fait l'une des bibliothèques Python les plus populaires dans ce domaine. Notamment des fichiers CSV, grâce à des fonctions telles que read-csv, qui permet de charger des données depuis

un fichier CSV, et head, qui affiche les premières lignes d'un DataFrame pour une inspection rapide. Son attrait réside également dans sa communauté active de contributeurs qui travaillent régulièrement à son développement et à son amélioration.

A.4.2 Numpy

NumPy, raccourci de "Numerical Python", est une bibliothéque essentielle pour les scientifiques et les programmeurs qui utilisant Python. Elle nous permet de manipuler facilement les données numériques, comme celles que l'on trouve dans les images. NumPy fusionne les points forts du langage C avec ceux de Python. Dans notre cas, nous l'utilisons pour extraire et manipuler les données numériques des expressions faciales dans les images, car il offre une multitude de fonctionnalités pour effectuer des opérations complexes sur ces données.

A.4.3 Torch

Torch est un framework de machine learning open-source qui offre une large gamme d'outils et de bibliothèques pour le développement de modèles d'apprentissage automatique qui privilégie les GPU. Son objectif principal est de fournir aux développeurs et aux chercheurs une plateforme flexible et efficace pour construire, former et déployer des modèles d'apprentissage automatique. Connu pour sa prise en charge native des GPU, Torch permet d'accélérer les calculs, ce qui le rend particulièrement adapté pour traiter de grands ensembles de données et des modèles complexes. En outre, Torch est accompagné d'un vaste écosystème de packages développés par la communauté, couvrant divers domaines de l'apprentissage automatique tel que le traitement du signal, le traitement paralléle et la vision par ordinateur , ce qui enrichit encore ses fonctionnalités et sa polyvalence.

A.4.4 Matplotlib

Matplotlib est une bibliothèque de visualisation de données très répandue et appréciée dans le domaine de la programmation Python. Elle offre une large gamme d'outils pour créer différents types de graphiques, tels que les graphiques linéaires, les nuages de points, les histogrammes, les graphiques à barres et les graphiques circulaires. Sa compatibilité avec les structures de données de NumPy en fait un choix idéal pour les utilisateurs qui travaillent

avec des données numériques. En utilisant le module 'matplotlib.pyplot', les développeurs peuvent créer des figures et personnaliser divers aspects de leurs graphiques, comme les couleurs, les légendes et les étiquettes. La fonction figure() est couramment utilisée pour créer une nouvelle figure sur laquelle les graphiques peuvent être dessinés. Matplotlib est un outil puissant qui permet aux utilisateurs de visualiser leurs données de manière efficace, offrant ainsi un support essentiel pour l'exploration et l'analyse de données dans Python.

A.4.5 Itertools

Une bibliothèque standard de Python qui propose une gamme d'outils permettant de travailler avec des itérateurs de manière efficace. Elle offre des fonctionnalités pour créer et manipuler des itérables, ainsi que des itérateurs spécifiques pour générer des combinaisons, des permutations et d'autres structures itératives de manière optimisée. L'une de ses fonctions les plus couramment utilisées est product, qui génère le produit cartésien de plusieurs ensembles d'entrées. Cette fonction est utile pour générer toutes les combinaisons possibles d'éléments provenant de ces ensembles.

A.4.6 Imblearn ou (Imbalanced-learn)

Est une bibliothèque Python qui offre des outils et des techniques pour gérer les données déséquilibrées dans le domaine de l'apprentissage automatique. Elle propose des méthodes de suréchantillonnage, de sous-échantillonnage et des approches combinées pour équilibrer les ensembles de données où certaines classes sont significativement sous-représentées. L'utilisation de cette bibliothèque permet d'améliorer les performances des modèles de classification en traitant efficacement les problèmes de déséquilibre des classes, ce qui est crucial pour obtenir des prédictions plus précises et fiables sur des ensembles de données déséquilibrés. Une des fonctions utilisées pour atteindre cet objectif est le :

RandomOverSampler : Cette fonction effectue un suréchantillonnage aléatoire des classes minoritaires en répétant de manière aléatoire des échantillons, afin d'équilibrer la distribution des classes dans l'ensemble de données.

A.4.7 Torchvision

Est une bibliothèque qui fournit des transformations courantes pour le traitement d'images et l'entraînement de modèles de vision par ordinateur. Cette bibliothèque est souvent utilisée en conjonction avec PyTorch pour faciliter le chargement, la préparation et la manipulation d'images dans le contexte de l'apprentissage automatique. Elle offre également un accès à des ensembles de données standardisés ainsi qu'à des modèles de réseaux de neurones pré-entraînés pour des tâches spécifiques de vision par ordinateur.

A.4.8 Tqdm

Cette bibliothèque fournit une interface utilisateur conviviale pour visualiser la progression des boucles et des opérations longues dans les scripts Python. Elle affiche des barres de progression animées ou des pourcentages en temps réel, ce qui permet aux utilisateurs de suivre facilement l'avancement des tâches en cours. Cette fonctionnalité est particulièrement utile lors de l'itération sur de grandes quantités de données ou lors de l'exécution de processus qui peuvent prendre du temps. En utilisant tqdm, les développeurs peuvent rendre leurs scripts plus interactifs et offrir une meilleure expérience utilisateur lors de l'exécution de tâches longues ou intensives en calcul.

