Individual Lab Report #04

Team D: Deeply Emotional

Ritwik Das

Luka Eerens
Luxing Jiang
Keerthana PG

November 10,2017

Individual Progress

Our FVE goals are to have 50% accuracy on the bimodal network and to track the user in 1 dimension. Overall the individual progress achieved include a revised implementation of the attention model and complete implementation with training of the bimodal network.

The first thing I did was to implement m-DAN and r-DAN implementation of the attention model. Keerthana had already written code for one of the above but it was not correct so I had to modify entirely what she had written and correct it. In addition to that, after talking to our sponsor we realized that r-DAN implementation is more suitable to the problem statement that we are handling over m-DAN which Keerthana had implemented before.r-DAN has been seen to successfully perform on tasks like Visual Question Answering while m-DAN has been used for Image-Text Matching. Since in our case we didn't want to match text with vocal features but rather find out similarities and focus points so the dual attention model for connecting our verbal and vocal cues is the r-DAN implementation.

For the other part I was working with bimodal network. There were a number of things that I did to get to better levels of performance which were not done before:

1. Batchnorm on VocalNet - The paper didn't have any batch-normalization by default implemented within it. We added batch-norm because we felt it would help with overfitting as well as faster convergence.

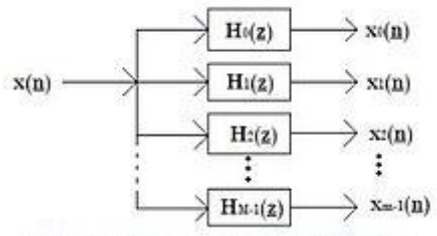

Fig 1. Multidimensional Analysis Filter Banks[1]

2. FilterBanks - Filterbanks(see Fig.1) consists of many filters which decompose a signal. It was proposed that we apply the filterbanks before the convolutional vocal net to better learn features from audio. One more thing which has been implemented is 2D convolutions rather than 1D convolutions. In the current network we are only doing 1D convolutions which are over time. Our sponsor has proposed to try out 2D convolutions, one over time and another over frequency. This has been implemented, but currently no tests

have been run on this network to check its performance.

Lastly I implemented a simple K-means algorithm to cluster all the 5D data in the entire training set for each frame into 6 emotions. I didn't classify on a human level as to which emotion means what but it would help if we want to report accuracy on a classification level. This means that accuracy = no. of correct predictions/no of total examples. On a regression level we are still metrics like mean absolute percent error. This step was just taken to report accuracy values and emotions which can be understood on a human level instead of saying how far we are from a particular 5D data point value which has no intuition in itself.

Challenges

The challenges faced were plenty. Mostly on the time side because we are a four person and additionally due to problems with Luxing. We had to completely finish the PCB design within 24 hours and I knew nothing about it because I hadn't worked on it from the start. So that was really difficult and fortunately with help from others we hopefully rectified all that was wrong before with our PCB and didn't need an extension for the same. Other than that on the technical side, I wouldn't really classify anything what I did as a challenge but rather routine steps which I had to take to make the model work.

Teamwork

Luka again worked on getting the speech pre-processing correct. There were some additional things that we realized since we were not working with the entire dataset until very recently which was not a very intelligent decision to take. Since its a dyadic dataset i.e two people are speaking, we now have audio waveforms and labels for both people which was done only for one person before. Also the intensity label is not present for all the annotated videos in the dataset. So we decided to get rid of the intensity label since valence and arousal are the two dimensions with the most important information anyways and because we can't discard annotated data. He also ordered parts required for the robot.

Keerthana wrote the bidirectional LSTM for text encoding, designed the PCB board and programmed the camera tracker. All that needs to be done now is to integrate this code with the hardware to turn the camera by the requisite amount.

Luxing has done no work since the last progress review.

Plans

My future work on the project till the next progress review is to improve the accuracy of the bimodal network through changing hyperparameters, and testing pre-feature extraction on vocal features like filterbanks , MFCC and also test the new bimodal network for 2D convolutional along those features for the vocal features.

References

[1]https://en.wikipedia.org/wiki/Filter_bank/