CS6300 - Speech Technology
Final Project - Speech E2E

Dhanush Krishna : EE16B009
Vignesh S : EE16B127

October 21, 2020

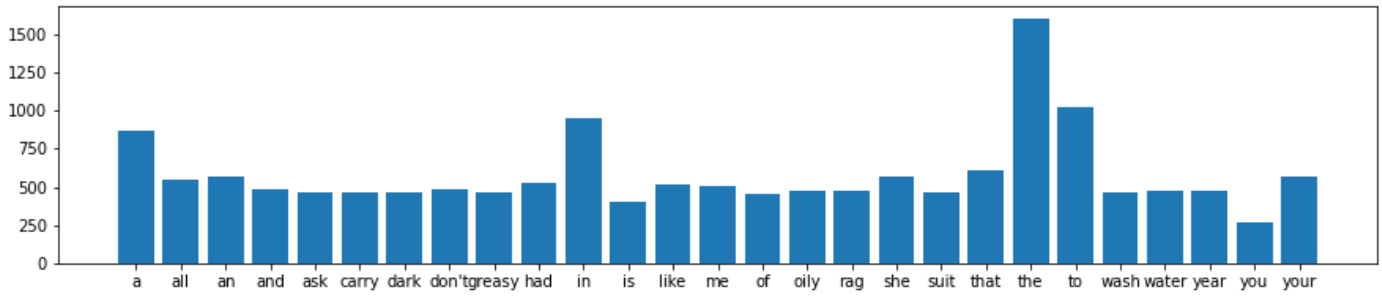


Figure 1: word frequencies

Input pre-processing

Steps

1. We collected all the words from the TIMIT dataset sentences and chose the words with a repetition of at least 250. This is done so as to get sufficient features for each word which we want to classify. We get a set of 27 words which have occurance frequency as shown in 1.

2. We go through each of the .wrd files for all the utterances and collect the frames (which are sampled 16kHz) which correspond to these words.

3. We then compute the MFCC features for each of these frames and divide the entire thing into 80% for train and remaining for test.

4. The training input contains the MFCC features of 38 coefficients stacked. Each of these datapoints have features which are calculated over windows of 0.008 seconds with an overlap of 0.004 seconds.

5. The number windows are made uniform for all the datapoints, and those with lesser than maximum number of windows are padded with zeros.

6. The output form is number between 1-27 which indicate the corresponding class of the word.

Encoder-Decoder

Without attention

The inputs so obtained upon pre-processing has to be first passed into in Encoder-Decoder (E-D) structure so that the representation and information of MFCC features can be compressed into a context/hidden vector of dimension 256 (for each datapoint). The E-D network is made up of GRU units. The output and input of the E-D network is same. Upon training, the test inputs are also made to pass through the E-D and we obtain the hidden vectors for all of them. These hidden vectors are passed through the ANN and SVM for classification. The training for this was done with a learning rate of 0.001 and batch size of 20

With attention

In the case of attention, instead of just taking the context/hidden vector for the final stage, the hidden vectors of previous few stages are also considered. The previous hidden vectors are weighted so that some of them are given more importance than the others. This helps especially in case of translation related tasks where the words in a sentence of different languages are not in the same order. In this particular case, attention does not help significantly since the input and the output are the same. As a result the weightage will not affect significantly.

Classification

The accuracy for different models is as shown in the table. The accuracy for Isolated digits model was almost 100% using a single hidden layer and hence no more experiments were done on the same.

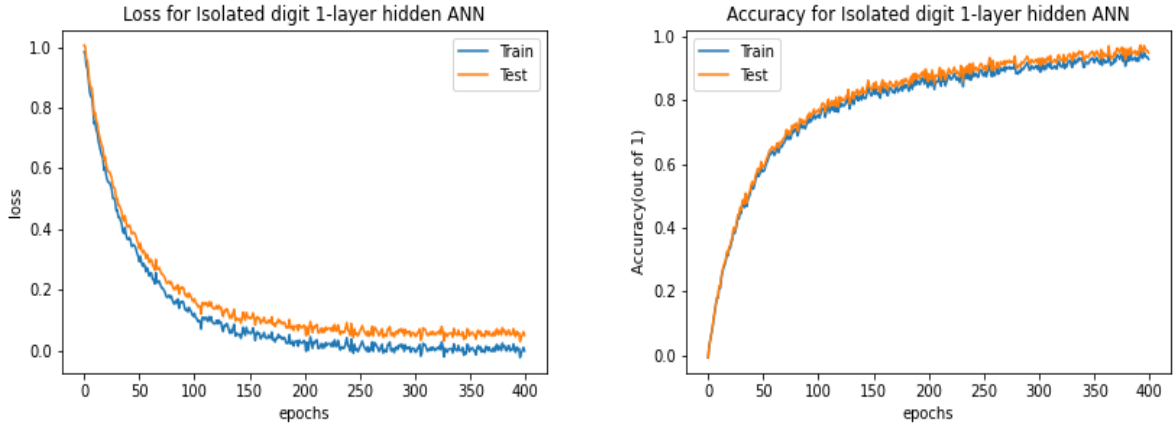
Network type	Accuracy for TIMIT
ANN with 1 hidden layer	59.7%
ANN with 2 hidden layers	61.9%
ANN with 1 hidden layer and attention	66.5%
SVM	57.2%

