الجمـــهوريـــة الجـــزائريـــة الديـمقراطيـــــــة الشـــعبيـــــة
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعـلـيـم الـعـالـــي والـبـحـث الـعـلـمي
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
جـامـعـة عمّـــار ثـلـيـــجـــي بالأغـــواط
UNIVERSITE AMAR TELIDJI LAGHOUAT

كـلـيـــــة التكنولوجيا
FACULTE DE TECHNOLOGIE
قسم الإلكترونيك
Département d'électronique

Mémoire de Master

Domaine : SCIENCES ET TECHNOLOGIES

Filière : TELECOMMUNICATIONS

Option : RESEAUX ET TELECOMMUNICATIONS

Présenté par

CHATTA Isra

BENBRIKA Chourouk Nacira

THEME

Reconnaissance de la langue des signes algérienne basée sur le deep learning et MediaPipe

DEVANT LE JURY

Président : M Mourad REGGUIRE

Encadreur : M Mohammed BELKHEIRI

Examinateur : M Abdelkader ZITOUNI

Promotion 2022/2023

Dédicaces

Grace à l'aide d'ALLAH ; Le tout puissant ; Ce travail est achevé ;

Je le dédie …

À mon père, ma source de soutien infatigable et de motivation,

À la mémoire de ma mère, qui m'a inspirée chaque jour de ma vie, ta présence bienveillante me manque profondément. Tu as été une source d'inspiration constante pour moi chaque jour, je me suis efforcée d'honorer ta mémoire en persévérant dans mes études, en me rappelant les valeurs que tu m'as inculquées et en faisant preuve de compassion envers les autres. Ton absence physique est déchirante, mais ton esprit et ta bienveillance m'accompagnent toujours.

À mes chers frères et sœurs ; pour leur aide et leur soutien moral.

À la doctorante Imane TOUHAMI, mes sincères remerciements pour tes encouragements et ton aide

À toutes mes amies ; et à tous ceux qui me sont chers….

CHATTA Isra

Dédicaces

A l'aide de DIEU, le tout puissant je dédie ce travail

A mes parents, ma chère maman pour son amour infini, pour son soutien incorporable, pour sa compréhension qui n'a pas d'équivalent, avec mes sentiments d'amour et de respect les plus chaleureux.

Mon cher père, à qui je dois tant et tout, symbole du courage et de sacrifice, sa patience et son aide qui m'ont toujours encouragée et soutenue au cours de la période de mes études, je souhaite que ce travail soit un témoignage de ma profonde affection et reconnaissance du sacrifice de mon père.

À mes chers frères et sœurs ; pour leur aide et leur soutien moral.

À toute ma famille, à tous ceux que j'aime,

BENBRIKA Chourouk Nacira

Remerciements

Au nom d'Allah SWT, le Très Miséricordieux et le Plus Miséricordieux.

Toutes les louanges à Allah et Sa bénédiction pour l'achèvement de ce travail.

Nous tenons particulièrement à remercier Professeur ***BELKHEIRI Mohammed*** à Université Amar TELIDJI Laghouat, pour avoir accepté de nous encadrer, pour la confiance qu'il nous a accordé, et les conseils prodigués tout au long de la réalisation de ce travail.

Nous lui exprimons notre gratitude pour nous avoir fourni les outils méthodologiques indispensables.

Nous tenons à exprimer toute notre reconnaissance et notre gratitude envers tous ceux qui nous ont soutenu et encouragé tout au long de notre parcours universitaire et nous ont permis de redoubler d'effort et de persévérance.

الملخص

خلال هذه المذكرة قمنا بتصميم نظامين مبنين على الذكاء الاصطناعي للتعرف على لغة الاشارة الجزائرية الساكنة و الحركية، صممنا اولا نظاما ذكيا لمعالجة الاشارات الساكنة كالحروف والتعرف عليها بعد تدريب الشبكة العصبية العميقة من نوع(الشبكات العصبية الالتفافية) **CNN** بقاعدة بيانات مكونة من **049 54** صورة وبعد تقديم صورة جديدة تحتوي على اشارة فان النظام يتعرف عليها بسهولة وتصل دقة التعرف لحوالي **98%** وفي الجزء الثاني لمشروعنا انجزنا نظاما ذكيا يتعرف على الإشارات المتحركة بقاعدة بيانات مكونة من **2040** فيديو فبعد تقسيم الفيديو الى صور(اطارات) ونطبق اداة **Mediapipe**التي تقوم بتحديد موقع الوجه و واحداثيات مفاصل اليد والقوام ونكتفي بعدد محدد من الصور من نفس الفيديو كجزء اول من العمل ونمرر هذه المميزات المستخلصة لعدة امثلة من نفس الاشارة ونقوم بادخالها للشبكة العصبية من نوع **LSTMs**الشبكات العصبية المتكررة ،**ResNet**شبكات الرواسب والتي تم تدريبها باستخدام خوارزمية **backpropagation** بفضل التصميم المعماري الذي اقترحناه ، في الاخير حققنا نتائج جد مرضية في التعرف على الإشارات الحركية

الكلمات المفتاحية: الشبكات العصبية المتكررة -الاشارة -الشبكات العصبية -الشبكات العصبية الالتفافية- شبكات الرواسب Mediapipe

Résumé

Dans ce mémoire, nous avons conçu deux systèmes basés sur l'intelligence artificielle pour identifier la langue des signes statique et dynamique algérienne. Nous avons d'abord conçu un système d'intelligence artificielle pour traiter et reconnaître les signes statiques tels que les "lettres arabes" après avoir entraîné le réseau neuronal profond de type CNN (réseaux neuronaux convolutifs) avec une base de données de54 049 images et après avoir introduit une nouvelle image contenant un signe, le système le reconnaît, et la précision de la reconnaissance atteint environ 98%. Dans la deuxième partie de notre projet, nous avons mis en œuvre un système intelligent qui reconnaît les signes dynamiques de base de données contient 2040 vidéos. Après avoir divisé la vidéo en images et appliqué l'outil MEDIAPIPE, qui détermine l'emplacement du visage, les coordonnées des articulations de la main, et la pose holistique, nous nous contentons d'un nombre spécifique fixe d'images de la même vidéo comme première partie et nous passons ces caractéristiques extraites à plusieurs exemples du même signe et nous les introduisons dans le réseau neuronal du type LSTMs (réseaux neuronaux récurrents) et ResNet (réseaux résiduels), qui ont été formés en utilisant l'algorithme de rétropropagation. Après un nombre suffisant d'époques d'entraînement, nous avons obtenu des résultats satisfaisants dans la reconnaissance de différents types de signes dynamiques.

Mots-clés**:** Signes, réseaux de neurones artificiels, caractéristiques, CNN, LSTMs, ResNet

Abstract

In this master project report, we have designed two systems based on artificial intelligence to identify the Algerian static and dynamic sign language. We first designed an AI system to process and recognize static signs such as “arabic letters” after training the deep neural network of CNN (convolutional neural networks) type with a dataset of 54 049 images and after introducing a new image containing a sign, the system recognizes it, and the recognition accuracy reaches about 98%. In the second part of our project, we implemented a smart system that recognizes dynamic signs with a dataset of 2040 videos. After dividing the video into frames and applying the MEDIAPIPE tool, which determines the location of the face, the coordinates of the hand joints, and the holistic pose, and we are satisfied with a fixed specific number of images from the same video as a first part and we pass these extracted features to several examples of the same sign and we introduced them to the neural network of the type LSTMs (recurrent neural networks) and ResNet (residual networks), which were trained using the backpropagation algorithm. After a sufficient training epoch, we achieved satisfactory results in the recognition of different types of dynamic signs

Keywords: Signs, artificial neural networks, feature, CNN, LSTMs, ResNet

Table des matières

Dédicaces i

Dédicaces ii

Remerciements iii

الملخص iv

Résumé v

Abstract vi

Table des matières ix

Liste des abréviations xi

Liste des tableaux xii

Liste des figures xiii

Introduction Générale 1

Chapitre 1 Langue des signes et outils de classification 3

1.1 Introduction 3

1.2 La langue des signes Algériennes 3

1.2.1 Communications entre les personnes sourdes-muettes 4

1.3 Le Deep Learning 5

1.4 Traitement d'image et vidéo 6

1.4.1 Définition du traitement d'images 6

1.4.2 Définition d'une image 6

1.4.3 Définition d'une vidéo 7

1.5 La sélection de signatures 8

1.6 Problème avec les réseaux neuronaux traditionnels 9

1.7 Réseaux neuronal convolutif (CNN) 10

1.7.1 Introduction 10

1.8 Réseau neuronal résiduel (ResNet) 14
1.9 Les réseaux de neurones récurrents LSTM 15
1.9.1 Définition 15
1.9.2 Architecture 15
1.10 Apprentissage par transfert 18
1.11 MEDIAPIPE 19
1.12 Conclusion 20
Chapitre 2 Préparation de la base de données 22
2.1 Introduction 22
2.2 Base de données statique 22
2.2.1 Description 22
2.2.2 Structure d'un réseau neuronal convolutif (CNN) 23
2.2.3 Augmentation des données 24
2.3 Préparation de la base de données dynamique 24
2.3.1 Extraction de séquences d'image 25
2.3.2 Extraction des points d'intérêt (keypoints) 28
2.4 Conclusion 30
Chapitre 3 Résultats de Classifications par DL 31
3.1 Introduction 31
3.2 Environnement matériel 31
3.3 Environnement logiciel 32
3.4 Résultats de classification Statique 32
3.5 Résultats de classification Dynamique 34
3.5.1 Les résultats d'extraction d'image 34
3.5.2 Les résultats du modèle 36
3.6 Conclusion 38
Conclusion Générale 39
Bibliographie 41

Liste des abréviations

ASL	Algerian Sign Language
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
DL	Deep Learning
ML	Machine Learning
IA	Intelligence Artificielle
ResNet	Residual Network
MLP	MultiLayer Perceptron
FC	Fully connected
TF	Transfert Learning
RGB	Red Green Bleu
NPY	NumPy array

Liste des tableaux

Tableau 1. Architecture du réseau résiduel à 50 couches. 14
Tableau 3. Les résultats d'extraction d'images 27
Tableau 4. Les résultats du Mediapipe 28
Tableau 5. Comparaison entre les différents types de réseaux de neurones utilisés 38

Liste des figures

Figure 1. Le dictionnaire de la langue des signes algérienne 4
Figure 2. Communication entre les sourds-muets 4
Figure 3. L'intelligence artificielle et ses sous-domaines 5
Figure 4. Systèmes à base de Traitement d'image 6
Figure 5. Visualisation des Pixels 7
Figure 6. Séquence d'images extraites de la Vidéo 7
Figure 7. Principe de sélection de signatures 8
Figure 8. Processus de sélection de signatures 8
Figure 9. Structure générale d'un réseau neuronal artificiel 9
Figure 10. Blocs de construction d'un réseau de neurones convolutifs (CNN) 10
Figure 11. Exemple d'une couche d'entrée 10
Figure 12. Opération de convolution appliquée à une images 11
Figure 13. Illustration des techniques de Max pooling & Average pooling 12
Figure 14. Principe de fonctionnement d'une fonction d'activation 12
Figure 15. Formules mathématiques des principales fonctions d'activation 13
Figure 16. Couches Fully Connected (FC) & Fonction Softmax 13
Figure 17. L'architecture ResNet 14
Figure 18. LSTM architecture 15
Figure 19. Fonctionnement d'une cellule LSTM 16
Figure 20. Fonctionnement de la porte d'oubli d'un LSTM 16
Figure 21. Fonctionnement de la porte d'entrée d'un LSTM 17
Figure 22. Fonctionnement d'état de la cellule d'un LSTM 17
Figure 23. Fonctionnement de la porte de sortie d'un LSTM 18
Figure 24. Schéma d'apprentissage par transfert 19
Figure 25. Aperçu de MediaPipe Holistique 20
Figure 26. Un échantillon des signes de la base de données utilisée 23

Figure 27. Extraction des séquences d'images à partir d'une vidéo 26
Figure 28. Les logiciels nécessaires pour notre projet 32
Figure 29. Couche de convolution après l'entraînement 33
Figure 30. Taux d’erreur et de reconnaissance CNN 33
Figure 31. Taux d’erreur et de reconnaissance LSTM.....36
Figure 32. Taux d’erreur et de reconnaissance ResNet.....37
Figure 33. Exemple d'un test.....37

Introduction Générale

Introduction Générale

En Algérie, un pays riche en diversité culturelle et linguistique, un écart important dans la façon dont les gens interagissent avec les sourds-muets est très perceptible [1]. La plupart d'entre nous avons rencontré au moins une fois des personnes ayant une déficience auditive ou des troubles de la parole. L'expérience que nous avons acquise dans ces situations nous a permis de constater que l'impossibilité de communiquer avec eux les prive de leurs droits les plus élémentaires d'intégrer la société [2]. Cela nous a profondément touchés en tant qu'étudiants en télécommunications, car la communication est au cœur de notre domaine d'études, et a renforcé notre désir d'apporter une contribution significative à notre communauté et inspiré du boom de l'intelligence artificielle, en particulier le 'deep learning' comme un outil précieux qui peut aider dans la résolution de nombreux problèmes et profite de ses progrès dans le traitement du langage naturel et les applications de vision.

