Luka Tom Eerens

Team D: Deeply Emotional

Teammates: Luxing Jiang, Keerthana P G, Ritwik Das

ILR02

October 20, 2017

Individual Progress

I began this project by surveying the landscape for deep learning based multimodal emotion recognition publications and their corresponding datasets. By carefully examining the nature of the datasets, the performance of the algorithms and the direct relevance to our project, I was able to narrow down our search space to about a dozen publications and datasets. Next I went through the formalities and paperwork associated with acquiring these datasets. I emailed researchers around the world who built those datasets and asked for copies of their dataset (eg. CREMA-D dataset). I also organised and held meetings with Dr. Jeff Cohen to look into collaboration and for the possibility of using some of his pertinent datasets. To add, I convinced Dr Louis Phillip Morency to let us sit in on his Advanced Multimodal Machine Learning class as I felt that he explained some of these research papers in a very digestible manner. I was also able to download several datasets that were open-source.

With our literature and datasets available, we began further literature review and Ritwik ended up picking the ideal research paper to emulate in code. This paper was titled "End-to-End Multimodal Emotion Recognition using Deep Neural Networks". We vaguely assigned roles for code implementation of the various subsystems but most of these roles have been very flexible and loosely defined, which meant that there was a lot of overlap into each others work. As it pertains to this implementation, my role thus far was data pre-processing. I wrote several python scripts in order to parse through the natural language scripting in .txt files and to convert the .wav audio files into a format that could be input into the neural networks. As we are using a different dataset from the one used in the research paper, I wrote a data pre-processing script that converts the audio input format to be exactly like in the research paper. In the research paper a 6 second long audio sample is fed into the network at a time, and this from a microphone that has a 13kHz sample rate, which means a vector of size 96000 elements is fed into the network. Our dataset had a sampling rate of 48kHz (3 times that value) and so I chose to append every 3rd value of the waveform in the input vector so as to retain a size of 96000 elements. Appending every 3rd value instead of appending the average of every 3 values reduced the computing time by an order of magnitude, and avoided some overflow errors that for some reason were happening when averaging over so many values.

I also debugged an issue that was associated with waveforms that were 24-bit that simply could not be read by the code unless they were in 16 format. I tried various approaches for hours which included converting them in code using ffmpeg but found that the easiest approach involved converting it through a NCH audio switching software. This is something that is not done at run-time but rather manually done to all datasets before running any code and it works like a charm.

Despite all these things, tasks that I carried out individually, I was heavily involved in discussions that we had as a group in order to demystify certain aspects of LSTMs or Multimodal Machine Learning. As a team we were completely unfamiliar with the field of multimodal machine learning, which means that we all invested some time to read up on literature, attend Dr Morency's lectures and become conversant in it. I also brushed up on deep learning as a whole myself because I am still a neophyte when it comes to it.

I have volunteered to become the project manager in order to ensure that there is good information flow from person to person so as to avoid wasting precious time developing things that are not needed. This is a decision that I made in the past 2 days so I do not have any productive updates on this since then.

Challenges

Starting off on the topic of dataset acquisition, it only became truly palpable just how much more important the datasets were than the networks. Before this project, it was a big habit of mine to superficially glance through publications and skim through the abstract of papers that related sort off to my query. This project however thought me to read in between the lines and to rigorously scrutinize the datasets, apparatus used, architectures used to un-tick all of the boxes that should by default warrant us not to use them. An example of this includes me spotting the perfect research paper plus related dataset and only after further inquiry, learning that this dataset is not available to the public at all for the foreseeable future.

I also learnt that most datasets that appear in recent papers have not been organised for data pre-processing and most that we hear about today are still going through that tidying up phase and are not yet available to the public. This is especially frustrating because well written papers with really interesting architectures and state of the art results cannot be implemented if we do not have a clear view of how the dataset is organised and structure. Sure it may say it is emotion recognition, but this may be the emotion of the listener who never speaks, or it could be an interactive dataset with 2 people and no amount of description can replace actually scrolling through it and checking videos, audio files and text. Most multimodal datasets also have bimodal (audio and video, audio and speech, video and text) data, which doesn't make them that helpful for this trimodal problem of emotion recognition.

On the topic of Audio, another big challenge is that sample rate is different for almost all datasets and so requires a different append spacing. Some datasets have 13kHz like the one from the paper we are emulating in code, but most of our datasets do not have this sample rate. I outlined how I approached this problem in the individual progress section above.

Still on the topic of Audio is that the dataset we chose to start of with is a Diadic dataset that features 2 speakers. This is a challenge because the waveform of the conversation is by default not partitioned into the audio waveform of each speaker. It is represented jointly. Thankfully in the case of the SEMAINE dataset, there is not too much overlap in conversation and there are some audio files where the microphones are placed closer to one speaker than the other. I tackled this dyadic dataset by parsing through the script and finding the time interval listed where person A speaks and where person B replies and so on… The waveform was then nulled at each time interval where person B replied or vice versa and so the result is an audio waveform that has an acoustic wave moving sideways with 0 amplitude when another person speaks. This is probably not the correct way to approach this problem for a neural network, and a possible better substitute would be to then replace all regions of the waveform with values of 0 with background noise that lies within the range of the background noise in between the word pauses when person A speaks a sentence. This will probably be explore next.

Teamwork

Ritwik:

Drove most of the technical high-level progress by outlining some tasks we need to do as a group and suggesting which approaches we should follow. He wrote code for our speech net analysis neural network, and modified the initial LSTM provided by Keerthana He also quickly implemented dlib face recognition functions in order to detect faces in an image. He also along with me, sent emails to professors in order to acquire these datasets. He also helped set up the workhorse computer in the MRSD lab by installing most of the dependencies.

Luxing:

Helped Ritwik in writing up the speech analysis network and helped us in finding datasets. He conducted a lot of literature review in order to help narrow down the right papers to follow and also looked for datasets. Luxing attended pretty much every single meeting and was almost never absent from any scrum. He also helped modify and update our website. He also helped in setting up the workhorse computer for use by testing code that made use of the dependencies we were going to use.

Keerthana:

Liased quite a bit with our sponsor Emotech as well as academics such as Dr Jeff Cohen. She helped survey the landscape of research papers and found many useful datasets. She has also been responsible for implementing much of the LSTM algorithms and has made the initial contribution to the speech network. She also helped explain several deep-learning concepts to me and helped me understand how LSTMs work.

Plans

I hope to make the audio waveform pre-processing work for more general datasets (if we are going to use other datasets than SEMAINE down the line). I also hope to implement the substitute of nulled waveform with background noise that is consistent with the training data.

After this I hope to assist in the integration of the subsystems and debug any reasons for unexpectedly poor performance or breaking. Given the challenging nature of software project integration I believe that this may take quite a bit of time to do and so am quite content with saying that this is all that I hope to be able to contribute to by next review from a technical standpoint.

I believe that as a team, our efforts will be most spent around this, and this will be something that will require all of us to contribute.