ANN

Upon obtaining the hidden vectors from the E-D network we proceed to classify the utterances using a Deep Neural Network. We experiment for both single and double hidden layers. We also classify the attention based model also. The plots for the loss v/s epochs and the Accuracy v/s epochs are shown in the plots below. If we look at the tSNE plots, we see that initially (input/hidden layer 1) there is no clear boundaries. However for latter 2 layers we see that there are visible boundaries that emerge. These show that at the output layer it is much easier to classify the words. Again we use a learning rate of 0.001 and batch size of 20.

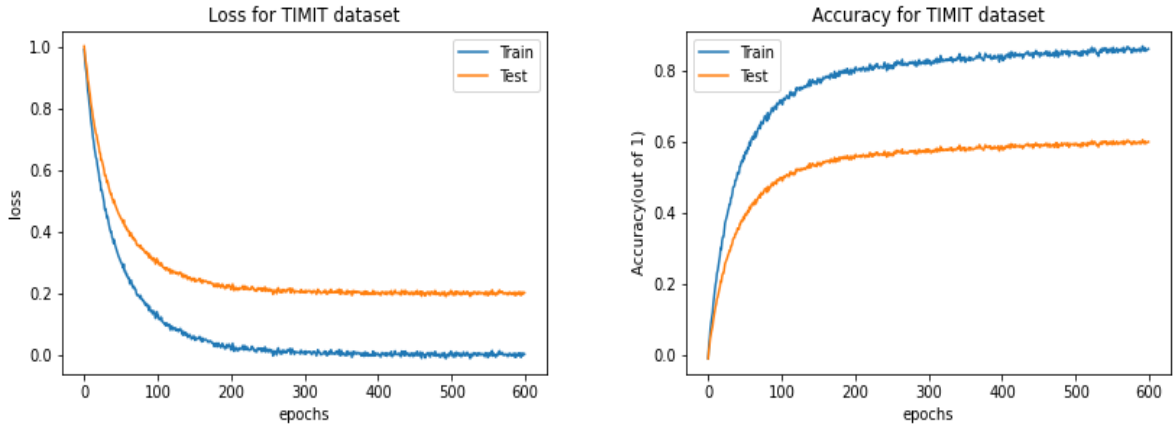
SVM

The feature vectors are passed through an SVM Classifier. The variation of regularization v/s SVM accuracy is shown. Linear kernel is used here for classification, since it is the simplest and gives accurate. Non-linear kernel might overfit the data in case of large number of features, so the linear kernel would give the best performance in this scenario. As you can see in Figure 8, the accuracy increases with regularization parameter. This is to be expected since increasing the regularizatoin parameter would allow more leeway for samples to be further from the margin of the classifier. As the restriction is relaxed, the accuracy increases both in the case of training and validation data. The regularization punishes the misclassfication.



