Individual Lab Report #02

Team D: Deeply Emotional

Ritwik Das

Luka Eerens
Luxing Jiang
Keerthana PG

October 20,2017

Individual Progress

Starting off with a brief introduction, all of us decided to implement a bi-modal emotion detection paper due to less research on tri-modal machine learning as compared to the former. This paper takes in audio waveform and video only and uses two convolutional nets(see fig 1 and fig 2) for feature extraction from audio and video and an LSTM(see fig 3) to process the temporal relationship of the concatenated feature vector from above.

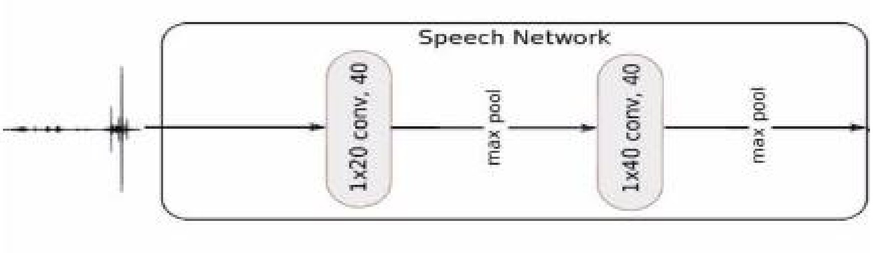

Figure 1: Speech Convolutional Net to extract 1280 features[1]

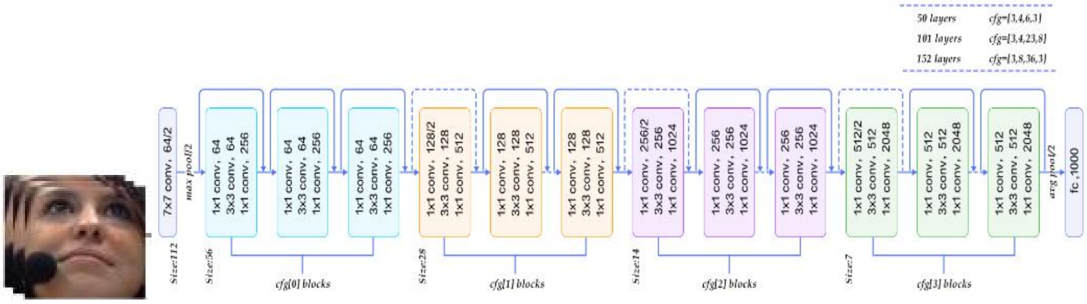

Figure 2 : Resnet-50/101/152 [2]

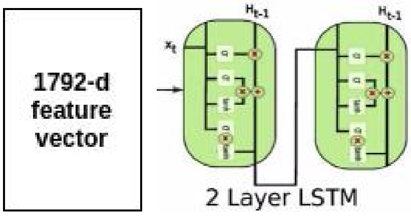

Figure 3 : 2 Layer LSTM takes in 150 sequences of 1792-d feature vector[1]

I wrote code for pre-processing the video data into a stream of normalized, re-sized images to feed it into the resnet-18 network. I initially had decided to use the resnet-50 but then resnet-50 had 2048 features before the fully connected layer(see fig.2) while resnet-18 had 512 features. Although the numbers can be modified, but that would require additional layers to get it down to less features. Therefore, to save time with further training I went ahead with resnet-18 with pre-trained layer weights.

I also wrote code to get the resnet-18 working on a dataset of facial images labeled with emotion to fine-tune(not train from randomly initialized weights) the pre-trained layer weights to facial images only since resnet-18 was trained on ImageNet. I did this by removing the last fully connected layer and then adding a new layer which goes from 512 features to 6 classes(due to 6 emotions). The network is yet to be trained, but the requisite steps have been completed. The main objective of this network is not to use the network for uni-modal emotional recognition but rather to fine-tune the weight matrices so that the visual network is disabled for back-propagation when networks from all modalities are put together along with the LSTM. This decision was taken since traditionally it has been observed that in the case of LSTMs/RNNs + CNNs using a pre-trained CNN as a feature extractor and feeding it to the LSTM/RNN achieves much better convergence and performance as compared to training the huge entire network together.

Further, I modified the speech convolutional net and the LSTM written by my teammates.

For the speech convolutional net, there were quite a few things in the paper which were not mentioned regarding the network like the stride size and padding on each of the convolutional filters, so I had to play with these hyper-parameters to get to 1280 extracted features. Further tuning of these hyper-parameters will be done in the future during training and validation cycles. For the LSTM, I corrected some minor bugs in the batch size and sequence length.

I also wrote code for a sequence to sequence attention model. Attention models are fairly new to the domain. This was written separately which will be integrated later when we work on the encoder decoder network for all three modalities. Currently I wrote it only for the text modality. I will replicate it for other modalities soon.

Challenges

The main challenge I faced since the last progress review was time-management along with the other MRSD assignments,midterms and job interviews. I am also sitting in the Multimodal Machine Learning class on Tuesdays and keeping up with the paper reading assignments for the course, so it was a huge time crunch for me. The only way I could address this issue was to work late nights and sleep less. However, in the future I hope to be more productive to have a proper balance.The other hurdle was writing deep learning architectures using Pytorch. I am used to the static graph implementation of Tensorflow w/o Keras and had never used Pytorch before. I read the documentation carefully and it was pretty much pattern matching after that. None of these challenges were large enough to dent our timeline. Another smaller hurdle was that the paper we were trying to implement was incomplete in some of its parameters in the convolutional network for the speech modality. So I had to use my own hyper-parameters to get the desired output at the end.

Teamwork

Luka completely worked on getting the speech pre-processing correct for the SEMAINE dataset. This was a very critical and a tedious task. He got everything working perfectly for our progress review. The main aim was to convert the raw audio waveform into a 96000 dimensional feature vector which could be fed into the speech convolutional network.

Keerthana took the responsibility of working on the LSTM part of the paper. The LSTM part works on the concatenated feature vector from the speech and the visual networks.

Luxing worked on the speech convolutional network with me.

Despite ups and downs due to disagreements and miscommunications, I think we got it very well together for the progress review and are on track as per our deadlines as a team. There is always room for improvement though and I hope that we can come to mutually agreeable decisions much faster in the future rather than going on in circles. We have also appointed Luka as the project manager within the team to prevent such miscommunications and to mediate work which is not discussed at length during our scrum meetings. As for everyone keeping up to date with new code pushed and tasks due we have integrated Github and Trello with Slack so that everyone in the team is instantly notified of new commits,pull requests made to the repository as well as new tasks and whom are they assigned to.

Plans

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

1. Integrating the LSTM and both the CNNs into a single network and start training the network.
2. Debug any network issues that occur during training and/or convergence.

References

[1]http://book.paddlepaddle.org/03.image_classification/
[2]https://arxiv.org/abs/1704.08619v1