EE599 Deep Learning – Revised Project Proposal

April 19, 2020

Project Title: Multimodal sentiment analysis

Project Team: Anjana Niranjan, Po Yu Yang, Prajakta Karandikar

Project Summary: Sentiment analysis is the interpretation and classification of emotions. Emotions can be analysed by the words a person speaks, the tone, the facial expressions and so on. By combining voice modulations and facial expressions to the text obtained from a person's speech, the feature learning process can be enriched. In this project, we propose sentiment analysis using text and audio. We will be comparing the accuracy of a model which analyzes the sentiment based on text only with one that analyzes the sentiment based on multimodal features. We will also compare different architectures for the implementation based on factors like speed, accuracy and size of the model.

Data Needs and Acquisition Plan: The data required to train our models are audio and the corresponding text. So we are looking at using a dataset of audios, which we can convert to text. We are going to be working with audio files that have only the English language. The dataset we will be using is this. It it consists of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.

Primary References and Codebase: We have referred to the following sources to get an understanding -

- Blog Posts: Introduction to hate speech detection, Multimodal sentiment classification, Deep NLP for hate speech detection
- GitHub codebases: Audio to text conversion, Offensive text detection with GRU, Hate speech detection with LSTM, Audio sentiment analysis

Architecture Investigation Plan: We plan to utilize the above codebases with some modifications in the following ways:

Data preparation - we will make use of the speech to text conversion method as referenced in the codebase above, to get text data from the audio files.

Model with unimodal information - the text data will be used as input (unimodal) for sentiment analysis and an estimate of the performance will be made.

Model with multimodal information - both text and audio data as input to the model for the analysis, thereby expanding the input space. The performance of model will thereafter be observed.

We will then compare the unimodal and multimodal approaches for their performance.

Estimated Compute Needs: Since we are going to be working with text, our computing needs are going to be low. Put together, we have about $150 AWS credit left. This along with Google Colab should be sufficient for our training runs. We also expect that we will be running the model for a considerable number of times to get the best hyperparameters for our model.

Team Roles: The following is the rough breakdown of roles and responsibilities we plan for our team:

- Anjana Niranjan: Data collection, data preprocesssing, working on the model with text input
- Po Yu Yang: Data preprocessing, working on the model with audio input
- Prajakta Karandikar: Training on AWS instance, working on combining the two models for multimodal information

All team members will work on the final presentation, slides, and report.