A.4.9 Os

Cette bibliothèque en Python fournit des fonctions pour interagir avec le système d'exploitation (d'où le nom "os" pour "Operating System" en anglais). Elle permet d'accéder à des fonctionnalités telles que la gestion des répertoires et des fichiers, la manipulation des chemins de fichiers, et l'exécution de commandes système. Elle offre une interface portable pour effectuer des opérations liées au système d'exploitation, ce qui la rend très utile pour le développement d'applications et de scripts Python multiplateformes.

A.4.10 Collections

Collections est une bibliothèque Python qui implémente des types de données spécialisés comme les chaînes ordonnées, les dictionnaires à valeurs par défaut, et les compteurs. Ces structures de données avancées offrent des fonctionnalités supplémentaires par rapport aux

types de données standard de Python, permettant de gérer plus efficacement des collections de données complexes. Counter est une fonction qui permet de compter les occurrences d'éléments dans une collection, facilitant ainsi le comptage rapide et facile des éléments dans une séquence ou un itérable.

A.5 Implémentaation

Dans se qui suit nous présentons les étapes principales d'implémentation

A.5.1 Importation des bibliothèques

Dabord, nous importerons tous les modules nécessaires pour entraîner notre modèle. La figure (A.1) présente un bout de code pour importer les bibliothèques nécessaires.

```
import gc
import numpy as np
import pandas as pd
import itertools
from collections import Counter
import matplotlib.pyplot as plt
```

FIGURE A.1 – Les nécessaires bibliothèques

A.5.2 Prétraitement des données

Avant d'intégrer les données d'image dans le modèle ViT, nous devons les prétraiter et effectuer certaines opérations pour les preparer.

```
normalize = Normalize(mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5])
class GrayscaleToRGB:
    def __call__(self, img):
        return img.convert("RGB")

_train_transforms = Compose(
    [
        GrayscaleToRGB(),
        Resize((size, size)),
        RandomRotation(90),
        RandomAdjustSharpness(2),
        RandomHorizontalFlip(0.5),
        ToTensor(),
        normalize
    ]
)
```

FIGURE A.2 – Les opération de prétraitement des données

A.5.3 La division des donnéees

Nous avons réparti cette base de données en trois sous ensembles distincts : un pour l'entraînement, lautre pour la validation et le dernier les tests (train/val/test). Cette répartition a été réalisée en utilisant le code présenté ci-dessous.

```
dataset = dataset.train_test_split(test_size=0.1,
                                   shuffle=True,
                                   stratify_by_column="label")
train_data = dataset['train']
test_data = dataset['test']
train_val_split = train_data.train_test_split(test_size=0.1,
                                              shuffle=True,
                                              stratify_by_column="label")
train_val_split = DatasetDict({
    'train': train_val_split['train'],
    'validation': train_val_split['test']})
train_data = train_val_split['train']
val_data = train_val_split['validation']
```

FIGURE A.3 – La division des données

A.5.4 Création de modèle vit

Nous avons utilisé le modèle pré-entraînés vit pour classifier les émotions à partir des expressions faciales dans les images, comme indiqué sur la figure ci-dessous.

```
#'google/vit-base-patch16-224-in21k'
model_str = 'dima806/face_emotions_image_detection'
processor = ViTImageProcessor.from_pretrained(model_str)
```

FIGURE A.4 – Chargement du modèle vit

A.5.5 Les arguments d'entraînement

Les arguments d'entraînement pour entraîner le modèle vit ont été définis dans ce qui suit (figure A.5), en spécifiant des paramètres comme le nombre d'époques et le pas d'apprentissage.

```
num_train_epochs = 20
args = TrainingArguments(
    output_dir=model_name,
    logging_strategy="steps",
    logging_steps=20,
    evaluation_strategy="epoch",
    learning_rate=0.00001,
    per_device_train_batch_size=32,
```

FIGURE A.5 – Les arguments d'entraînement

A.5.6 Apprentissage

Le code suivant réalise l'entraînement de notre modèle.

```
#Commencez l'entraînement du modèle
trainer.train()
```

FIGURE A.6 – Fonction d'entraînement

A.5.7 Test

Dans cette phase nous avons évalués les performances de notre modèle vit en utilisons la fonction 'trainer.predict()' et en analysant l'accuracy et la perte (loss).

```
#Utilisez 'trainer' pour faire des prédictions sur bdd test.
outputs = trainer.predict(test_data)
```

FIGURE A.7 – Évaluation des performances sur la base de test

A.6 Conclusion

Dans cette annexe, nous avons abordé les divers aspects du développement de notre système. Nous avons fourni des détails précis sur les outils utilisés et les étapes d'implémentation de notre système, en incluant quelques captures d'écran de notre code Python.