Speech Processing
Exercise 4
Oren Cohen - 305164295
Liron Harazi - 066743253

In this exercise we choose to work with STFT (the Gcommand_loader.py code was very helpful).

Data samples were wav files, each file with 1 sec recording, then we do STFT => data with 4 dimensions: [64, 1, 161, 101]
- 64: is number of batches (hyper parameter)
- 1x161x101 dimensions of the 3dim “image” (stft),
notice that 101 is the timestamps (time-slice) and 161 is the “features”.

Architecture:
we build a NN with the next layers:
- 2 2Dconvonotional layers (extract features from stft), and then Max-pooling
- bidirectional LSTM with 4 layers (“learn with context”)
- FC NN (predict letter from 26 alphabet)

some important properties:
- the filter window size of conv layer was: 3x3
- we added batch-norm and dropout (2D NORM, and 0.3 0.5 dropout) to the conv layers
- the output of first conv layer was activate with RELU function (input to 2’nd conv)

basic dimensions throw the NET:
data – 161x101
conv layer_1 (3x3 filter, 10 feature maps) – 10x159x99
conv layer2 (3x3 filter, 1 feature map) – 1x157x97
maxPooling (2x2) - 1x78x48
lstm with hidden of 300 – so we get: 48x600 (48 is time-slices, each time slice 600 features)
fc1 – 48x150
fc2 – 48x27

and in the end instead of softmax, we used log_softmax (for usage of CTC)
and that it! We have a neural Net that do ASR!

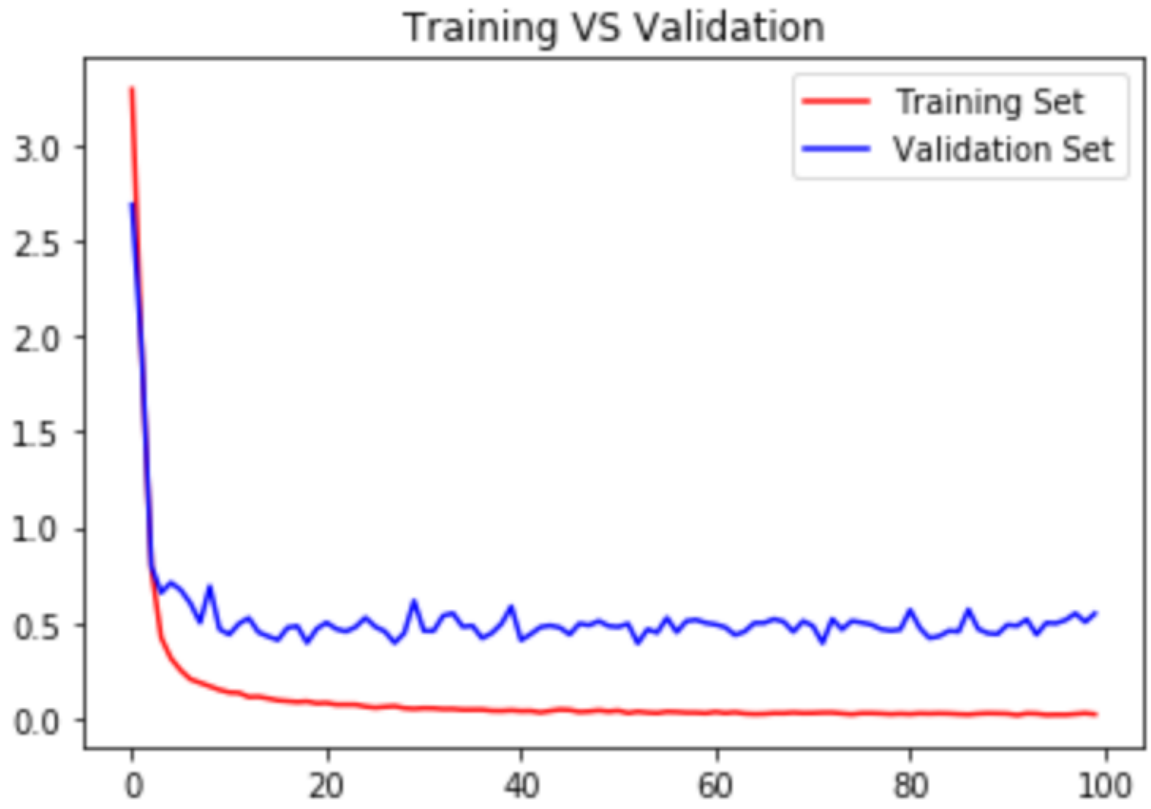
Some of the hyper parameters:
batch_Size = 64
learning rate = 0.001
# epochs = 100
optimizer = Adam

Speech Processing
Exercise 4
Oren Cohen - 305164295
Liron Harazi - 066743253

Results:

here is the graphs of our results:

Avg loss per epoch (train and validation)



accuracy (measures with CER – less is more;)) - we got the amaizing avg CER on validation: **0.11**