DUJARDIN Thomas

ENSAE 2e année

Stage d'application

Année 2021-2022

Deep Learning pour la reconnaissance multimodale d'émotions dans des extraits vidéos

Note de synthèse (français)

MICS-IECL

Gif-sur-Yvette, 91190

Marianne CLAUSEL et Myriam TAMI

Du 01/06/2022 au 30/09/2022

Grâce à M. Chopin, qui proposait des sujets de stage orientés recherche au début du second semestre de l'année dernière, j'ai pu être mis en contact avec Mmes Marianne Clausel et Myriam Tami, des laboratoires Mathematics and Informatics (MICS) de CentraleSupélec et de l'Institut Elie-Cartan de Lorraine respectivement, qui m'ont proposé un stage de trois mois fort intéressant en apprentissage statistique appliqué à la détection d'émotions, et qui l'ont encadré. J'ai ainsi passé les trois mois de ce stage dans les locaux du laboratoire MICS, dans un open space dédié aux stagiaires. J'avais régulièrement des points d'avancement et de cadrage avec mes encadrantes, pour qu'elles puissent décrire les différentes étapes de mon stage, et m'assister dans le déroulement de celles-ci.

La problématique était la suivante : à partir d'une base de données recensant des extraits vidéo de personnes s'exprimant sous le coup d'une émotion – donc à partir de pistes audio, vidéo, et de la transcription des paroles des interlocuteurs présents – il me fallait prédire l'émotion présente, parmi un ensemble de quelques émotions basiques (joie, colère, tristesse, …). L'intérêt de l'utilisation de plusieurs types de données est d'inférer plus d'informations qu'avec une seule modalité. Si, par exemple, l'on se retrouvait dans une situation où un personnage disait quelque chose dont l'émotion dominante est « joie », mais que la vidéo montrait un visage en colère ou grimaçant, on accèderait à un supplément d'information qu'aucune des modalités prises séparément n'arriverait à prédire : le personnage ment, ou est ironique. Ainsi que me l'ont recommandé mes encadrantes, j'ai procédé très progressivement, n'étant pas, au début de ce stage, très familier avec les méthodes d'apprentissage statistiques utilisées dans les principaux de papier de recherche traitant de cette problématique.

En premier lieu, je me suis remis en mémoire tout ce que j'avais pu apprendre sur le deep learning durant ma 2^{e} année, lors de mes cours et lors des quelques MOOCs que j'ai effectuées en parallèle, en utilisant quelques ressources données par mes encadrantes. Ensuite, j'ai commencé à constituer une bibliographie commentée de tout ce qui concerne l'apprentissage statistique appliqué à la détection d'émotions. Mes premières recherches ont rapidement montré que de tels modèles se décomposent systématiquement en deux grandes parties. La première d'entre elles consiste à trouver la meilleure façon possible de représenter une « modalité de donnée » (son, vidéo, texte, …) par un vecteur de nombres que la machine est capable d'interpréter. Il me fallait donc chercher, pour chaque modalité, les algorithmes issus de la théorie du deep learning capables de produire les meilleurs vecteurs possibles, c'est-à-dire les vecteurs qui représentent le mieux les données, en encapsulant le plus d'informations possible sur ce qu'ils représentent et sur ce que leurs « voisins » représentent

(par exemple, lorsqu'on encode une phrase par des vecteurs, il est nécessaire que ces vecteurs soient traités avec une notion d'« ordre », sinon on perd l'information sur l'ordre des mots durant cette numérisation), tout en étant suffisamment compacts pour être interprété par la machine. Pour trouver ces algorithmes, j'ai principalement utilisé le site Web « Paperswithcode » de Meta, qui propose des comparaisons chiffrées sur des tâches diverses (les vecteurs représentatifs étant produits par les algorithmes lorsqu'ils tentent de résoudre ces tâches), et j'ai demandé à mes encadrantes ce qu'elles en pensaient. Les algorithmes trouvés utilisent, dans les trois modalités, des « transformers », architectures de deep learning très populaires actuellement grâce au niveau inédit de représentativité que les vecteurs qui en sont issus offrent. Les « transformers » utilisent des « modules d'attention », qui résolvent un des problèmes que posaient les algorithmes utilisés auparavant : lorsque les données passées en entrée étaient trop longues (par exemple, une phrase longue avec beaucoup de mots entre un nom et un adjectif le qualifiant), des liens pertinents n'étaient pas faits à cause du problème du « vanishing gradient », qui empêchait de mettre en lien des données trop éloignées temporellement.

La seconde d'entre elles consiste à trouver la meilleure manière de fusionner ces vecteurs représentatifs, afin de produire une représentation qui incorpore de la manière la plus pertinente possible les éléments issus des différentes modalités. Ce sont les papiers de recherche dans le domaine du deep learning appliqué à la reconnaissance d'émotions qui m'ont aidé à comprendre cette partie. Ces papiers mettent en évidence que la simple mise bout à bout des vecteurs représentants chacune des modalités était déjà une manière assez pertinente et efficace de le faire, mais que la meilleure façon de le faire consiste à utiliser des architectures assez complexes à base de « transformers », en pondérant par priorité chacune des modalités, et en accordant la plus grande priorité aux vecteurs de nombres issus du texte, puisque le texte est la modalité la plus précise dans la détermination de l'émotion.

J'ai ainsi recensé dans le détail ces différents éléments dans des présentations LaTeX que j'ai exposées à mes encadrantes, ce qui m'a amené à la deuxième partie de ce stage : l'implémentation d'un modèle de prédiction d'émotions à l'aide du langage Python et du paradigme de programmation PyTorch, conçu spécialement pour le deep learning. Celle-ci a occupé plus de la moitié de mon stage. J'ai essayé d'implémenter le modèle « M2FNet » (Multi-modal Fusion Network), ce qui n'a pas été évident compte tenu de l'absence d'implémentation mise à disposition par les auteurs du papier. J'ai néanmoins pu produire une version simplifiée de ce modèle, avec quelques modifications que j'ai effectuées en

m’appuyant sur ma bibliographie. Ce modèle m’a donné environ 35 % de prédictions correctes sur un ensemble de vidéos que mon algorithme n’avait jamais vu. Ce résultat est assez décevant, puisqu’il est bien en dessous de ce qui est décrit dans le papier, mais cela s’explique à la fois par la courte durée de ce stage qui ne m’a pas permis de pousser plus loin mon implémentation, et ma simplification du modèle, faute de compréhension du papier.

Pour autant, mes travaux n’ont pas été vains, et la bibliographie commentée ainsi que la tentative d’implémentation seront, selon mes encadrantes, donnés à de futurs stagiaires, voire à de futurs doctorants. De plus, j’ai appris beaucoup de choses en deep learning qui me seront fort utiles durant ma 3^{e} année, et pour le reste de mon parcours scolaire et professionnel.