Chaque fois que nous nous trouvons dans une situation où nous ne pouvons pas converser avec les sourdes ou leur fournir de l'aide, ce sentiment d'urgence et ce désir d'aider grandissent en nous. C'est pourquoi nous sommes motivés à choisir le thème de la langue des signes algérienne pour aider ces personnes marginalisées s'intégrer dans la communauté. En misant sur cette forme de communication visuelle, nous pourrons contribuer à rendre l'information et les ressources plus accessibles, favoriser l'intégration des personnes sourdes-muettes dans la société algérienne, sensibiliser à l'importance de la communication inclusive et aider ces personnes à développer leur autonomie.

L'objectif principal de ce projet de master est de concevoir un système intelligent de reconnaissance et classification de langue des signes algérienne basé sur des techniques intelligentes et le deep learning. Lors de la conception d'un tel système, la tâche importante est de créer une base de données approprié pour la création et l'entrainement du classificateur ASL. Ensuite, certaines techniques issues de l'IA ou de la vision sont invoquées pour extraire les

principales signatures (features) qui serons utilisées dans la phase de classification. Deux systèmes sont conçus : en premier temps un système basé sur des images pour classer les signes statiques (alphabet et nombres) conçu basé sur Les réseaux de neurones convolutifs, ensuite un système plus complexe est mis en place basé sur la vidéo composée d'un ensemble donné d'images pour la classification d'un dictionnaire de langage de signes algériennes composé d'une centaine des gloses de trois principaux sujets (corps humain, animaux et membres de la famille).

Outre l'introduction et la conclusion ; ce mémoire de master est divisé en trois chapitres. Dans le premier chapitre, nous aborderons de la communication des sourdes par le langage des signes, puis nous présenterons les outils de classification par deep learning et le Mediapipe. Dans le deuxième chapitre, nous nous concentrerons sur les deux bases de données que nous avons utilisées pour créer et trainer les classificateurs. Dans le troisième chapitre, nous présentons les résultats de la reconnaissance de la langue des signes algérienne à l'aide des différents types de réseaux de neurones profonds à savoir les CNNS, RESNET et LSTM et ainsi on discutera les performances atteintes pour les deux bases de données (signes statiques : images et signes dynamiques : vidéos).

Chapitre 1

Langue des signes et outils de classification

Chapitre 1

Langue des signes et outils de classification

1.1 Introduction

Le langage des signes est une méthode de communication visuelle utilisée par les personnes sourdes-muettes pour exprimer leurs besoins. Il utilise des gestes des mains et du langage corporel pour transmettre le sens [3]. Son utilisation remonte à l'histoire, avec Pedro Ponce de Leon en 1500, qui a développé l'alphabet manuel pour l'éducation des élèves sourds-muets en Espagne [4]. Les utilisateurs du langage des signes peuvent choisir entre les alphabets manuels ou les signes représentant des mots spécifiques. Il existe plusieurs langues des signes utilisées dans différentes régions, chacune étant spécifique à la langue parlée localement. Dans le monde arabe, la langue des signes utilisée n'est pas la même que dans d'autres régions [5].

Dans ce chapitre, nous allons nous concentrer sur la communication gestuelle, plus précisément sur la langue des signes. Nous aborderons les outils de classification, tels que les réseaux de neurones et les réseaux pré-entraînés, ainsi que l'utilisation l'outil puissant de google Mediapipe pour l'extraction des signatures pour le suivi des mouvements gestuels.

1.2 La langue des signes Algériennes

La langue des signes algérienne (ASL) a été officiellement reconnue par la loi algérienne sur la protection des personnes handicapées en 2002. Cependant, elle n'est pas enseignée dans les écoles régulières, ce qui limite la capacité des citoyens algériens à communiquer avec les personnes sourdes-muettes. Les campagnes de sensibilisation lancées par ces personnes pour promouvoir la communication avec la population entendante sont souvent ignorées ou négligées. Les personnes entendantes peuvent considérer l'apprentissage de la langue des signes comme

inutile, en raison du manque d'interactions fréquentes avec les personnes sourdes-muettes et de la vie trépidante qu'elles mènent. [6]

Figure 1. Le dictionnaire de la langue des signes algérienne

1.2.1 Communications entre les personnes sourdes-muettes

Comme évoqué précédemment, les personnes sourdes-muettes ont la capacité d'adapter et d'enrichir la langue des signes. Dans cette langue dynamique, chaque lettre de l'alphabet arabe est représentée par un signe distinct, et ces signes peuvent ensuite être utilisés pour former des mots. De plus, les utilisateurs de la langue des signes ont la liberté de créer de nouveaux signes et d'utiliser des abréviations spéciales, facilitant ainsi leurs échanges dans divers contextes, y compris les interactions informelles en extérieur. Cette flexibilité linguistique leur permet d'exprimer leurs pensées et de communiquer efficacement entre eux.

Figure 2. Communication entre les sourds-muets

1.3 Le Deep Learning

Le deep learning et le machine learning font tous deux appels à l'intelligence artificielle celle-ci vise à reproduire l'intelligence humaine par des moyen informatiques par exemple pour conduire une voiture autonome ou pour faire dialoguer un CHATBOT.

Le machine learning est un sous domaine de l'intelligence artificielle, il vise à explorer des données pour apprendre à faire une tache ou établir un modèle prédictif, c'est donc un humain qui décide quelles données fournir à la machine.

Le deep learning est une forme particulière de machine learning, il utilise des réseaux de neurones artificiels pour établir ses propres décisions. Le deep learning fait souvent appel à des données non structurées : cela signifie qu'elles sont hétérogènes (images, sons, texte…).

C'est l'algorithme que décide quelles données sont utiles pour son apprentissage. Le deep learning a besoin d'un volume de données considérable pour être performant. Il nécessite aussi une énorme puissance de calcul. [7] [1]

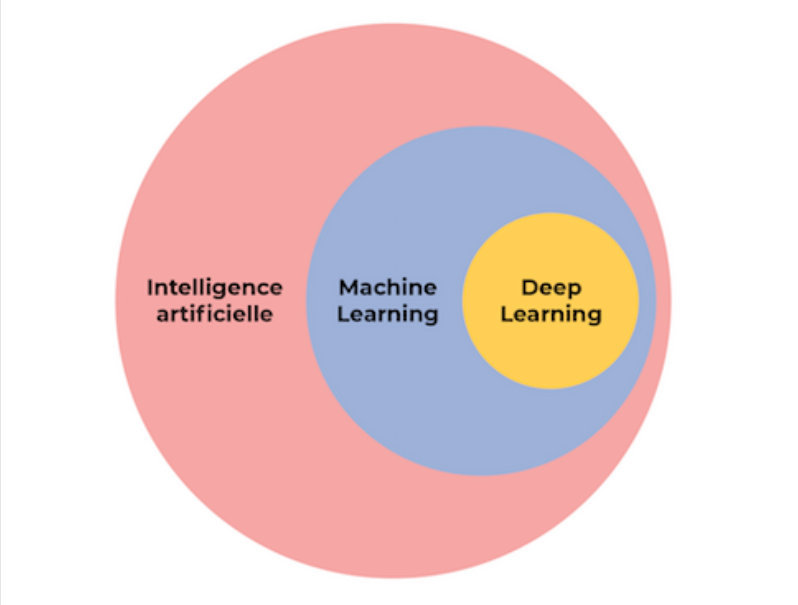


Figure 3. L'intelligence artificielle et ses sous-domaines

1.4 Traitement d'image et vidéo

1.4.1 Définition du traitement d'images

Le traitement d'image englobe les différentes opérations qui peuvent être effectuées sur des images. Avant de pouvoir appliquer ces opérations, il est nécessaire de capturer les images, c'est-à-dire de les convertir du monde réel en une représentation numérique. Une fois cette étape accomplie, les images sont transformées en matrices qui peuvent ensuite être traitées de différentes manières [9].

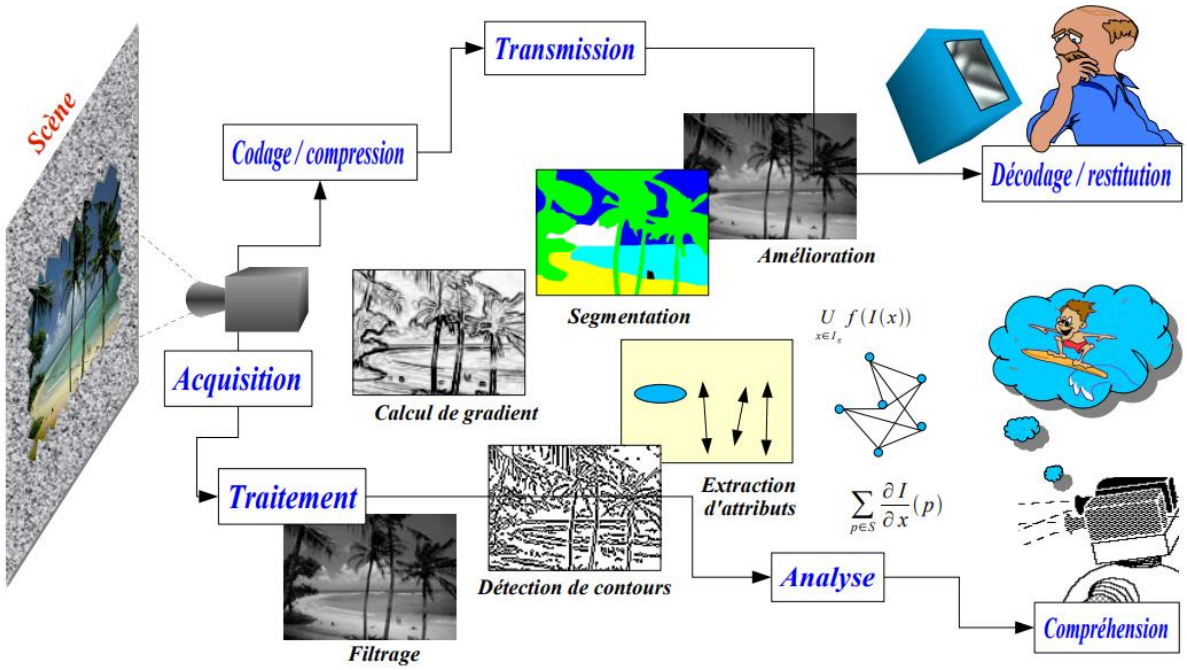


Figure 4. Systèmes à base de Traitement d'image

1.4.2 Définition d'une image

L'image est une représentation visuelle statique composée de pixels, où chaque pixel contient des informations sur la couleur et l'intensité lumineuse. Elle peut être en couleur (utilisant les canaux rouge, vert et bleu) ou en niveaux de gris (utilisant un seul canal d'intensité lumineuse). Le traitement d'images consiste à analyser, améliorer et extraire des informations à partir de ces images en utilisant différentes caractéristiques telles que la texture et les contours. [10]

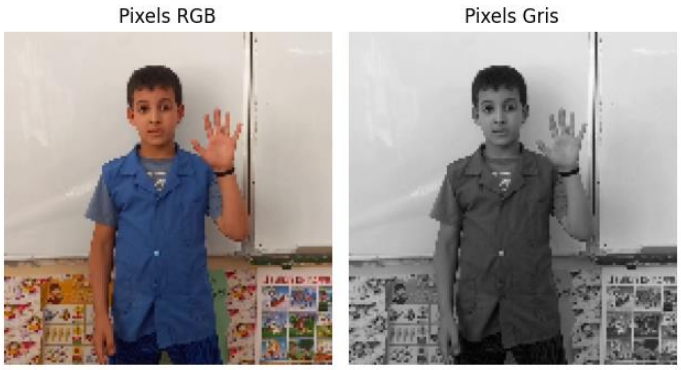


Figure 5. Visualisation des Pixels

1.4.3 Définition d'une vidéo

Une vidéo est un ensemble séquentiel de frames, où chaque frame est une image individuelle capturée à un moment précis. Elle est composée d'une séquence d'images en mouvement qui, lorsqu'elles sont lues à une vitesse suffisante, créent l'illusion du mouvement. Dans le traitement de vidéos, les mêmes principes de traitement d'images s'appliquent, mais en exploitant également les informations temporelles pour des tâches telles que la détection de mouvement et le suivi d'objets. [11]

Figure 6. Séquence d'images extraites de la Vidéo

1.5 La sélection de signatures

La sélection de signatures (feature selection en anglais) est une méthode de réduction de la dimensionnalité utilisée en apprentissage automatique et en traitement de données. Il consiste, dans un espace de grande dimension, à trouver un sous-ensemble de variables pertinentes. C'est-à-dire que l'on cherche à minimiser la perte d'information venant de la suppression de toutes les autres variables [12].

