DUJARDIN Thomas

ENSAE 2nd year

Practical internship

Class of 2021-2022

Deep Learning for multimodal emotion recognition in video clips

Summary (English)

MICS-IECL

Gif-sur-Yvette, 91190

Marianne CLAUSEL and Myriam TAMI

01/06/2022 - 30/09/2022

Thanks to Mr. Chopin, who proposed research-oriented internship topics at the beginning of the second semester of last year, I was put in contact with Ms. Myriam Tami and Ms. Marianne Clausel, from the Mathematics and Informatics (MICS) laboratories of CentraleSupélec and the Elie-Cartan Institute of Lorraine, respectively, who offered me a very interesting three-month internship in statistical learning applied to emotion detection, and who supervised it. I spent the three months of this internship in the MICS laboratory, in an open space dedicated to interns. I had regular meetings with my supervisors, so that they could describe the different stages of my internship, and assist me in the development of these stages.

The problem was the following: from a database listing video excerpts of two people talking, along with their audio and the text transcription, I had to predict the emotion present at each moment, among a set of some basic emotions (joy, anger, sadness, ...). The interest of using several types of data is to infer more information than with a single modality. If, for example, we find ourselves in a situation where a character says something whose dominant emotion is "joy", but the video shows an angry face, we would gain access to additional information that none of the modalities taken separately would be able to predict: the character is lying, or is being ironic. Studying the history of the conversation (before the moment when we try to predict the emotion) also allows us to infer more information about the predominant emotions at each moment. As recommended by my supervisors, I proceeded very gradually, not being, at the beginning of this internship, very familiar with the statistical learning methods used in the main research papers dealing with this issue.

First, I remembered everything I had learned about deep learning during my 2nd year, during my courses and during the few MOOCs I did in parallel, using some resources given by my supervisors. Then, I started to build a commented bibliography of everything related to statistical learning applied to emotion detection. My first research quickly showed that such models systematically break down into two main parts. The first one consists in finding the best possible way to represent a "data modality" (sound, video, text, ...) through time, by a vector of numbers that the machine is able to interpret. I therefore had to look for algorithms, for each modality, based on the theory of deep learning, capable of producing the best possible vectors, i.e. vectors that best represent the data, by encapsulating as much information as possible on what they represent and on what their "neighbors" represent. For example when encoding a sentence by vectors, it is necessary that these vectors be processed with a notion of "order", otherwise we lose the information given by the words order during

this digitization. These vectors should also be small enough to be used by the machine. To find these algorithms, I mainly used Meta's "Paperswithcode" website, which offers numerical comparisons on various tasks (the representative vectors being produced by the algorithms as they attempt to solve these tasks), and I asked my supervisors what they thought of them. The algorithms found use, in all three modalities, "transformers", deep learning architectures that are currently very popular thanks to the unprecedented level of representativeness that the vectors produced by them offer. The transformers use "attention modules", which solve one of the problems of the previous algorithms: when the input data was too long, relevant links were not made because of the "vanishing gradient" problem. For example, in a long sentence with many words between a noun and an adjective qualifying it, it was difficult to establish that this noun and this adjective were related. In our case, this problem made it hard to link elements that were too far apart in time in a video clip.

The second one is to find the best way to merge these representative vectors, in order to produce a representation that incorporates in the most relevant way possible the elements coming from the different modalities. The research papers in the field of deep learning applied to emotion recognition helped me with this part. These papers show that simply putting end to end the vectors representing each of the modalities was already a rather relevant and efficient way to do it, but that the best way to do it was to use rather complex architectures based on "transformers", weighting by priority each of the modalities, and giving the highest priority to the vectors of numbers coming from the text, since the text is the most precise modality in the determination of the emotion.

I then listed in detail these different elements in presentations that I exposed to my supervisors, which led me to the second part of this internship: the implementation of an emotion prediction model using the Python language and the PyTorch programming paradigm, designed specifically for deep learning. This one occupied more than half of my internship. I tried to implement the "M2FNet" (Multi-modal Fusion Network) model, which was not an easy task given the lack of implementation made available by the authors of the paper. Nevertheless, I was able to produce a simplified version of this model, with some modifications that I made based on my bibliography. This model gave me about 35% of correct predictions on a set of videos that my algorithm had never seen. This result is quite disappointing, since it is far below what is described in the paper, but this is explained both by the short duration of this internship which did not allow me to push my implementation further, and my simplification of the model, due to my lack of understanding of the paper.

However, my work was not in vain, and the commented bibliography as well as the implementation attempt will be, according to my supervisors, given to future interns, or even to future PhD students. Moreover, I learned a lot of things in deep learning that will be very useful during my 3rd year, and for the rest of my academic and professional career.