Rapport SAM
Modèle multi-modaux

Cléa Han, Yanis Labeyrie et Adrien Zabban

janvier 2024

1 Introduction

2 Données

parler des données (d'où elles viennent) dire que les vidéos sont déjà prétraités (analyse des landmarks) parler (et redéfinir) les IPUs comment sont agencées les données parler des problèmes (temps de chargement du csv) et comment on l'a résolu : avec des skiprows

3 Traitement unimodale

3.1 Traitement du texte

Pour parvenir à détecter le changement de tour de parole dans les données textuel, nous avons choisis d'utiliser un modèle considéré comme l'état de l'art pour la tâche de classification de texte : le modèle BERT. Le modèle BERT (Bidirectional Encoder Representations from Transformers) offre une utilité significative dans la prédiction et la détection des changements de tour de parole dans un script de conversation. En raison de sa capacité à comprendre le contexte bidirectionnel dans lequel les mots apparaissent, BERT excelle dans la capture des nuances sémantiques et des relations complexes entre les différents segments de texte. Lorsqu'il est appliqué à l'analyse de scripts de conversations, BERT peut saisir les subtiles transitions entre les locuteurs, en tenant compte des indices contextuels tels que le ton, le style et les références antérieures. Cette capacité à contextualiser l'information permet à BERT d'améliorer la précision de la prédiction des changements de tour de parole, offrant ainsi une méthode robuste pour l'analyse automatisée des dialogues et des échanges conversationnels.

Par ailleurs, BERT étant un modèle comprenant un nombre important de paramètres (environ 100 million) il est innenvisageable avec nos moyens de l'entraîner "à partir de 0". Nous avons donc opter pour l'utilisation d'un modèle pré-entrainé sur la langue Française appelé CamemBERT. En utilisant un modèle pré-entraîné sur la langue française tel que CamemBERT, qui a déjà appris des représentations linguistiques riches, nous capitalisons sur cette connaissance préalable pour la tâche spécifique de détection des tours de parole. Le fine-tuning permet d'adapter le modèle à la tâche spécifique tout en conservant les caractéristiques générales apprises lors du pré-entraînement, améliorant ainsi la

performance du modèle pour la tâche de prédiction des tours de parole dans le contexte d'un script de conversation en français. Cette approche hybride permet de tirer parti des avantages du pré-entraînement tout en adaptant le modèle à la spécificité de la tâche souhaitée.

Pour ce faire nous importons le modèle pré-entrainé à partir de la platerforme HuggingFace, et nous changeons uniquement la couche finale du modèle pour l'adapter à notre tâche de classification binaire (changement de parole/pas changement de parole).

3.2 Traitement de l'audio

Afin de détecter les changements de parole dans des données audio, nous avons choisi d'exploiter une approche similaire en utilisant un modèle avancé dans le domaine de la représentation audio : le modèle Wave2Vec. Wave2Vec, basé sur les Transformers, s'est établi comme une référence pour la tâche de traitement du signal audio, en particulier pour la détection des variations dans la parole. Grâce à sa capacité à encoder de manière bidirectionnelle les représentations des signaux sonores, Wave2Vec excelle dans la compréhension des nuances acoustiques et des transitions subtiles entre les locuteurs. En utilisant ce modèle dans l'analyse des données audio, Wave2Vec peut capturer les changements de parole en tenant compte des indices contextuels tels que l'intonation, le rythme et les caractéristiques acoustiques spécifiques à chaque locuteur. En raison du nombre considérable de paramètres dans Wave2Vec, une approche pragmatique consiste à utiliser un modèle pré-entraîné spécifique à la tâche. Nous avons ainsi choisi d'adopter un modèle pré-entraîné sur des données audio en , capitalisant sur ses connaissances préalables pour la détection des changements de tour de parole. En procédant au fine-tuning de ce modèle, en ajustant uniquement la couche finale pour la classification binaire (changement de parole/pas changement de parole), nous adaptons efficacement Wave2Vec à notre objectif spécifique, offrant ainsi une méthode robuste pour l'analyse automatisée des variations vocales dans un contexte audio francophone.

parler de en quelle langue est wave2vec

3.3 Traitement de la vidéo

Pour traiter les données issues de la vidéo qui sont sous la forme des landmarks des visage des deux personnes en conversations, nous avons choisis d'utiliser un réseau de neurone adapté à cette tâche d'analyse de série temporelle : le réseau LSTM (Long-Short Term Memory). En effet ce modèle est pertinant pour la détection du changement de locuteur car il est capable de conserver des informations sur de longues séquences temporelles, ce qui est essentiel pour analyser les variations subtiles dans les mouvements des landmarks faciaux. Les réseaux LSTM présentent une architecture spécifique qui permet de gérer les dépendances temporelles complexes dans les données vidéo, tout en évitant les problèmes de disparition du gradient associés aux réseaux de neurones classiques. En utilisant un LSTM, nous pouvons exploiter la mémoire à long terme du modèle pour suivre les patterns caractéristiques des changements de tour de parole, qui peuvent être reflétés dans les déplacements et expressions faciales des interlocuteurs.

4 Traitement multimodale

Notre approche vise à maximiser l'utilisation des dépendances et complémentarités entre différents types de données (vidéo, texte, audio). Pour effectuer des prédictions en utilisant ces modalités diverses, nous avons opté pour l'utilisation de chaque réseau de neurones présenté précédemment afin d'extraire des représentations compressées (ou features) des différents types de données. Pour fusionner ces informations, nous employons une couche dense qui prend en entrée ces trois représentations latentes, produisant en sortie une classification binaire (changement de parole/pas de changement de parole). Cette démarche est illustrée dans la figure suivante :

Insérer ici un schéma illustrant la fusion latente

Les modèles individuels sont initialement pré-entraînés, puis fusionnés en une seule architecture multimodale qui subit ensuite un nouvel entraînement. Le pré-entraînement des modèles individuels présente l'avantage de réduire le temps de convergence de l'entraînement de cette architecture multimodale comprenant un nombre important de paramètres.

4.1 Maximum de vraisemblance

on prend la solution la plus prédite

4.2 Apprentissage basique

on enlève la dernière couche dense de chaque modèle et on concatène tout et on refait passer dans une couche dense (on utilise les entrainements déjà faits) en gelant les poids précédant

4.3 Entraînement du gros modèle

on ne gèle pas les poids

5 Conclusion