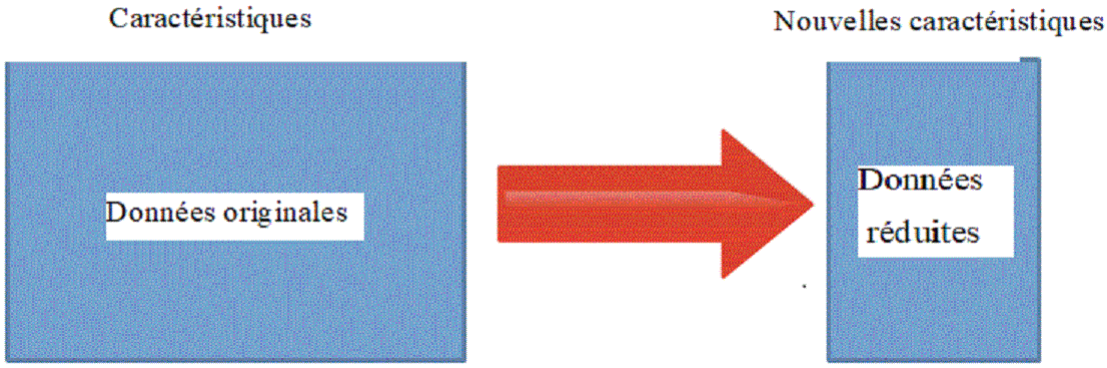


Figure 7. Principe de sélection de signatures

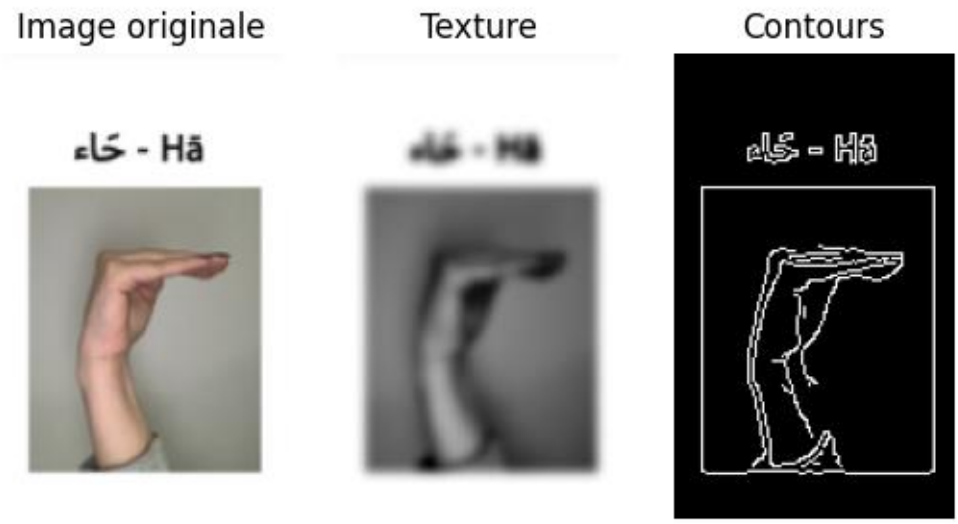


Figure 8. Processus de sélection de signatures

Avec l'avènement des réseaux neuronal convolutif (CNN), la sélection de signatures a été révolutionnée. Les CNN peuvent apprendre automatiquement et extraire les caractéristiques les

plus importantes à partir des données brutes, éliminant ainsi la nécessité d'une sélection manuelle. Grâce à leur architecture profonde et à leurs couches de convolution, les CNN détectent les motifs, les textures et les contours pertinents à différentes échelles et niveaux d'abstraction. Cela simplifie le processus d'analyse des données, accélère le flux de travail et améliore souvent les performances des modèles. En résumé, les CNN permettent une sélection de signatures automatique et adaptative, offrant une approche plus puissante pour traiter les données complexes.[13]

1.6 Problème avec les réseaux neuronaux traditionnels

Les réseaux neuronaux traditionnels, tels que les perceptron multicouche **(multi-layer perceptron, MLP)**, rencontrent des problèmes dans le traitement d'images. Les poids deviennent ingérables pour les grandes images, ce qui entraîne des difficultés de calcul et de mémoire. De plus, les MLP réagissent différemment aux images et à leurs versions traduites, perdant ainsi les relations spatiales. En aplatissant les images, on perd également des informations spatiales. Par conséquent, les MLP ne sont pas adaptés au traitement d'images. Les architectures spécialisées comme les CNN sont préférables, car elles préservent les informations spatiales, capturent les motifs locaux et offrent de meilleures performances pour la classification, la détection et la segmentation d'images [14].

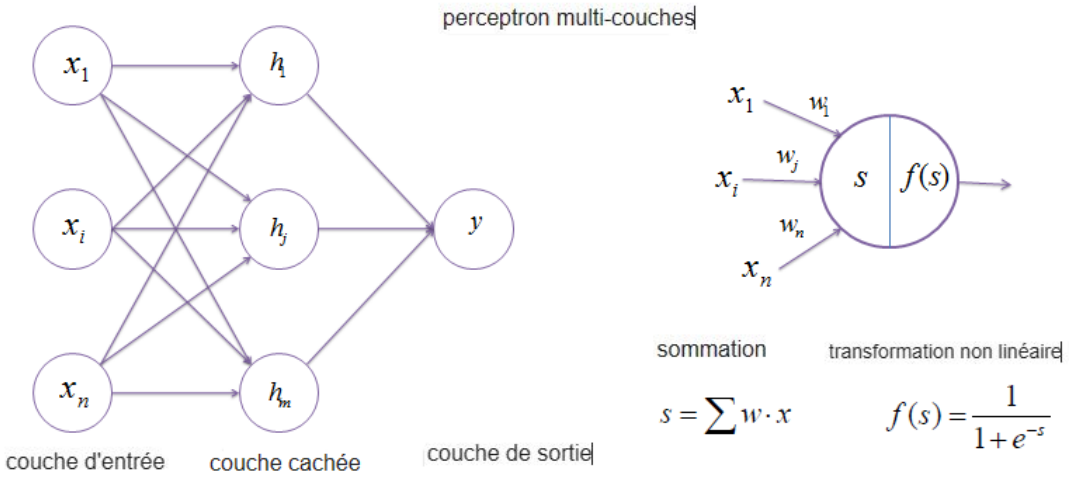


Figure 9. Structure générale d'un réseau neuronal artificiel

1.7 Réseaux neuronal convolutif (CNN)

1.7.1 Introduction

Comme le réseau neuronal traditionnel, le CNN est composé d'une couche d'entrée, de couches cachées et d'une couche de sortie. La différence réside dans le fait que l'entrée du CNN est une image (une matrice de pixels), et la sortie est la caractéristique de l'image obtenue par le calcul de convolution. [15][16]

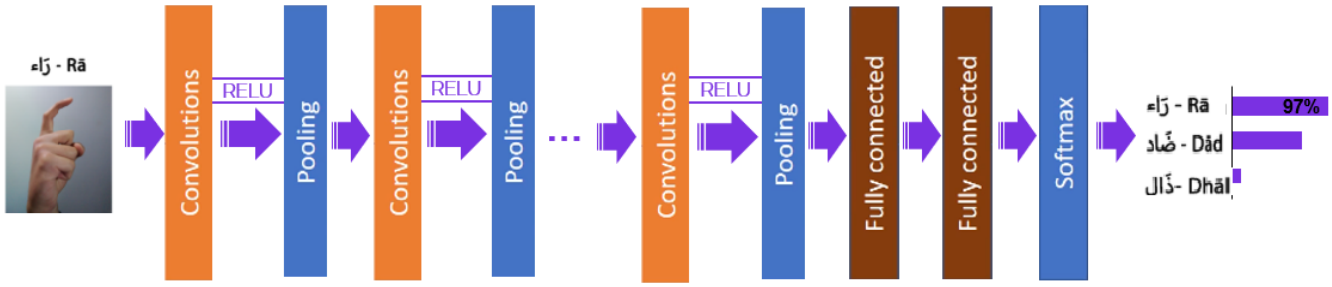


Figure 10. Blocs de construction d'un réseau de neurones convolutifs (CNN)

1. Couche d'entrée (Input Layer)

Comme son nom l'indique, c'est notre image d'entrée, qui peut être en niveaux de gris ou en RGB. Chaque image est composée de pixels dont les valeurs vont de 0 à 255. Avant de les transmettre au modèle, nous devons les normaliser, c'est-à-dire convertir leur plage de valeurs entre 0 et 1.

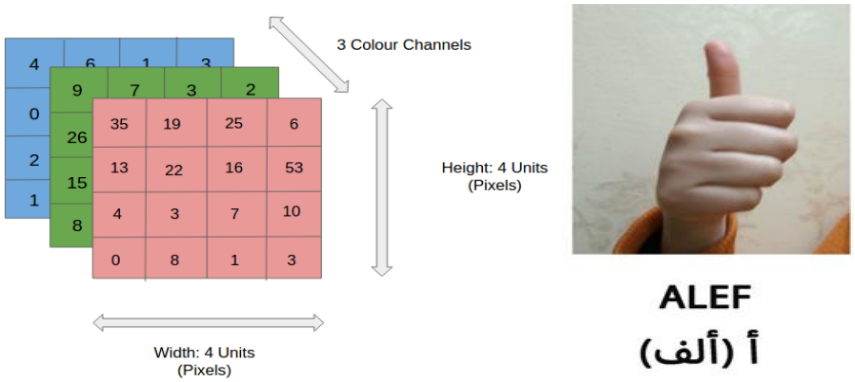


Figure 11. Exemple d'une couche d'entrée

2. Couche de convolution (Convolutional Layer)

- La couche de convolution

La couche de convolution applique un filtre à une image d'entrée afin d'extraire ses caractéristiques. Cette opération est répétée plusieurs fois pour créer une carte de caractéristiques qui aide à classifier l'image. Dans notre exemple, nous avons utilisé une image 2D de taille 6x6 et appliqué un filtre de taille 3x3 pour détecter certaines caractéristiques. En pratique, de nombreux filtres similaires sont utilisés pour extraire des informations de l'image. Le résultat est une carte de caractéristiques de taille 4x4 qui contient des informations spécifiques à l'image. De nombreuses cartes de caractéristiques similaires sont générées dans des applications réelles.

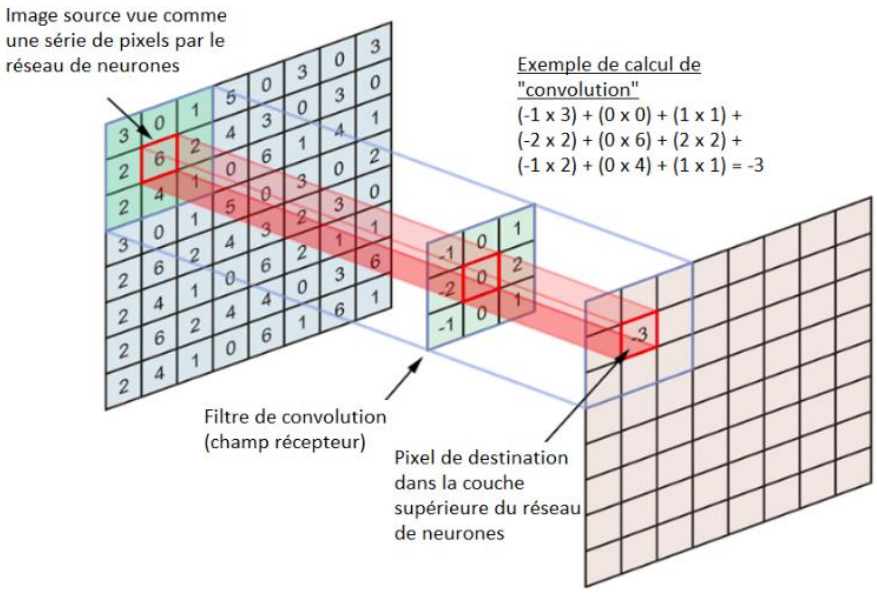


Figure 12. Opération de convolution appliquée à une images

- La couche de pooling

La couche de pooling est utilisée pour réduire la taille des représentations et accélérer les calculs, ainsi que pour rendre certaines des caractéristiques détectées un peu plus robustes.

Les types de pooling courants sont le pooling maximum (max pooling) et le pooling moyen (avg pooling), mais de nos jours, le pooling maximum est plus courant.

