CentraleSupélec

université PARIS-SACLAY

ST7 Project Specification

Authors:

Daniel Stulberg Huf
Lawson Oliveira Lima
Luan Rocha do Amaral
Lucas Vitoriano de Queiroz Lira
Rajeh El Saddi
Tomas Gonzalez Villaroel

Supervisor:

Simon Leglaive

February 27, 2023

Contents

Chapter 1

Introduction

1.1 Context and Issues

It is very well-known that humans are social animals whose relationships are strongly influenced by emotions. Until a few decades ago, it might have seemed unlikely that we could decipher and automatically analyze an individual's emotion, especially one related to non-verbal behaviors. Today, thanks to the progress in the fields of technology, psychology, medicine, among others, us humans can become more proficient in deciphering non-verbal behaviors, allowing for more accurate and efficient communication between humans and machines.

In this context, our group was assigned to participate in an international challenge of research on the automatic analysis of human behaviors. In order to help us in this task, we will work with IEMOCAP, the Interactive emotional motion capture database [1]. Created by the University of Southern California, such database comprehends thousands of labeled data concerning emotions in both audio, text, and motion capture of several communication scenarios performed by ten actors on dyadic sessions. Differently from other conventional databases that only contain one speech channel, IEMOCAP can be extremely useful when dealing with non-verbal multi-modal emotions.

After gathering the data provided by IEMOCAP, we will be responsible for preprocessing it, and then building and testing a model that will automatically extract human emotions according to certain details, which are specified below.

Chapter 2

Objectives

2.1 Goals

In order to fulfill the requirements specified in the contexts and also face the issues that accompany him, our group has decided to focus both on audio and text representations aimed at detecting emotions in discrete categories. That means that in the end we will have created an algorithm that is hopefully capable of classifying an emotion based on the sound and transcription of someone's speech. In order to perform such task, we have initially split in 3 fronts: audio processing, text processing, and AI modeling.

2.2 Processing

2.2.1 Audio processing

For the sound processing, our goal is to transform the wave form of someone's speech into an easier representation for performing a classification algorithm. As we already have the content of the speech, our processing will focus on non-verbal characteristics.

The post-processing data consists on the log-mel-spectogram of the speech, which is a non-linear representation to simulate the way human beings understand and process the sound in our brains [3]. After producing the log-mel-spectogram of each audio record, the final length of each result will be unified and then converted to an image, which will be the input for a convolutional neural network (CNN).

2.2.2 Text processing

For the text processing part our goals are divided into two main ideas:

- Extract and store the data from several Forced Alignment files into different csv files, divided into training and validation datasets in order to feed them to the AI model. Each of these files will have 3 columns: emotion, path and words. Note that we have to remove all rows of "xxx" emotion.
- Process the words using techniques such as word lemmatization, removing stop words, punctuation, etc [2]. After that, the text data will be ready to build the AI model.

We will start by testing the accuracy of the AI model with the processed words as input, and based on that, we will then decide whether or not to check the notion of time of the sequences, or to add "reaction words" to the datasets such as <laugh> or <sil>. We also need to adjust the class imbalance in order to have approximately the same number of data words in each class emotion.

2.2.3 AI Model

The artificial intelligence part is divided into two sub-models, which will be combined later to make a more general model that uses audio and text input. That said, the first sub-model to be built corresponds to the audio classification, which will be built using a CNN, due to the large number of architectures and parameters available for transfer learning. Moreover, after a literature search, we see that such architecture has great results when combined with log-mel spectrogram.

The second sub-model will deal with the classification of the texts. Therefore, we chose to use Bert, a state-of-the-art model based on Transformers to deal with NLP problems. This model has been chosen because of its big precision and speed when dealing with such kind of problem.

After building the two submodels, we will freeze all parameters of both, and build a more general architecture that has as input the concatenation of audio and text. This model will process each input independently, but at the end, we will add a fully connected layer to do the classification.

The work being done, we will additionally study how to improve the accuracy obtained by varying the hyper parameters.

Bibliography

1. BUSSO, Carlos et al. IEMOCAP: Interactive emotional dyadic motion capture database. **Language resources and evaluation**, v. 42, p. 335-359, 2008.

2. HARSHITH. 2022. Text preprocessing in Natural Language Processing using python, Medium. **Towards Data Science**. Available at: <https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8>. Accessed: February 27, 2023.

3. ROBERTS, Leland. 2020. Understanding the Mel Spectogram. **Medium**. Available at: <https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53>. Accessed: February 27, 2023.