EMOTION IN MOTION: A Deep Learning Model for Speech Emotion Recognition

by MPG: Matan Gans, Peter Theodores, George Rusu

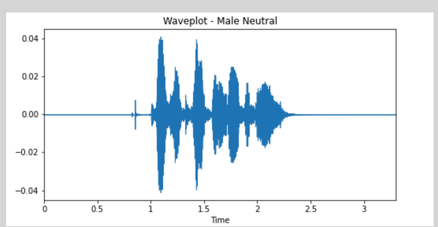
CSCI 1470 (Deep Learning), Department of Computer Science, Brown University

Introduction

The problem we are addressing is the difficulty of classifying human emotion, given variable length speech. Speech emotion recognition systems struggle with classifying emotion due to the abstract nature of emotion and the fact that human emotion can only be detected in small parts during long segments of speech. This project is based on the methodology described in the paper *Attention Based Fully Convolutional Network for Speech Emotion Recognition* by Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, Yanhui Tu

Methodology: Data

We used audio-only speech recordings from the open-sourced RAVDESS dataset to train our model, and extracted four different emotion classes. From each of these wav files, we developed a speech spectrogram representation with a series of 259 time steps representing average frequency values of the audio. We ended up with 510 such representations. Given the limited amount of data, we set up a five-fold cross validation sampling procedure so that we could average results over five unique test samples.



Results

Each trial is run on a dataset over 200 epochs. The unweighted accuracy is the percentage that the model predicts correctly from the test set. The weighted accuracy is the average of the percentage of samples of each emotion that the model predicted correctly.
We managed to get fairly high average accuracies, but the model occasionally falls into a trap of learning to predict only one emotion. This suboptimal strategy would likely be resolved with different hyperparameters, but every iteration of our model that we tried had this issue to some degree. Using the weighted average helped us examine when the model was learning bad strategies, and even though we were not able to completely resolve this, we have a better understanding of why the model is doing this and why its performance varies. This problem is not visible when we only look at unweighted averages. This fact reinforces the idea that we need to examine results more carefully than just looking at an unweighted accuracy, as the discrepancy also hints at potential issues in the data distribution.

Sample Number	Unweighted Accuracy	Weighted Accuracy
1	59.8%	50.2%
2	72.5%	61.5%
3	36.3%	25.0%
4	59.8%	47.8%
5	27.5%	25.0%
Average	**51.2%**	**41.9%**

Figure 4: Unweighted and Weighted Accuracies Over 5 Samples (Following a 5-Fold Cross Validation Mechanism)

Methodology: Model Architecture

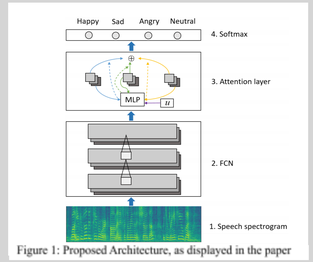


Figure 1: Proposed Architecture, as displayed in the paper

Our model architecture is split into two parts: a Fully Convolutional Network (FCN), and an Attention Layer. The FCN is composed of 1 dimensional convolution layers, maxpool layers, and nonlinear activation layers. Our attention layer is used to construct an utterance emotion vector by extracting the elements most relevant to the utterance emotion and aggregating them. This is done because not all parts of the utterance, meaning not all average frequency values within the spectrogram, are equally significant.

Discussion and Next Steps

Our project turned out reasonably well, as our trained network identifies sample emotions around 51% of the time, which is significantly better than guessing randomly. We were not able to reach the accuracy achieved in the paper, which makes sense as we had to deviate from their implementation given the time constraint on preprocessing. Our next steps involve revising our preprocessing approach to better align our model inputs with the proposed architecture. We gained an appreciation for the challenges of training models on small datasets and learned not to underestimate the importance of having a good amount of knowledge about our data and its distribution before running such tasks.