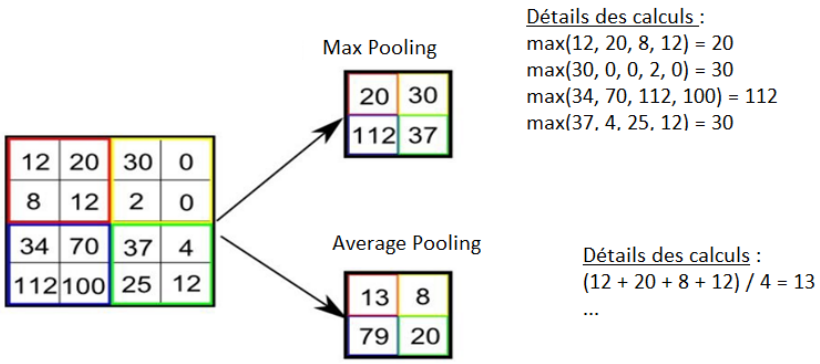


Figure 13. Illustration des techniques de Max pooling & Average pooling

- La fonction d'activation

La fonction d'activation est une technique pour résoudre un problème majeur lié au processus de rétropropagation de l'erreur dans le réseau de neurones.

Il y a 2 problèmes majeurs rencontrés lors de la rétropagation de l'erreur :

- Il se peut que les gradients d'erreurs disparaissent, c'est-à-dire qu'au fur et à mesure que l'information progresse dans les couches du réseau, les gradients deviennent très petits et par conséquent, la mise à jour des poids de connexion n'est presque plus effectuée. Alors, le modèle ne converge jamais vers une bonne solution.
- Ou alors au contraire, les gradients d'erreurs deviennent de plus en plus grands et explosent. Alors, les poids de connexion deviennent très grands à leur tour et le modèle diverge.

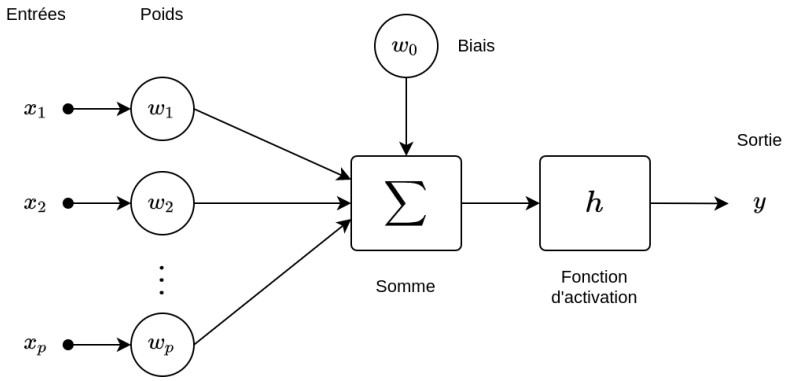


Figure 14. Principe de fonctionnement d'une fonction d'activation

Les fonctions d'activation les plus couramment utilisées :

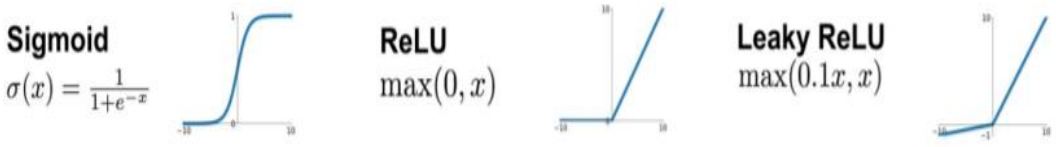


Figure 15. Formules mathématiques des principales fonctions d'activation

- Couche Fully Connected (FC)

La couche "fully connected" est la dernière étape d'un réseau de neurones. Elle est complètement connectée, ce qui signifie que tous les neurones sont connectés les uns aux autres. La fonction d'activation utilisée dans cette couche est généralement la fonction softmax, qui attribue à chaque sortie du réseau un score, converti ensuite en probabilité pour chaque label d'image.

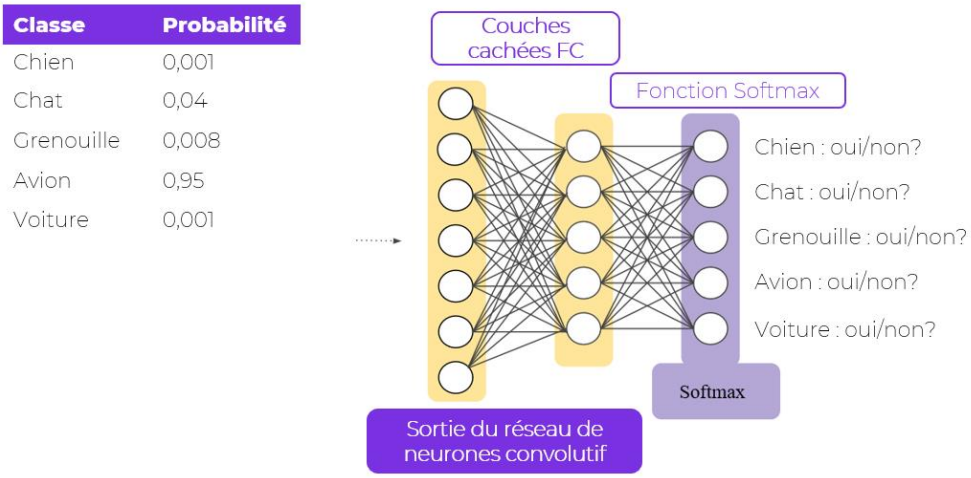


Figure 16. Couches Fully Connected (FC) & Fonction Softmax

1.8 Réseau neuronal résiduel (ResNet)

Les ResNets (Residual Networks) sont des modèles pré-entraînés très populaire dans le domaine de la vision par ordinateur, développés par Kaiming He et d'autres chercheurs, dans l'article "Deep Residual Learning for Image Recognition" en 2015, et sont considérés comme les modèles de réseaux neuronaux convolutionnels les plus avancés à ce jour. Ils sont largement utilisés dans la pratique comme choix par défaut pour les ConvNets (réseaux neuronaux convolutionnels) depuis mai 2016. Les ResNets ont surpassé les autres architectures de réseau neuronal grâce à leur utilisation de connexions résiduelles, ce qui a permis de résoudre les problèmes de formation des réseaux profonds et d'obtenir de meilleures performances en termes de précision de classification. [17]

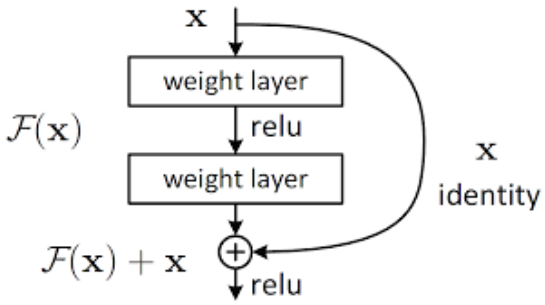


Figure 17. L'architecture ResNet

Tableau 1. Architecture du réseau résiduel à 50 couches.

Noms des couches	Taille de sortie	Organisation des 50 couches
Conv1	112x112	7x7, 64, stride 2
Conv2_x	56x56	3x3 max pool, stride 2
		1x1, 64 3x3, 64 x 3 1x1, 256
Conv3_x	28x28	1x1, 128 3x3, 128 x 4 1x1, 512
Conv4_x	14x14	1x1, 256 3x3, 256 x 6 1x1, 1024
Conv5_x	7x7	1x1, 512 3x3, 512 x3 1x1, 2048
	1x1	Couche Fully Connected

Les avantages des ResNets sont les suivants :

a. Les performances ne se dégradent pas avec des réseaux très profonds.
b. Moins coûteux en termes de calcul.
c. Capacité à entraîner des réseaux très profonds.

1.9 Les réseaux de neurones récurrents LSTM

1.9.1 Définition

LSTM, ou Mémoire à Long Terme - Court Terme, est un type de réseau neuronal récurrent qui se distingue par sa capacité supérieure à stocker et à rappeler des informations sur une longue période. Les LSTM surpassent les réseaux neuronaux récurrents traditionnels en termes de mémoire, ce qui leur permet d'obtenir de meilleurs résultats, notamment dans la reconnaissance de motifs. [18] [19]

1.9.2 Architecture

L'architecture d'un modèle élémentaire de LSTM est représentée dans la figure suivante :

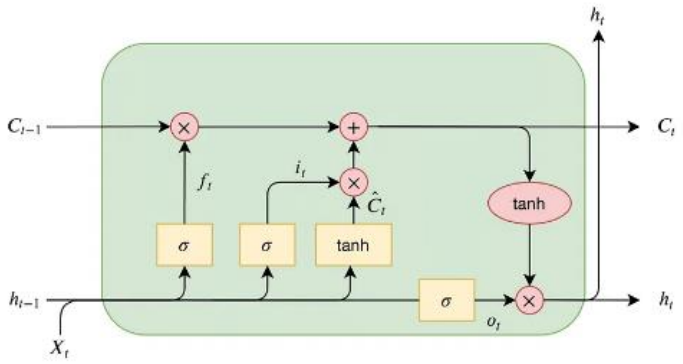


Figure 18. LSTM architecture

Ou Xt : pas de temps d'entrée, ht : sortie, Ct : état de la cellule, it : porte d'entrée
, ft : porte d'oubli, Ot : porte de sortie, Äˆt : état interne de la cellule. Les opérations à l'intérieur du cercle rouge clair sont effectuées élément par élément.

Un LSTM fonctionne en utilisant des cellules spéciales appelées "cellules LSTM". Chaque cellule LSTM possède des portes qui permettent de contrôler le flux d'informations à l'intérieur de la cellule. Ces portes comprennent une porte d'oubli (forget gate), une porte d'entrée (input gate) et une porte de sortie (output gate).

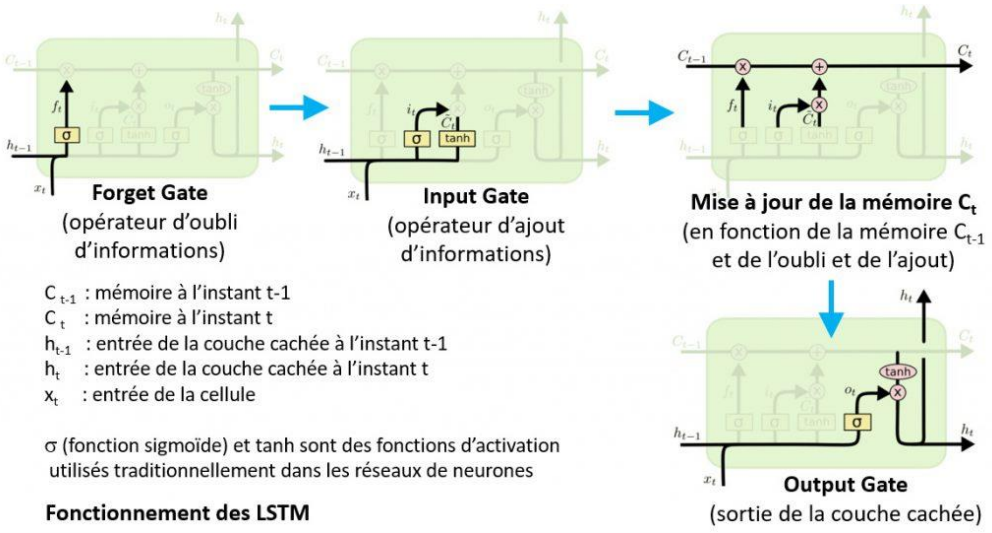
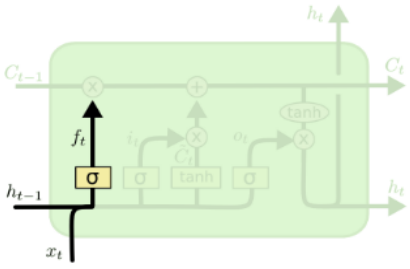


Figure 19. Fonctionnement d'une cellule LSTM

- **Porte d'oubli (forget gate)**

Cette porte décide de quelle information doit être conservée ou jetée : l'information de l'état caché précédent est concaténée à la donnée en entrée (par exemple le mot « des » vectorisé) puis on y applique la fonction sigmoïde afin de normaliser les valeurs entre 0 et 1. Si la sortie de la sigmoïde est proche de 0, cela signifie que l'on doit oublier l'information et si on est proche de 1 alors il faut la mémoriser pour la suite.

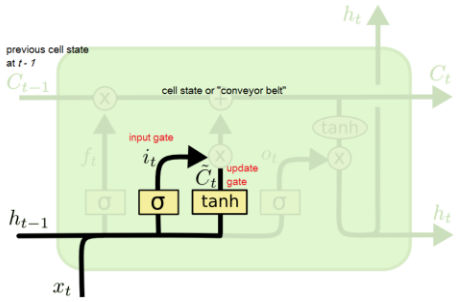


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Figure 20. Fonctionnement de la porte d'oubli d'un LSTM

- Porte d'entrée (input gate)

La porte d'entrée a pour rôle d'extraire l'information de la donnée courante (le mot « des » par exemple) : on va appliquer en parallèle une sigmoïde aux deux données concaténées (cf porte précédente) et une tanh.

