Individual Lab Report #03

Team D: Deeply Emotional

Ritwik Das

Luka Eerens

Luxing Jiang

Keerthana PG

October 27,2017

Individual Progress

Starting with a brief overview of the project's current state; after meeting with our sponsor and going back and forth on a few things we have decided on the architecture of the tri-modal network. The cyber-physical architecture of the network is shown in Fig 1.

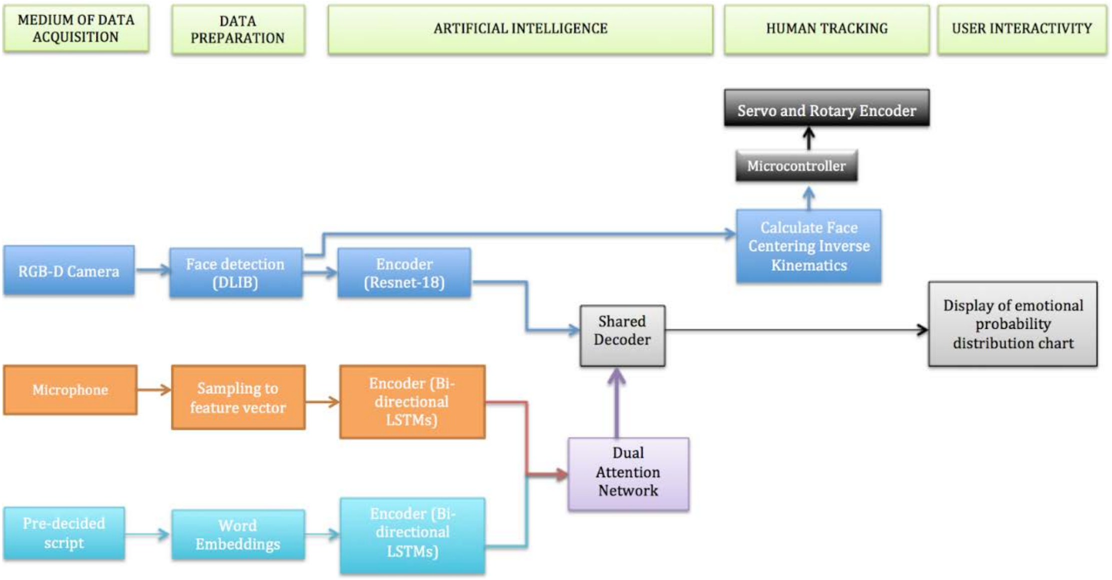

Fig 1. Revamped Cyber Physical Architecture

The pre-processing steps include cropping facial images from image frames, preprocessing text into word embeddings, and converting raw audio waveforms into vectors. After that we use bi-directional LSTM as encoders for both audio and speech. We use resnet-18 as the encoder for vision. After all the feature extraction has been completed by text plus speech encoders we feed it into a dual attention network which learns similarity and focus points in both text and speech. The output of this combined with the output of the resnet-18 feature extraction is fed into a shared LSTM decoder which will predict the emotion.

I completed writing code for training and testing the bimodal network [1] and combining all the different networks into 1 single network. This involved putting all the classes together through a single train function and a single test function, and writing data-loader and normalization classes.

I made sure that the backpropagation on the entire network was working correctly once combined. Defining the back-propagation and corresponding gradient derivations properly was important since there is a concatenation of vectors before the LSTM feed and we didn't need to backpropagation on the resnet-18 for obvious reasons. The parameters for gradient descent were therefore only chosen to be the speech-net parameters and the LSTM parameters. I tested the backpropagation using a few random vectors and tried to over-fit the network for those samples to see if the loss goes down. The test was successful and so we at-least made a sanity check that our network has been implemented correctly and debugged all major parts where issues could have occurred with high probability.

I defined the target folders (i.e. containing the labels) for training and testing, loading the data and normalization of data before feeding it in the network. Image normalization was done for the resnet-18 with a particular mean and standard deviation which has been pre-computed by the respective authors of the paper.

It is also important to note that the knowledge of putting together different types of individual networks, getting the backpropagation working correctly across the entire huge network. writing functions to do routine tasks for the bimodal network will be reused for the tri-modal network saving us time.

I also wrote code for storing all of the post-processed audio waveforms and video image frames into HDF5 file datasets which could be quickly accessed without needing to refer to the dataset again and again.

Other than that, my time was invested in understanding the dual attention models paper [2] which was published in CVPR 2017 and fitting it together to complete the new cyber-physical and functional architecture of the tri-modal network.

I helped Keerthana with the attention models to solve some of the bugs that she encountered. The dual attention model paper uses a two-layer feedforward neural network to calculate a scalar similarity value between two input vectors. A custom loss function was defined for the same purpose. I worked on getting the forward and backpropagation running with the other parts of the code that she wrote.

Challenges

I would say that I didn’t face any major challenges this time around. Apart from a few bugs while combining the network and a bit of math on the CVPR papers it was mostly very fluid. It was only a matter of time to get through the bugs and the paper to figure out how everything works. The luxury of a challenge free week was mostly because the things I did this week were although vital but mostly trivial to accomplish like reading papers, implementing routine function calls etc.

Teamwork

Luka further worked on the speech preprocessing part. He had a few minor bugs which he corrected and wrote scripts to parse through the entire dataset, divide each audio waveform into pre-decided sequence lengths(for the LSTM) and finally output a 96000 vector for each of them. He also worked on doing trade studies for the parts that we need to order. He also made a CAD version1.0 for the fall validation experiment.

Keerthana took the responsibility of working on word embeddings as well as the attention models. This will be later merged together with other parts of the tri-modal network which we will be building till our next progress review.

Luxing worked on the network combination along with me. We got it working and we will start training on this bimodal network over the weekend.

Overall, this week was better coordinated and planned than the last week. We met our goals and we think we are on track for the next progress review as well. We postponed one of the things which was training the bimodal network in exchange for understanding the attention models, changing our architecture and working on CAD models based on talks with our sponsor and feedback by John on hardware requirement during the FVE. However, we did finish putting it all together, doing sanity checks and unless we have a lot of bugs in the current network (which is unlikely) which show up later, we should mostly be good.

Plans

My future work on the project till the next progress review includes the following:

1. Debug network issues for the bimodal network during training or convergence.

2. Putting together all individual parts of the tri-modal network.

References

[1]https:/arxiv.org/abs/1704.08619v1

[2]http://openaccess.thecvf.com/content_cvpr_2017/papers/Nam_Dual_Attention_Networks_CVPR_2017_paper.pdf