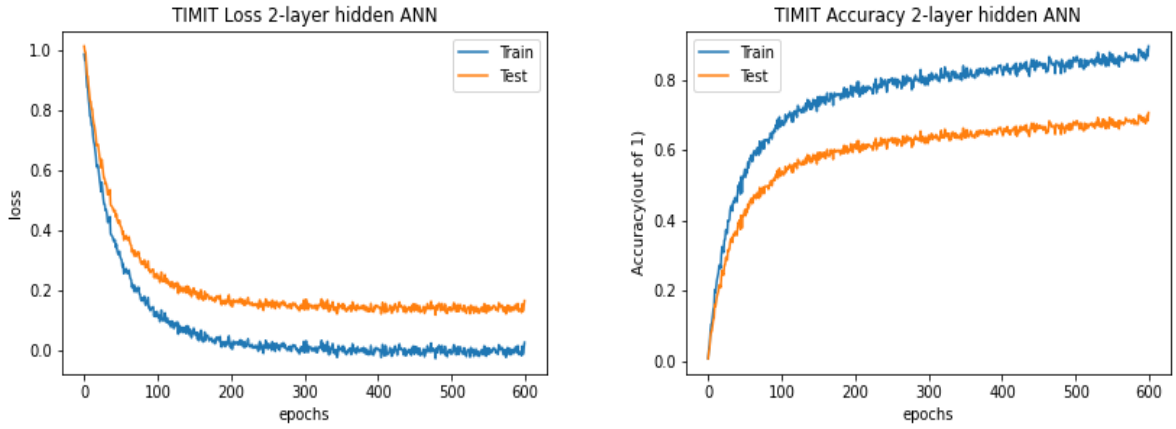
(a) (b)

Figure 2: Single hidden layer ANN - Isolated digit



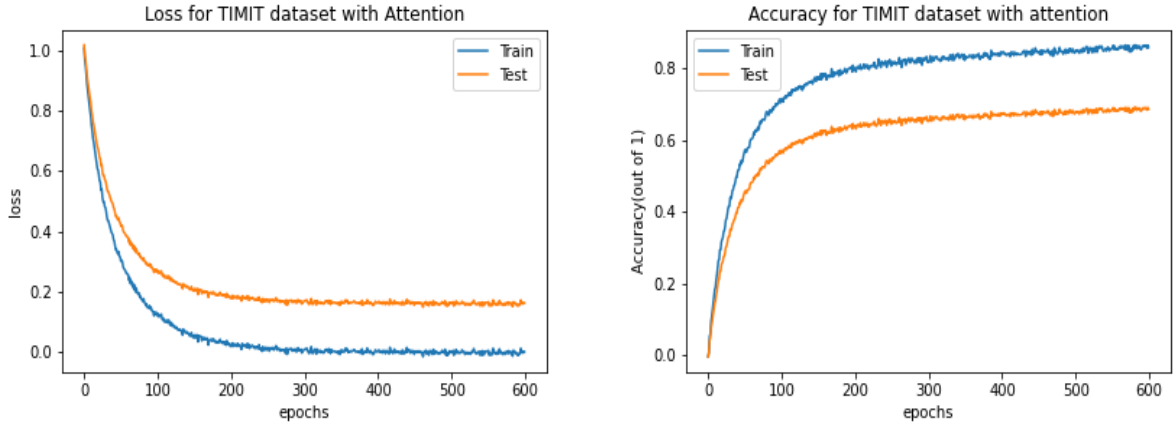
(a) (b)

Figure 3: Single hidden layer ANN - TIMIT



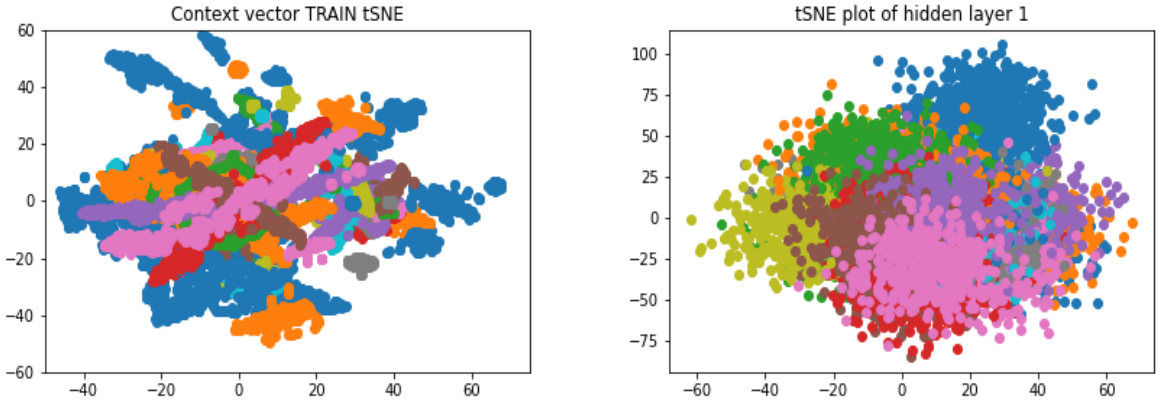
(a) (b)