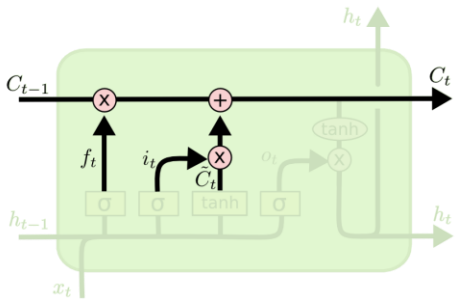


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Figure 21. Fonctionnement de la porte d'entrée d'un LSTM

- Etat de la cellule (cell state)

L'état de la cellule est calculé en utilisant la porte d'oubli et la porte d'entrée. Il est obtenu en multipliant élément par élément la sortie de la porte d'oubli avec l'ancien état de la cellule, ce qui permet d'oublier certaines informations non pertinentes. Ensuite, on ajoute (élément par élément) la sortie de la porte d'entrée à ce résultat, ce qui permet d'enregistrer dans l'état de la cellule les informations jugées pertinentes par le LSTM parmi les entrées et l'état caché précédent.

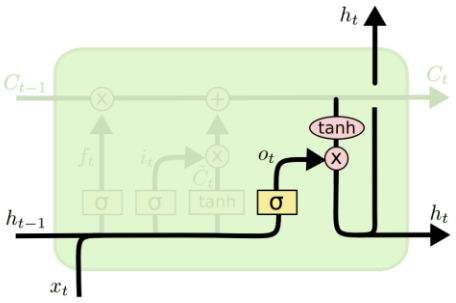


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure 22. Fonctionnement d'état de la cellule d'un LSTM

- Porte de sortie (output gate)

Dernière étape : la porte de sortie doit décider de quel sera le prochain état caché, qui contient des informations sur les entrées précédentes du réseau et sert aux prédictions.



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Figure 23. Fonctionnement de la porte de sortie d'un LSTM

1.10 Apprentissage par transfert

Lorsque nous disposons d'un ensemble de données volumineux (énorme), l'entraînement d'un modèle d'apprentissage automatique peut nécessiter des ressources de calcul importantes. Dans de tels cas, L'apprentissage par transfert (ou Transfer Learning en anglais) est peut se révéler être une solution efficace.

L'apprentissage par transfert permet d'utiliser les connaissances acquises à partir de modèles pré-entraînés sur de grands ensembles de données pour accélérer l'apprentissage sur l'ensemble de données volumineux. Cela réduit le besoin de ressources de calcul puissantes et permet d'obtenir de bonnes performances même avec des ensembles de données limités ou étiquetés de manière insuffisante.

L'adaptation d'un modèle pré-entraîné en modifiant principalement les dernières couches pour s'adapter à une nouvelle tâche cible est appelée L'apprentissage par transfert. Par rapport à un entraînement complet à partir de zéro, il est possible d'améliorer les performances sur la tâche cible tout en économisant du temps et des ressources en réutilisant les connaissances préexistantes du modèle et en se concentrant sur les ajustements spécifiques aux dernières couches. Cette méthode permet de bénéficier des connaissances générales du modèle pré-entraîné tout en ajustant les sorties finales pour répondre aux exigences spécifiques de la tâche cible. [20]

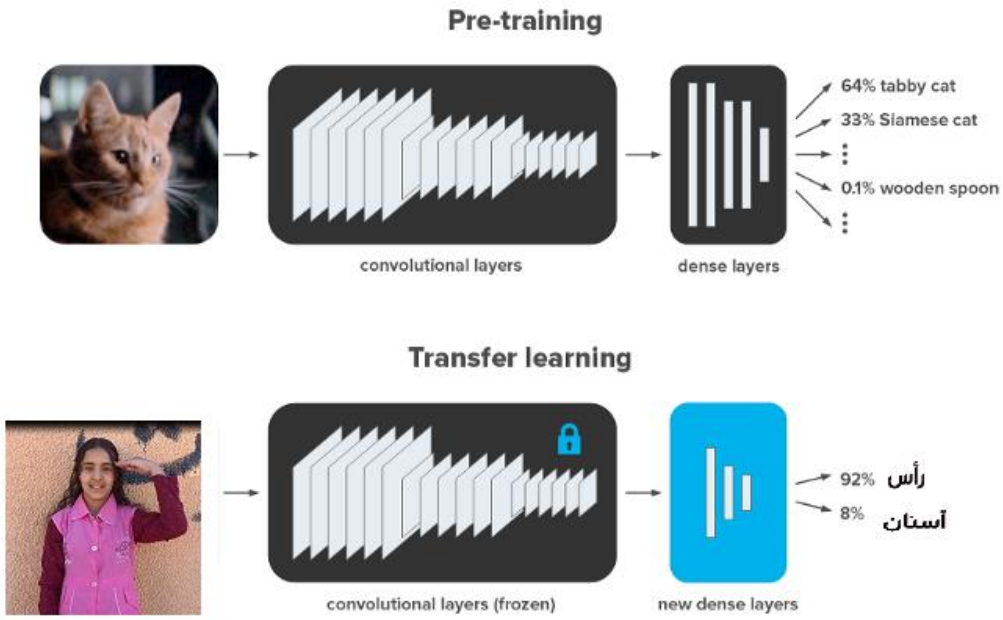


Figure 24. Schéma d'apprentissage par transfert

1.11MEDIAPIPE

Le MEDIAPIPE est un framework open-source, conçu par google qui permet de créer des pipelines pour le traitement de données perceptuelles telles que les images, les vidéos et l'audio. Il est utilisé pour le suivi des mains, la reconnaissance des gestes et l'analyse faciale en temps réel. MediaPipe utilise des modèles de réseaux neuronaux pour détecter et localiser les points d'intérêt (ou keypoints en anglais), du visage, des mains et de la pose du corps. Ces points d'intérêt sont des points 3D qui fournissent des informations précises sur la position et les mouvements des différentes parties du corps. L'utilisation de MediaPipe permet d'obtenir une localisation précise des points d'intérêt, ce qui facilite la reconnaissance des gestes, la pose et le suivi des mains, ainsi que l'analyse faciale. [21] [22]

Figure 25. Aperçu de MediaPipe Holistique

1.12 Conclusion

Dans ce chapitre on a parlé tout d'abord sur la diversité de la langue des signes, qui varie d'un pays à un autre. Nous avons constaté que la langue des signes égyptienne diffère de la langue des signes algérienne et tunisienne ...etc. Ensuite on a parlé du deep : sa définition et ces domaines d'application Learning. Nous avons également exploré les différents outils de classification utilisés, tels que les réseaux de neurones CNN, LSTM et ResNet, qui sont essentiels pour l'analyse et la reconnaissance des signes. De plus, nous avons souligné l'importance de

Mediapipe, une bibliothèque logicielle qui offre des fonctionnalités puissantes pour la détection et le suivi des mouvements des mains dans les vidéos. Ces outils et technologies jouent un rôle crucial dans le développement de notre système.

Dans le chapitre suivant, nous présenterons la méthode que nous avons choisie pour notre système de reconnaissance, ainsi que la préparation de la base de données.

Chapitre 2

Préparation de la base de données

Chapitre 2

Préparation de la base de données

2.1 Introduction

Pour construire un système de détection doté d'une grande précision, il est essentiel de disposer de données d'entraînement riches, comprenant un nombre suffisant de classes et d'exemples d'entraînement. La qualité et la diversité des données jouent un rôle crucial dans l'apprentissage des modèles et leur capacité à généraliser efficacement.

Dans ce chapitre, nous aborderons la préparation de deux types de bases de données distincts pour soutenir la construction de systèmes de détection performants.

2.2 Base de données statique

2.2.1 Description

L'ArSL2018 est un nouvel ensemble de données complet et entièrement étiqueté d'images de la langue des signes arabe, lancé à l'Université Prince Mohammad Bin Fahd à Al Khobar, en Arabie saoudite. L'ensemble de données ArSL2018 comprend 54 049 images en niveaux de gris, de dimensions 64 × 64, représentant 32 signes de la langue des signes arabe. Ces images ont été réalisées par plus de 40 personnes différentes, garantissant ainsi une diversité dans les styles et les perspectives de capture. Une particularité de cet ensemble de données est que le nombre d'images par classe varie d'une classe à l'autre, ce qui suggère une répartition inégale des données. Cependant, pour faciliter la classification, un fichier CSV accompagne les images, fournissant une étiquette pour chaque image correspondant à la langue des signes arabe. Ces étiquettes sont basées sur le nom du fichier image, offrant ainsi un moyen pratique d'associer

chaque image à sa classe respective. Cette ressource sera d'une grande valeur pour les projets liés à la reconnaissance et à l'interprétation des signes de la langue des signes arabe. La Figure 26 présente un échantillon des signes et des alphabets de la langue des signes arabe inclus dans cet ensemble de données. [23]

Figure 26. Un échantillon des signes de la base de données utilisée

2.2.2 Structure d'un réseau neuronal convolutif (CNN)

Le réseau neuronal convolutif (CNN) est utilisé pour traiter la base de données d'images ArSL2018 (Arabic Sign Language), utilise trois couches de convolution pour extraire les caractéristiques des images. Chaque couche de convolution est suivie d'une couche de pooling qui réduit la dimensionnalité des caractéristiques en utilisant un max pooling avec un noyau de taille 2x2. Les dimensions des caractéristiques après les couches de convolution et de pooling sont 128 x 3 x 3. Ensuite, les caractéristiques sont aplaties et alimentées dans une couche entièrement connectée avec 512 neurones et une fonction d'activation ReLU. Une couche de dropout avec un taux de 0,25 est ajoutée pour prévenir le surapprentissage. Enfin, une deuxième

couche entièrement connectée avec 32 neurones est utilisée, suivie d'une fonction de softmax pour obtenir les probabilités de classe. La fonction d'activation ReLU est utilisée pour toutes les couches de convolution, sauf pour la dernière couche où une fonction de softmax est appliquée pour obtenir les probabilités de classe. Les images en entrée sont des images RGB avec 3 canaux.

2.2.3 Augmentation des données

Puisqu'il y a un déséquilibre des données dans ArSL2018, il est important de noter que l'augmentation de données peut être utilisée pour équilibrer cet ensemble de données. L'augmentation de données est une méthode pour atténuer le problème de surapprentissage (overfitting) consiste à réaliser une augmentation de données en effectuant des transformations aléatoires sur les images d'entraînement, telles La rotation, Le flou et bruit, Le contraste et le zoom ou rognage.

Du fait que ces transformations d'augmentation de données sont appliquées de manière aléatoire pendant l'entraînement, une même image peut être présentée de manière différente d'un lot à l'autre, ce qui crée davantage de variations dans les données d'entraînement. Cela aide le modèle à apprendre les caractéristiques des mêmes objets dans différentes orientations ou échelles.

2.3 Préparation de la base de données dynamique

Notre projet s'est inspiré de la base de données Word-Level American Sign Language (WLASL), qui a remporté le prix du meilleur article lors de la conférence d'hiver de l'IEEE en mars 2020. La WLASL comprend plusieurs sous-ensembles, tels que le WLASL2000, WLASL100, WLASL300 et WLASL1000, qui diffèrent en termes du nombre total de vidéos disponibles dans chaque sous-ensemble. Dans notre cas, nous avons utilisé un sous-ensemble basé sur l'article WLASL contenant 100 mots. Le dictionnaire de la langue des signes algérienne constitué de 1560 mots les plus usités et composé de 29 thèmes en utilisant 100 mots différents répartis dans trois classes distinctes : le corps humain, les membres de la famille et les animaux. Pour garantir une représentation complète, chaque vidéo de notre base de données a été répétée 20 fois, ce qui signifie que nous disposons de 2040 vidéos au total. Ces vidéos ont été réalisées par 20 personnes différentes, ce qui assure une diversité d'interprétations et de styles de communication en langue

des signes. De plus, notre base de données se concentre sur une tranche d'âge spécifique, avec des participants âgés de 10 à 23 ans.

Nous avons pris la décision de visiter une école pour sourds-muets afin de mieux comprendre leur environnement et leurs besoins. Malheureusement, en raison de contraintes légales, nous n'avons pas été autorisés à prendre des captures d'écran ou à enregistrer des vidéos pendant notre visite. Cependant, nous avons eu l'opportunité d'observer et d'interagir directement avec les étudiants et le personnel de l'école, ce qui nous a permis de recueillir des informations précieuses pour notre projet. L'école de Bait Ben Jeddou, qui a grandement contribué à notre projet. Grâce à cette collaboration avec l'école publique, nous avons pu bénéficier d'une plus grande participation et capturer une plus grande diversité de la langue des signes. Cette collaboration a été essentielle pour créer un ensemble de données complet dont nous disposons aujourd'hui. Nous exprimons notre profonde gratitude envers l'école publique pour son rôle central dans la facilitation de ce processus.

