Reconnaissance de la langue des signes à partir de vidéos

1. Introduction

La reconnaissance automatique de la langue des signes constitue un enjeu majeur pour l'accessibilité et l'inclusion des personnes sourdes et malentendantes. Elle permet de faciliter la communication entre les personnes pratiquant la langue des signes et le reste de la population, notamment à travers des interfaces homme–machine intelligentes.

Ce projet a pour objectif de mettre en œuvre un pipeline complet de reconnaissance de la langue des signes américaine (ASL) à partir de vidéos, en utilisant des techniques de Deep Learning. Le travail se concentre sur l'extraction de points clés (keypoints) des mains à l'aide de MediaPipe, puis sur la classification des signes à l'aide d'un réseau de neurones récurrent de type LSTM.

L'objectif principal n'est pas d'atteindre une performance maximale, mais de comprendre et d'implémenter l'ensemble des étapes nécessaires à un système de reconnaissance de gestes réaliste, à partir de données réelles et bruitées.

2. Présentation du dataset

Le dataset utilisé est WLASL v0.3 (Word-Level American Sign Language). Il s'agit d'un dataset de recherche largement utilisé dans la communauté scientifique pour l'étude de la reconnaissance de la langue des signes.

Caractéristiques principales :

- Vidéos issues de sources réelles (YouTube, ASLBrick, ASL SignBank, etc.)
- Forte variabilité des conditions d'acquisition (éclairage, angle de vue, résolution)
- Présence de vidéos corrompues ou incomplètes
- Déséquilibre important entre les classes

Dans le cadre de ce projet, seules les 30 classes les plus fréquentes ont été sélectionnées afin de limiter la complexité du problème. Chaque classe contient entre 11 et 18 instances, ce qui reste relativement faible pour un entraînement supervisé.

3. Prétraitement des données

3.1 Téléchargement des vidéos

Les métadonnées du dataset sont fournies sous forme de fichiers JSON. Un script Python a été développé afin de télécharger automatiquement les vidéos associées à chaque instance.

Certaines vidéos n'ont pas pu être téléchargées ou lues correctement (erreurs de type *moov atom not found*), ce qui reflète la nature non contrôlée et réaliste du dataset.

3.2 Extraction des keypoints

Pour chaque vidéo téléchargée, les points clés des mains sont extraits à l'aide de la bibliothèque MediaPipe Hands. Pour chaque frame, les coordonnées 3D des 21 points clés par main sont récupérées.

Les caractéristiques finales par frame sont constituées de :

- 2 mains × 21 points × 3 coordonnées = 126 features

Chaque vidéo est ensuite représentée comme une séquence temporelle de longueur fixe (60 frames). Les séquences plus courtes sont complétées par du padding, tandis que les séquences plus longues sont tronquées.

Les données finales sont sauvegardées sous forme de fichiers .npy afin d'accélérer l'entraînement du modèle.

4. Modèle proposé

4.1 Architecture

Le modèle utilisé est un réseau de neurones récurrent de type LSTM (Long Short-Term Memory), adapté au traitement de données séquentielles.

Architecture du modèle :

- Entrée : séquence de taille (60, 126)
- LSTM :
    - 1 couche
    - 256 unités cachées

- Dropout (0.5)
- Couche entièrement connectée (Fully Connected)
- Sortie : probabilité sur les 30 classes

Afin de mieux exploiter l'information temporelle, une moyenne temporelle (mean pooling) est appliquée sur les sorties du LSTM.

4.2 Fonction de perte et optimisation

- Fonction de perte : CrossEntropyLoss
- Optimiseur : Adam
- Taux d'apprentissage : 1e-3

5. Entraînement

L'entraînement est réalisé sur l'ensemble des données disponibles, avec un batch size de 32. Le modèle est entraîné pendant 20 époques.

En raison du temps limité et de la complexité du dataset, aucune stratégie avancée de validation croisée ou de rééquilibrage des classes n'a été mise en place.

6. Résultats

Les performances obtenues varient en fonction de la stratégie de pooling temporel utilisée.

- Accuracy observée : entre 2 % et 13 %

Ces résultats peuvent sembler faibles, mais ils sont cohérents avec :

- le faible nombre d'échantillons par classe,
- la forte variabilité des vidéos,
- le bruit introduit lors de l'extraction des keypoints,
- l'utilisation d'un modèle volontairement simple.

7. Discussion et limites

Plusieurs facteurs expliquent les performances obtenues :

- Dataset très bruité et non contrôlé
- Échecs fréquents de détection des mains par MediaPipe
- Déséquilibre important entre les classes
- Absence de pré-entraînement ou de data augmentation
- Architecture simple ne modélisant pas explicitement les relations spatiales

Cependant, ces limites mettent en évidence les défis réels de la reconnaissance de la langue des signes à partir de données du monde réel.

8. Perspectives d'amélioration

Plusieurs pistes pourraient être explorées pour améliorer les performances :

- Utilisation de modèles plus avancés (BiLSTM, GRU, Transformers)
- Intégration des points clés du visage et du corps
- Data augmentation sur les séquences de keypoints
- Rééquilibrage des classes
- Utilisation de modèles pré-entraînés sur des données vidéo

9. Conclusion

Ce projet a permis de mettre en œuvre un pipeline complet de reconnaissance de la langue des signes à partir de vidéos, allant du téléchargement des données jusqu'à l'entraînement d'un modèle de Deep Learning.

Malgré des performances modestes, le travail réalisé met en évidence la complexité du problème et les nombreux défis liés à l'utilisation de données réelles. Il constitue une base solide pour des travaux futurs plus avancés.