Individual Lab Report #05

Team D: Deeply Emotional

Ritwik Das

Luka Eerens
Keerthana PG

November 22,2017

Individual Progress

My individual progress was to converge the loss function on the entire training set. All of my time was vested into getting the end to end network working. While doing this I had to re-initialize the weights on the resnet, the reasons for which I have detailed in the challenges section. Due to this, to introduce diversity within the dataset for the vision part, I have also trained the resnet separately on CREMA-D dataset which contains emotional labels visual segments only for 91 diverse actors across 7500 video clips. I wrote code for the entire pipeline from preprocessing to training and testing. These weights are now being finetuned on the SEMAINE dataset which has significantly less diversity. Fig 1 shows the output of the network on a frame of video of a person it has not seen or heard before.

Fig 1 : Top Left Hand Corner Text showing Predicted and Actual Output of the Network

I have also started doing ablation studies to answer why our model works the way it works. This includes writing code for batch variance checks across different critical parts of the network i.e between the concatenation of the encoder features and the decoder and changing hyper parameters on the

LSTM to check temporal behavior. Most of my “progress” was spent on getting things to work so from the perspective of deliverables I am behind schedule on my tasks

Challenges

There were a major challenge that I faced in training the end to end network on the entire training set. The loss function wasn’t converging. Within 10% data being seen on the first epoch the loss saturates to a particular value and irrespective of future mini-batches. And the saturated loss wasn’t low enough to give us any suitable predictions. I changed the LSTM architecture and there was no difference whatsoever. So it was clear that the vocalnet or the resnet is outputting features which have very little difference within them. This was true when I did a Variance check across a batch of sequences coming out of the vocalnet and the resnet concatenated vector. So the next step which I did was that I unfroze the pre-trained Resnet and allowed the gradients to flow through it. Before this point I was just doing feature extraction on the Resnet and concatenating that vector to the vocalnet output vector. Even then the loss function didn’t decrease and it showed the same behavior as before. I changed every hyperparameter, the optimizer, switched from minibatch gradient descent to stochastic gradient descent and back to minibatch but it wouldn’t budge. After 5 days of trying out different things I reinitialized all the weights on the resnet and started training using the architecture only. Finally it worked 1 day before the progress review. My conclusion is that it was stuck in a really deep minima and even very high learning rates could not take it out of it. This was really unexpected since transfer learning has been touted in the machine learning community. Intuitionally it makes sense why even pre-trained weights with gradient flowing would not solve the problem because a lot of neurons will be getting activated on arbitrary images like cats,dogs,birds while we are supplying only faces. The resulting outcome is that we have to now train the resnet on more faces to introduce diversity which is why now we are training it on CREMA-D dataset which is a newly released emotional database on 6 emotions containing 91 actors. The diversity factor was a major risk for our project and it has manifested itself. I have also come to the realization that emotion recognition is a hard task because of the huge subjective factor involved in it and perpetuated through the datasets used. For instance a particular dataset of videos may have vision inputs which are really subtle and a network trained on more obvious visual cues won’t do really well although it might perform really well in the test set of the particular dataset. Text is the best

leveler but text trained on a particular dataset might have more key words which clearly show a emotion rather than another dataset or in real life. I am expecting a significant increase in performance once we include text in our model. We will see how it goes for the FVE. Right now our accuracy on the resnet alone without the vocal-net and the LSTM for the test-set stands at 32%.

Teamwork

Luka 3D printed the parts for the robot and there were a few small surprises with the dataset that I noticed which were corrected by pre-processing differently.
Keerthana worked on the getting the robot hardware and tracking working. She integrated the python and the arduino code.
Working in a three person team is difficult for a project of this scale. Luka has to work on and off on the preprocessing/normalization/data-engineering/ part because of the various mathematical problems that show up while training the network, or deciding to train the network in a different way and do the hardware part. Keerthana has been working on the hardware aspects of the project. It was supposed to be really simple but there were problems that arose as demonstrated during the progress review presentation. So majority of the core deep learning part is only left to me for quite a long time now which is the most critical aspect of the project deliverable wise.

Plans

My work till the FVE is to train the resnet on more faces, use those pretrained weights on the end to end network and improve the accuracy of the bimodal network.