En complément, il est important de noter que nous avons utilisé uniquement un téléphone portable pour capturer les vidéos de notre ensemble de données. Nous tenons à préciser que ces enregistrements n'ont pas été validés par des experts en langue des signes. Nous avons fait tout notre possible pour garantir l'exactitude et la qualité des signes capturés. Cependant, il est possible qu'il y ait des variations ou des erreurs lors de l'interprétation des signes. Par ailleurs, il est pertinent de mentionner que nous avons inclus dans notre ensemble de données un exemple de vidéos provenant du dictionnaire algérien de la langue des signes. Ces vidéos ont été utilisées dans le but d'enrichir la diversité et la représentativité des signes dans notre ensemble de données.

2.3.1 Extraction de séquences d’image

L'idée principale est d'extraire les séquences d'une vidéo tout en tenant compte de leurs différentes durées, qui varient entre 78 et 175 séquences. Pour résoudre ce problème, nous proposons d'adopter une approche consistant à fixer une taille standardisée pour toutes les vidéos, indépendamment de leur longueur d'origine. Cette approche permettra de traiter toutes les vidéos de manière cohérente, en les découpant en séquences de taille égale.

Pour l’extraction les séquences d'une vidéo, nous avons utilisé deux méthodes différentes.

- **La première méthode**

Appelée "espacement égal", consistait à sélectionner 30 frames de la vidéo en calculant un intervalle régulier entre chaque image. Cependant, nous avons découvert que cette technique ne capturait pas l'intégralité du mouvement, ce qui rendait la vidéo incomplète.

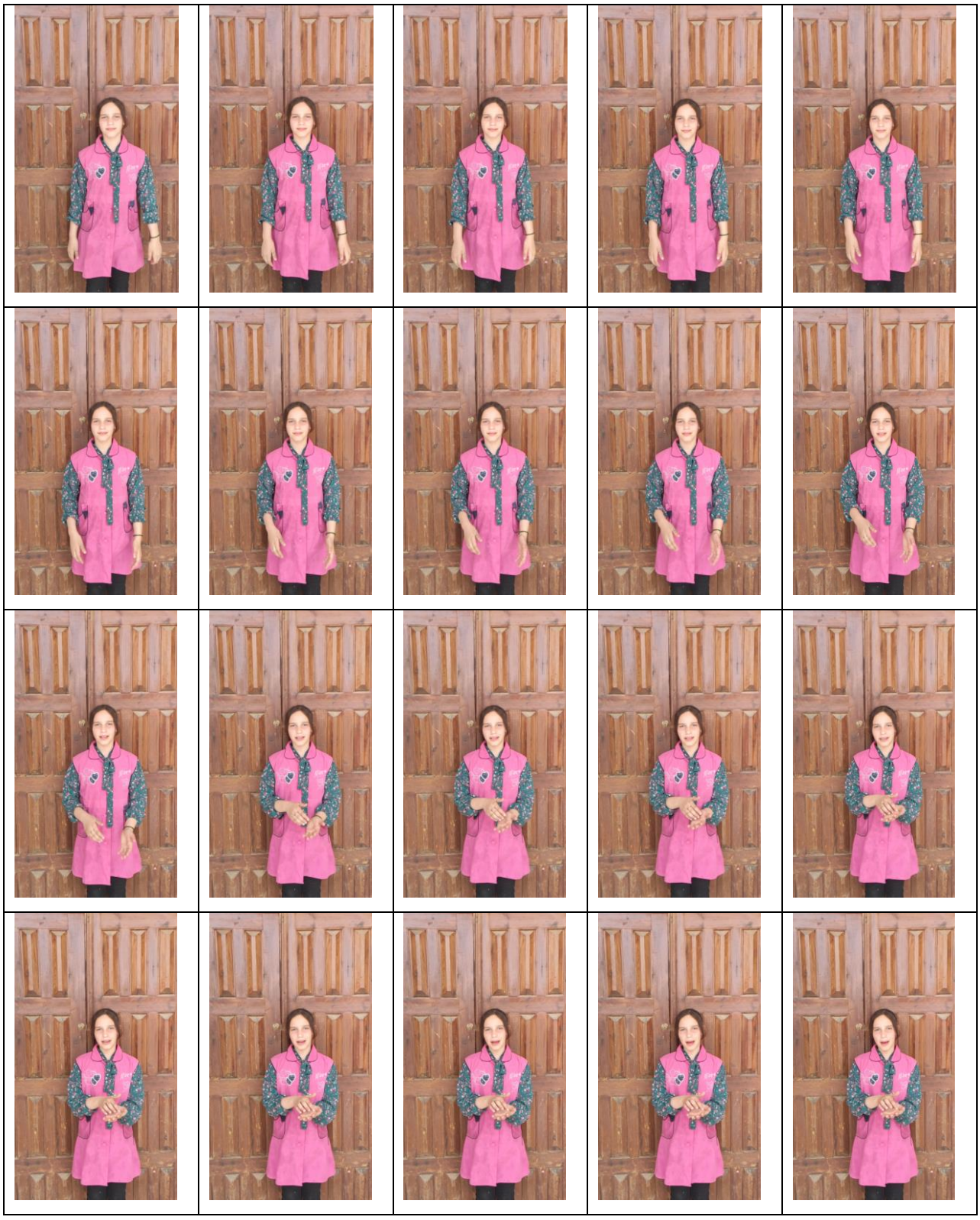
- **La deuxième méthode**

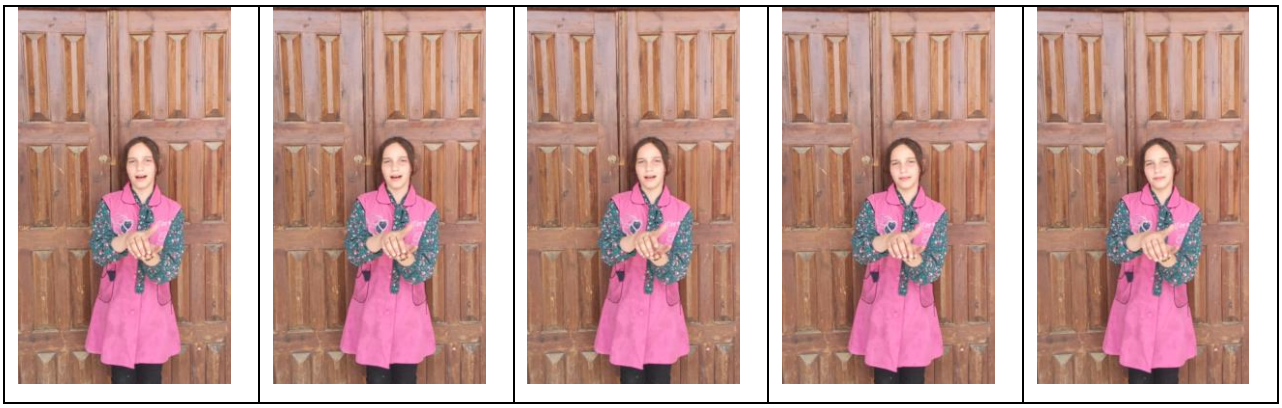
Pour remédier à ce problème, nous avons adopté une deuxième méthode, basée sur "le maximum de la norme de la différence des points clés entre deux images adjacentes. ". Cette approche nous a permis de sélectionner les frames présentant les mouvements les plus importants de chaque signe, garantissant ainsi une représentation complète de la vidéo du glose (signe). Après avoir comparé les résultats obtenus avec les deux méthodes, nous avons conclu que cette deuxième méthode du maximum était plus efficace.

Par conséquent, nous avons pris la décision de travailler exclusivement sur cette deuxième méthode en utilisant 25 frames au lieu de 30. Nous avons constaté que réduire le nombre de frames n'affectait pas les résultats, ce qui nous a conduit à privilégier une approche plus concise et efficace.

Figure 27. Extraction des séquences d'images à partir d'une vidéo

Tableau 2. Les résultats d'extraction d'images

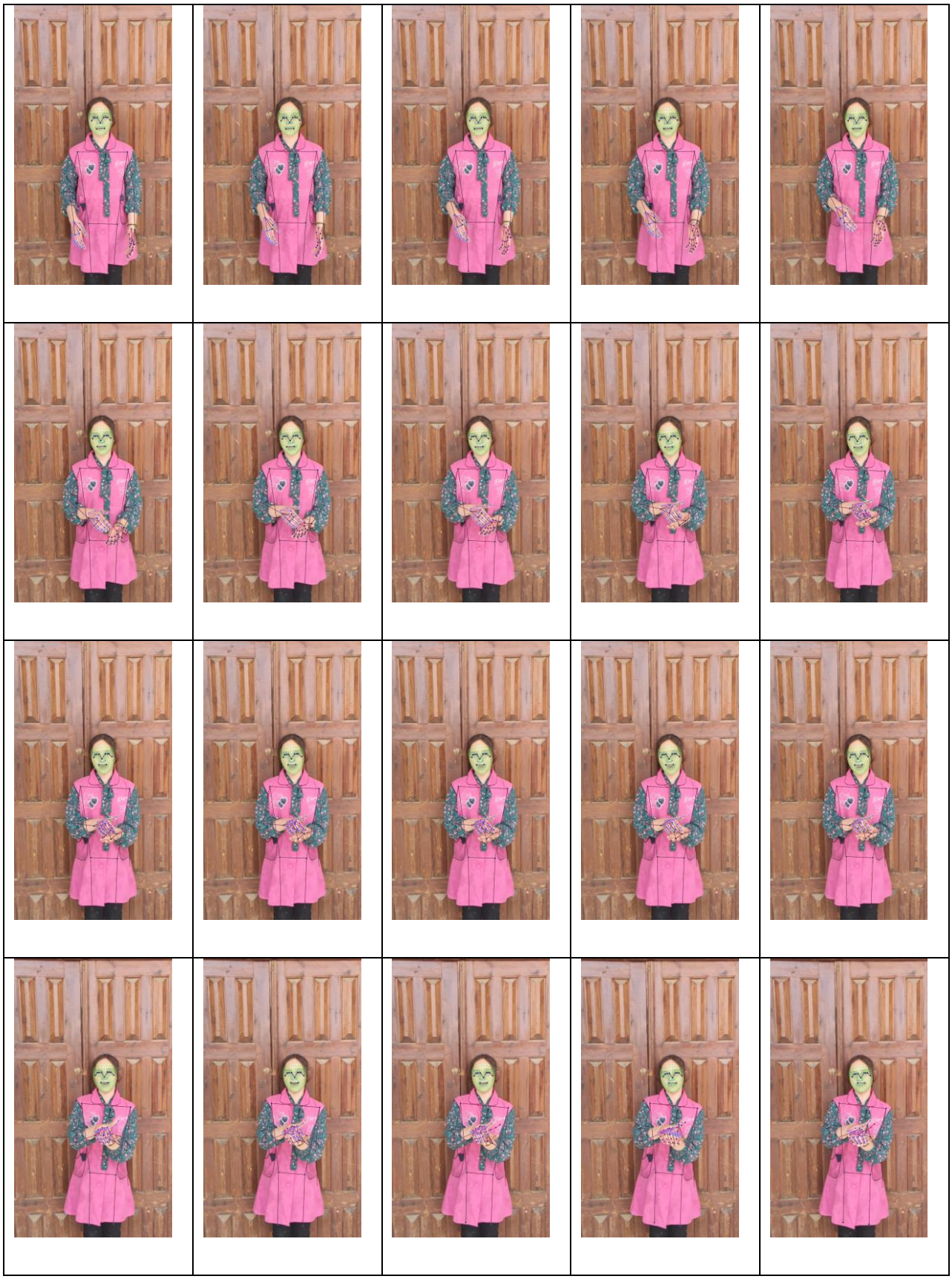




2.3.2 Extraction des points d'intérêt (keypoints)

L'équipe d'intelligence artificielle de Google a créé la compétition « Sign Language Medipipe » sur Kaggle février 2023, cette compétition a été une expérience transformative pour nous, c'est à ce moment-là que nous avons découvert l'utilisation de MediaPipe et nous avons été impressionnés par les résultats obtenus. L'efficacité, la flexibilité et la puissance de cette plateforme innovante développée par Google étaient évidentes, ce qui nous a incités à explorer davantage cette technologie et à l'incorporer dans notre propre projet.

Tableau 3. Les résultats du Mediapipe



Après avoir préparé la base de données dynamique, les données d'entrée ont été converties au format npy (NumPy array) et nous avons choisi deux types de réseaux différents pour comparer les résultats à la fin.

Pour le réseau LSTM :

Nombre de couches : Le réseau LSTM est composé de 3 couches LSTM empilées les unes sur les autres.

