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 avec CamemBERT

Pour parvenir à détecter le changement de tour de parole dans les données textuel, nous avons choisi 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. Par ailleurs, BERT étant un modèle comprenant un nombre important de paramètres (environ 100 millions) il est inenvisageable avec nos moyens de l'entraîner "à partir de 0". Nous avons donc opté pour l'utilisation d'un modèle pré-entrainé sur la langue Française appelé CamemBERT et nous avons gelé les poids de CamemBERT.

On prend alors les données de textes qui sont une liste d'entiers, correspondant à l'indice de chaque mot du tokenizer de CamemBERT. Ces données passent alors dans CamemBERT, elles ressortent avec une dimension de 768. Elles passent alors dans une activation ReLU puis une couche dense de taille 768, puis une couche de dropout (avec un taux d'oublie de 10%), pour enfin passer un dernier ReLU et une dernière couche dense à 2 dimensions.

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.

On possède deux données en entrée qui sont les enregistrements audio des deux interlocuteurs. On fait alors passer leurs audio dans Wave2Vec puis l'on concatène la sortie. On fait alors passer la concaténation dans une couche dense pour n'avoir qu'une sortie de taille 2.

3.3 Traitement de la vidéo

Pour traiter les données issues de la vidéo qui sont sous la forme des landmarks du visage des deux personnes en conversations, nous avons choisi 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 pertinent 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.

Comme pour l'audio, la vidéo contient un ensemble de deux fois 10 images pour chacun des interlocuteurs. On les fait alors passer dans deux couches LSTM et on concatène les sorties pour ensuite les faire passer dans du relu et une couche dense de taille 2.

4 Traitement multimodal

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 diverses modalités, 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.

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.

Au cours de ce processus, pour des raisons de contraintes de ressources, on gèle les poids des modèles unis-modale.

4.1 Maximum de vraisemblance

Ce premier modèle multimodal fait passer chaque donnée dans son modèle pré-entraîné associé et fait une moyenne des sorties. Le but est alors de faire une moyenne des probabilités selon chaque canal. Nous appellerons ce modèle *LIKELIHOOD* du fait que le but est de maximiser la log-likelihood.

4.2 Apprentissage basique

Ce deuxième modèle plus évolué consiste à faire passer chaque donnée dans son modèle pré-entraîné associé et de récupérer l'information avant qu'elle ne

passe dans la dernière couche dense (celle de taille deux). On concatène alors toute l'information et l'on la fait passer dans une grande couche dense de taille 2.

5 Métriques

6 Résultats

7 Conclusion