Figure 4: 2 hidden layers ANN - TIMIT



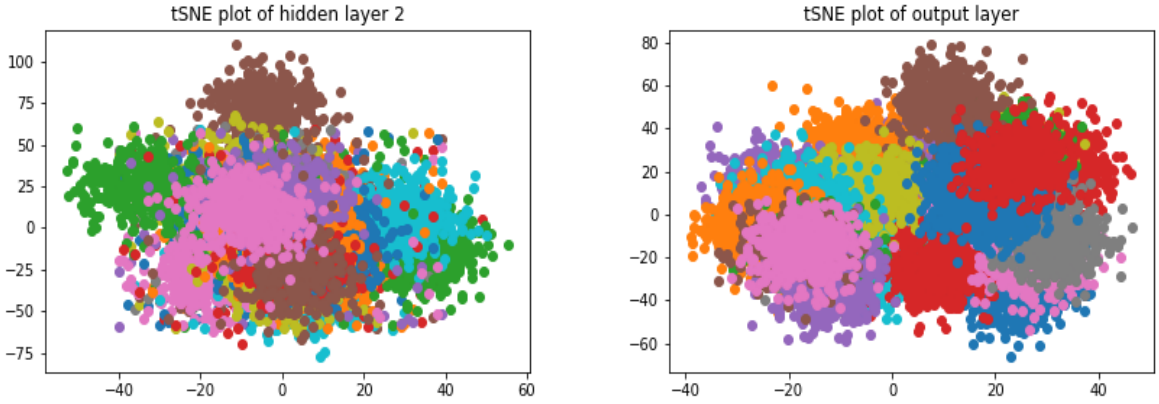
(a) (b)

Figure 5: Single hidden layer ANN - TIMIT with attention



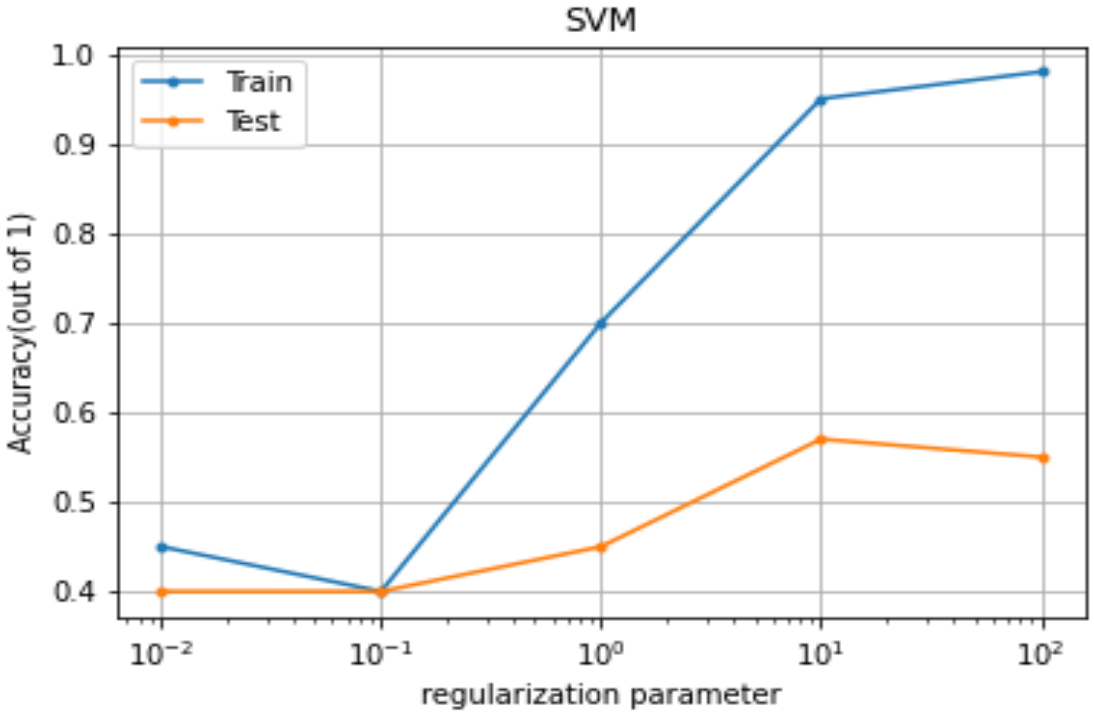
(a) (b)

Figure 6: tSNE plots of context vector and 1st hidden layer



(a) (b)

Figure 7: tSNE plots of 2nd hidden layer and output layer



(a)

Figure 8: SVM v/s Regularization Parameter