Fonction d'activation : La fonction d'activation 'relu' est utilisée pour les couches entièrement connectées (fully connected layers), et la fonction d'activation 'softmax' est utilisée pour la couche de sortie.

Pour le réseau ResNet :

Nombre de couches : Le réseau ResNet est composé de 3 blocs résiduels identiques. Chaque bloc résiduel est composé de deux couches de convolution.

Fonction d'activation : La fonction d'activation 'relu' est utilisée pour les couches de convolution, et la fonction d'activation 'softmax' est utilisée pour la couche de sortie.

2.4 Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes de préparation d'une base de données et les structures des réseaux de neurones utilisées. Dans le prochain chapitre, nous analyserons les résultats obtenus lors de la partie pratique de notre travail.

Chapitre 3

Résultats de Classifications par DL

Chapitre 3

Résultats de Classifications par DL

3.1 Introduction

Le chapitre précédent présentait deux types de bases de données utilisées dans l'étude : un ensemble d'images statiques et une base de données dynamique composée de vidéos. Nous avons utilisé MediaPipe pour extraire des séquences d'images à partir des vidéos.

Dans ce chapitre, nous exposons les résultats de notre travail en utilisant ces bases de données. Nous avons utilisé le réseau de neurones convolutifs (CNN) pour l'analyse des images statiques au format JPG.

Après avoir préparé la base de données dynamique, nous avons choisi deux réseaux ResNet de six couches et LSTM de trois couches. Ce chapitre a permis de détailler les performances et les résultats obtenus avec ces réseaux sur les bases de données utilisées. Nous avons également examiné l'architecture des réseaux ResNet et LSTM pour mieux comprendre leur fonctionnement et leur impact sur les résultats de notre étude.

3.2 Environnement matériel

La réalisation de notre projet s’est effectuée sur une station de calcul, relier à une connexion internet haut débit, doter d’un Processeur CPU **@** 2.10GHz, d’une mémoire vive Jusqu’à 384 Go de DDR4, 2 666 MHz, RDIMM (12 emplacements DIMM). Le tout sur une architecture 64 bits, processeur x64 du système d’exploitation Windows 10 professionnelle. La station de calcul utilisée pour notre projet était un Lenovo ThinkStation P720.

3.3 Environnement logiciel

Nous avons utilisé Python comme langage de programmation, Anaconda Jupyter Notebook comme environnement cloud, PyTorch pour le modèle statique et TensorFlow pour le modèle dynamique.

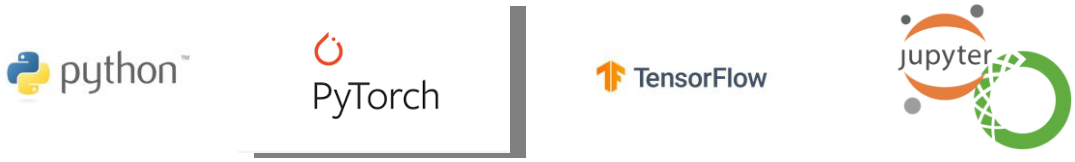


Figure 28. Les logiciels nécessaires pour notre projet

3.4 Résultats de classification Statique

Nous avons utilisé un CNN pour classifier les signes arabes à partir de notre base de données. Les images ont été mélangées aléatoirement et divisées en un ensemble d'apprentissage et un ensemble de test.

Dans cet ensemble d'apprentissage, Nous avons utilisé 70% des images, tandis que l'ensemble de test contenait les 30% restants des images. Cette division permet d'évaluer les performances de notre système sur des données qu'il n'a pas encore rencontrées, En plus de l'évaluation sur l'ensemble de test, nous avons testé le système avec de nouvelles images pour évaluer sa capacité à reconnaître les signes de la langue des signes arabe dans des situations réelles. Ces nouvelles images n'ont pas été utilisées lors de l'entraînement initial. Les tests avec les nouvelles images ont permis de mesurer la précision du système et sa capacité à classer correctement les signes dans des scénarios réels, tenant compte des variations de fond, d'éclairage et de pose. Ces résultats fournissent des informations importantes sur la généralisation du système et sa pertinence pour des applications pratiques d'interprétation des signes.

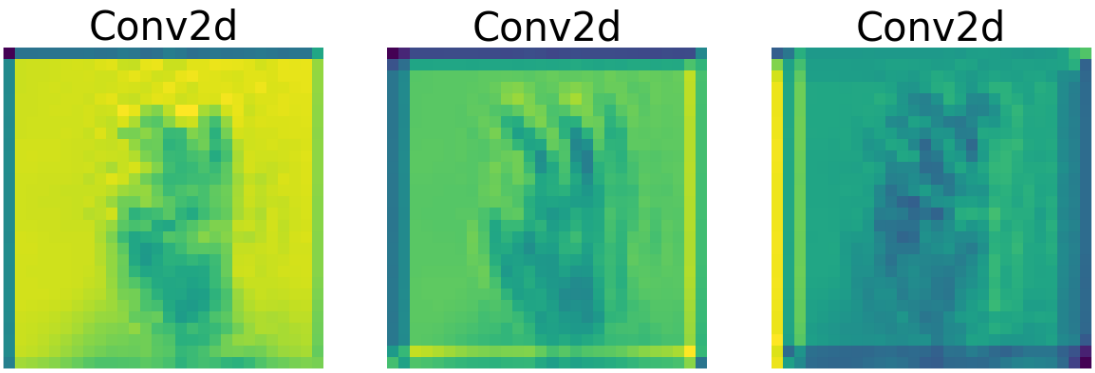


Figure 29. Couche de convolution après l'entraînement

Les images représentent les cartes de caractéristiques obtenues après l'entraînement du modèle CNN.

La figure 31 montre les courbes obtenues du taux de reconnaissance et de perte respectivement en fonction du nombre d'époques **(**50 époques**)** par notre modèle proposé**.**

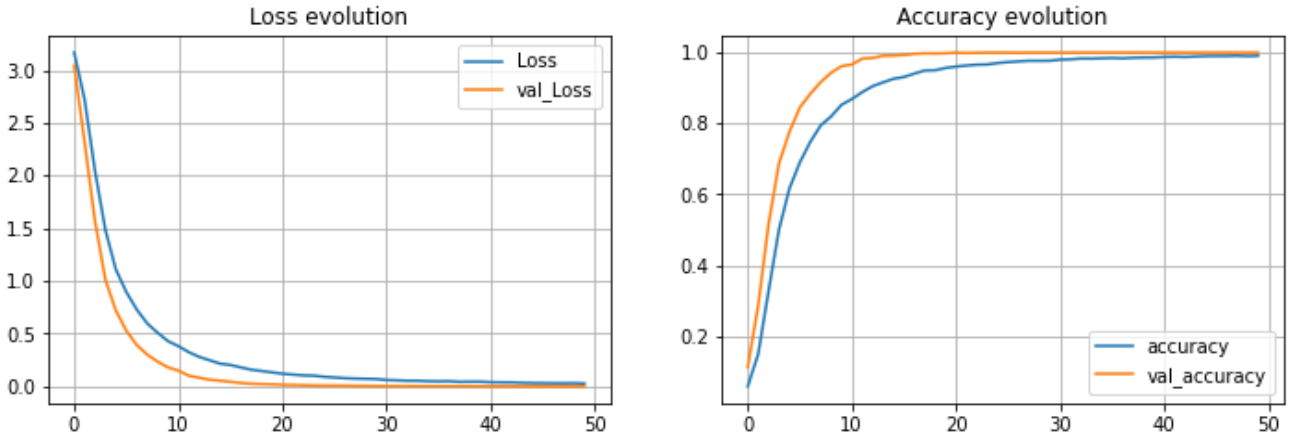


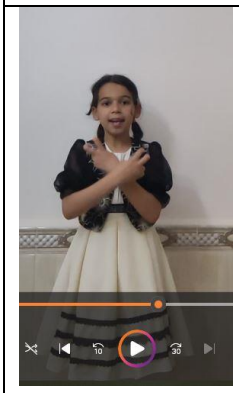
Figure 30. Taux d'erreur et de reconnaissance CNN

3.5 Résultats de classification Dynamique

3.5.1 Les résultats d'extraction d'image

Méthode 1

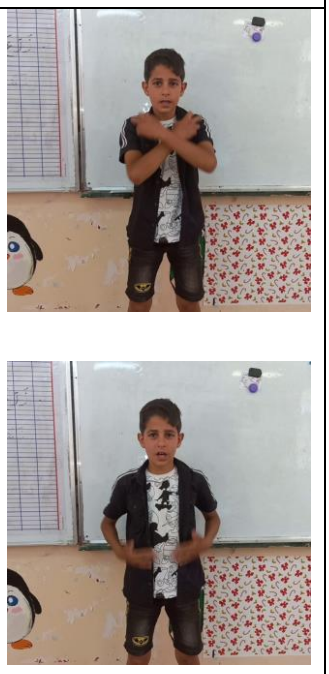
Nombre total dimages : 78

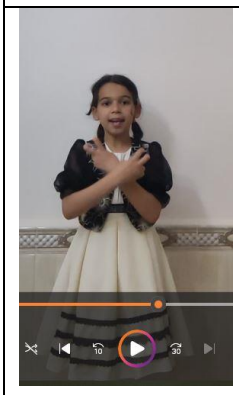


Nombre total d'images : 158

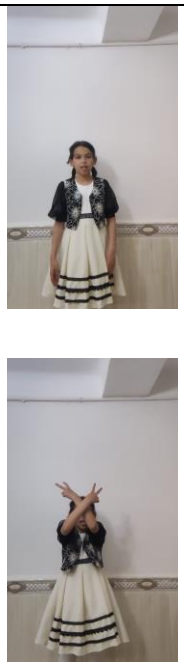
Méthode 2

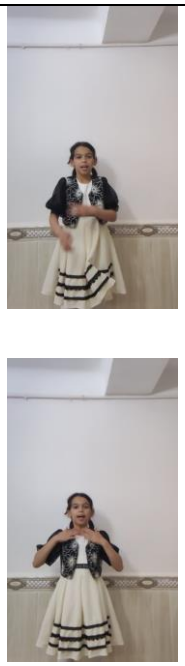
Nombre total d’images : 78

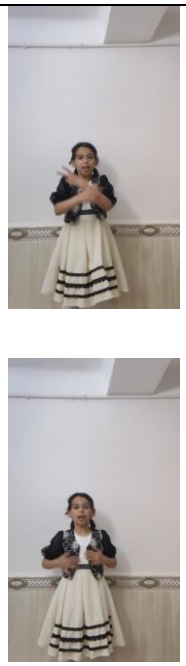




Nombre total d’images : 158







3.5.2 Les résultats du modèle

Nous avons mélangé de manière aléatoire les vidéos de notre base de données et les divisées en deux ensembles distincts : un ensemble d'apprentissage et un ensemble de test. Pour évaluer les performances de notre système, nous avons utilisé 95% des images dans l'ensemble d'apprentissage et 5% pour l'ensemble de test. En enregistrant le taux de reconnaissance obtenu à l'aide de ces ensembles, nous avons pu analyser les performances de notre système.

Les figures 32 et 33 montrent les courbes obtenues du taux de reconnaissance et de perte en fonction du nombre d'époques (400 époques) par le réseau LSTM et le ResNet. La Fig. 33 montre que notre réseau ResNet a obtenu de meilleurs résultats que le réseau LSTM car il s'agit d'un réseau pré-entraîné qui ont été formés sur de grandes quantités de données avant d'être utilisés pour une tâche spécifique.

Ainsi, ces figures offrent une visualisation des performances des deux architectures avec l'optimiseur Adam au fil des époques.

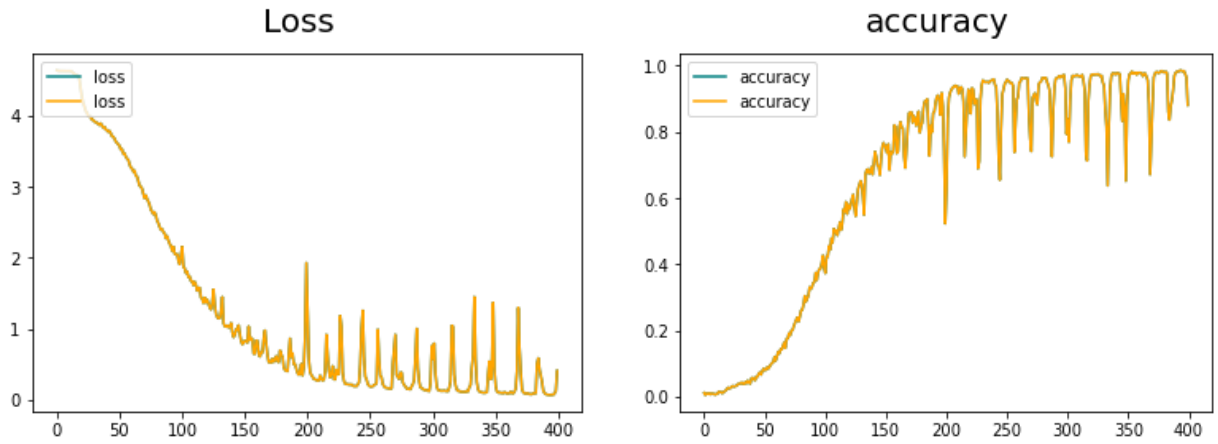


Figure 31. Taux d'erreur et de reconnaissance LSTM

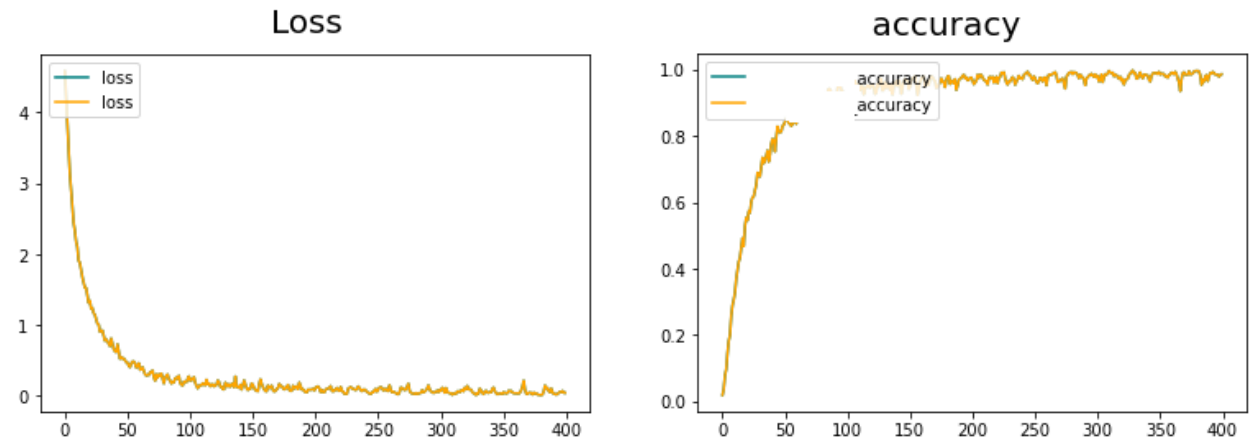


Figure 32. Taux d'erreur et de reconnaissance ResNet

Après avoir terminé l'entraînement de notre modèle, nous avons procédé à des tests pour évaluer ses performances. Nous avons d'abord effectué un test initial, nous avons sélectionné une seule classe (les animaux) de 20 exemples pour entraîner notre modèle. Les résultats obtenus étaient encourageants, ce qui montrait que notre réseau était capable de reconnaître et d'apprendre cette classe spécifique. Cependant, nous avons observé que notre modèle avait des difficultés avec des signes très similaires, tels que la fourmi et le ver, ainsi que le lapin et la vache, en raison de la similarité de leurs gestes respectifs. Il est important de souligner que les réseaux de neurones peuvent éprouver des difficultés lorsqu'il existe des similitudes subtiles entre différentes classes.

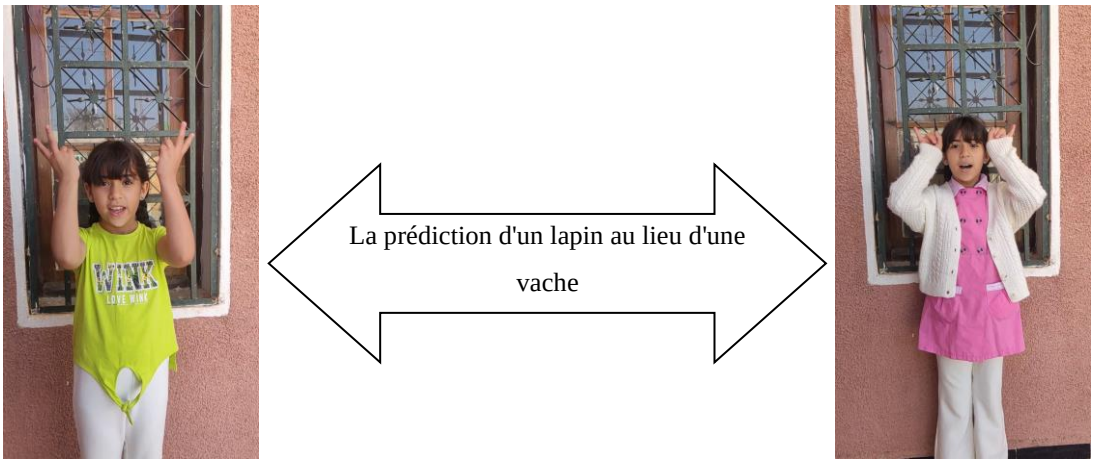


Figure 33. Exemple d'un test

Ensuite, nous avons réalisé un test final en utilisant la base de données complète. Pour ce faire, nous avons entraîné notre modèle sur 95% des vidéos de l'ensemble d'apprentissage, puis évalué ses performances sur les 5% restants de l'ensemble de test. Pour rendre ce test plus réaliste, nous avons également ajouté de nouvelles vidéos qui ne figuraient pas dans notre base de données initiale. Cela nous a permis de vérifier la capacité de généralisation de notre modèle sur des données inconnues et d'observer comment il se comportait face à des vidéos qu'il n'avait jamais rencontrées auparavant.

Tableau 4. Comparaison entre les différents types de réseaux de neurones utilisés

Classifieur	Taux de reconnaissance (Accuracy)	Taux d'erreur (Loss)	Nombre de couche	Type de réseau	Classification
CNN	98%	0. 032	3	Scratch	Statique
LSTM	88%	0.4	3	Scratch	Dynamique
ResNet	98%	0.04	6	TF	Dynamique

3.6 Conclusion

Dans ce chapitre nous avons fait un système de reconnaissance des lettres et des mots de la langue de signes algérienne basé sur les réseaux de neurones, les valeurs de taux d'erreur proches de zéro et une précision d'environ 95% sont des indicateurs positifs dans l'apprentissage automatique. Un taux d'erreur proche de zéro signifie que le modèle a minimisé les erreurs entre ses prédictions et les valeurs réelles. Une précision d'environ 95% indique que le modèle classe correctement la plupart des exemples. Ces résultats montrent que le modèle est bien entraîné, capable de capturer les motifs des données et de généraliser correctement

Conclusion Générale

Conclusion Générale

La langue des signes joue un rôle essentiel dans la réduction des barrières entre les personnes sourdes-muettes et les personnes entendantes. Cependant, il est difficile pour les personnes entendantes d'apprendre cette langue en raison de son utilisation moins fréquente et de la frénésie de leur vie quotidienne qui limite souvent le temps disponible pour son apprentissage. Heureusement, l'intelligence artificielle offre une solution prometteuse pour combler ce fossé en exploitant les connaissances des personnes maîtrisant la langue des signes afin de les transmettre à ceux qui ne la connaissent pas.

Dans le cadre de notre travail, nous avons utilisé trois types de réseaux différents. Tout d'abord, nous avons développé un système de reconnaissance des lettres du langage des signes arabe basé sur les CNN (Convolutional Neural Networks) avec une base de données de 54 049 images, et après avoir introduit une nouvelle image contenant un signe, le système le reconnaît, et la précision de la reconnaissance atteint environ 98%. Les résultats obtenus ont démontré l'efficacité de notre système dans la reconnaissance des lettres.

Nous avons également créé une base de données comprenant 2040 vidéos, cette base de données a été entraînée à l'aide de deux types de réseaux : le LSTM (Long Short-Term Memory) et le ResNet (Residual Neural Network). De plus, nous avons utilisé le framework Mediapipe pour suivre les mouvements des mains, des visages et du corps. Les résultats obtenus après l’entraînement par Le réseau ResNet plus efficaces que le LSTM.

Cependant, nous avons rencontré de nombreuses difficultés pour réaliser cette base de données. La permission de prendre des séquences vidéo avec les élèves du primaire ne nous a pas été accordée, y compris par l'école des sourds-muets. Par conséquent, nous avons dû recourir à des connaissances personnelles pour obtenir ces enregistrements.

Pour les futures recherches dans ce domaine, il serait intéressant d'élargir la portée de l'étude en incluant une tranche d'âge plus large. Notre étude s'est concentrée sur des personnes âgées de 10 à 23 ans, ce qui représente une tranche d'âge relativement restreinte. En incluant des participants

de différents groupes d'âge, nous pourrions obtenir une compréhension plus complète de l'apprentissage de la langue des signes et de son impact sur différentes populations.

Par ailleurs, il serait intéressant de capturer plusieurs signeurs simultanément dans une seule vidéo, ce qui permettrait d'économiser du temps et d'enrichir la base de données avec des exemples de conversations plus naturelles et dynamiques.

En élargissant la base de données et en explorant ces perspectives, nous pourrions renforcer les performances des modèles d'intelligence artificielle dans la reconnaissance et l'interprétation des signes, facilitant ainsi davantage la communication entre les personnes sourdes-muettes et les personnes entendantes. Ces avancées contribueraient à l'inclusion sociale et à la réduction des barrières linguistiques.

Bibliographie

[1] Brentari, Diane (1995) Sign language phonology, in J. Goldsmith (ed.), The Handbook of Phonological Theory, Cambridge, MA: Blackwell. 615-139

[2] E. Costello, American Sign Language Dictionary, Random House, New York, NY, USA, 2008.

[3] Delaporte, Y. (2007). Dictionnaire étymologique et historique de la langue des signes française. Origine et évolution de 1200 signes.

[4] [enligne].disponible:https://www.sense.org.uk/get-support/information-and advice/communication/sign-language/?msclkid=1ef67f4db46f11ec8ef5f3df9427194a. Consulter le 13/05/2023

[5] [enligne].disponible : https://www.nationalgeographic.com/history/history magazine/articl consulter le 13/05/2023

[6] A.Foltz.[enligne].disponible:https://theconversation.com/sign-languages-are-fully-fledgednatural-languages-with-their-own-dialects-they-need-protecting109388?msclkid=24cbb96ab47111ec9bb80bd45e99a487. Consulter le 16/06/2023

[7] [enligne]. disponible: https://www.ethnologue.com/language/asp. Consulter le 10/04/2023

[8] K. J. Shivani Kaushal, "2018 7th International Conference on Reliability," in Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2018.

[9] [en ligne]. disponible AI vs Machine Learning vs Deep Learning vs Data Science(1) AI vs Machine Learning vs Deep Learning vs Data Science - YouTube consulter le 16/06/2023

[10] Livre Introduction au traitement d'images février 2008, DIANE LINGRAND auteur,edition Vuibert

[11] Bouillot, R. (2005). Cours de traitement numérique de l'image. Dunod.

[12] [en ligne]. disponible Feature Selection Techniques in Machine Learning - GeeksforGeeks consulter le 16/06/2023

[13] Schmitt, A., Le Blanc, B., Corsini, M. M., Lafond, C., & Brůžek, J. (2001). Les réseaux de neurones artificiels. Un outil de traitement de données prometteur pour l'anthropologie. Bulletins et mémoires de la Société d'Anthropologie de Paris. BMSAP, 13(13 (1-2)).

DOI : https://doi.org/10.4000/bmsap.4463 consulter le 16/06/2023

[14] Lei, X,Pan, H, & Huang, X. (2019). A dilated CNN model for image classification. *IEEE Access*, *7*, 124087-124095.

[15] L'apprentissage profond Broché – Livre grand format, 18 octobre 2018, Florent Massot Eds, de Ian Goodfellow (Auteur), Yoshua Bengio (Auteur), Aaron Courville (Auteur), Francis Bach

[16] Student Notes: Convolutional Neural Networks (CNN) Introduction – Belajar Pembelajaran Mesin Indonesia (indoml.com) .Org GATE CS & IT 2024 (geeksforgeeks.org) consulter le 10/06/2023

[17] Olah, C. (2015). Understanding lstm networks.

[18] The Ultimate Guide to Building Your Own LSTM Models (projectpro.io) The Ultimate Guide to Building Your Own LSTM Models (projectpro.io) consulter le 16/06/2023

[19] Livre,Transfer Learning for Natural Language Processing Broché – 3 novembre 2021 Édition en Anglais de Paul Azunre (Auteur).

[20] Grishchenko, I. & Bazarevsky, V. Mediapipe holistic. Retrieved from Google AI Blog (googleblog.com) consulter le 16/06/2023

[21] MediaPipe Team GitHub - google/mediapipe: Cross-platform, customizable ML solutions for live and streaming media.consulter le 16/06/2023

[22] Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., & Presti, P. (2011, November). American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 279-286).

[23] [en ligne]. disponible Welcome to WLASL Homepage | WLASL (dxli94.github.io) consulter le 16/06/2023