Project 2 Handwriting Recognition
CSE 574: Introduction to Machine Learning

Kishan Dhamotharan
Person# 50287619

Introduction And Data Preprocessing

Source of dataset:

- Human Observed
- Gradient Structural Concavity

All the images which where used to generate this dataset are paired up and compared, Also marked if they are matching or not. The number of features for each image in human observed data are 9 and 18 for Gradient Structure Concavity. Each of the image is mapped to a unique ID. The data was divided into three files one which contained the features of each image, second file with all the matching pairs and the last one with the pairs of images which do not match. It was observed that the file which had images which do not match, had lots more data than the matching one. First step was to map the id of images with their feature vector. Now we replace the id of the images with their feature vector to get meaningful data. We came up with two approaches of doing this, one was to concatenate the features vector of the two images, which resulted in 18 and 1024 features respectively for human observed and GSC, and the other approach was to take the absolute difference between each feature. Initially tried to achieve this through iterating the dataset, which turned up to be a very expensive operation. Hence used pandas merge in a very efficient manner, such that the preprocessing time was brought down to seconds. As mentioned earlier we had lots more sample for not matching scenario. Which could make our system to be biased. Hence we picked equal number of matching and not matching samples. Hence after this we will be have 4 data sets:

1. Human Observed Dataset with feature concatenation
2. Human Observed Dataset with feature subtraction
3. GSC Dataset with feature concatenation
4. GSC Dataset with feature subtraction

Now we use each of these four dataset on linear regression, logistic regression and neural network.

Results

Following are the observation and results which were obtained on training the model with different datasets. Overall it is observed that the GSC dataset across all the models performed better than the human observed dataset. And the best results were observed when the dataset was trained on neural networks. Following are the results obtained across training different dataset over different model, and trying out and finding the best hyper parameters for different models.

Other important steps which were carried out:

- Identification of columns which had variance as zero had to be removed as we had to take inverse of a matrix
- As the data set was massive, due to computational constraints used 7000 rows of matching and non matching images for GSC dataset.
- Data set was divided into training, validation and testing in the ratio of 8:1:1
- Also had to properly shuffle the dataset so that the training could happen in a proper manner.

Linear Regression

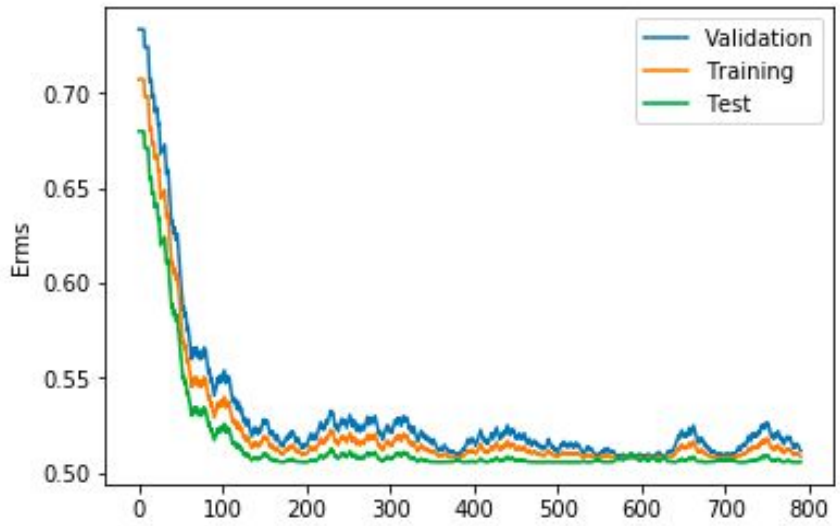
Obtained best results using the following configuration:

No of cluster: 15
Learning rate: 0.001
Regularization term: 2
Epoch: 700

a) Human Observed Dataset with feature concatenation

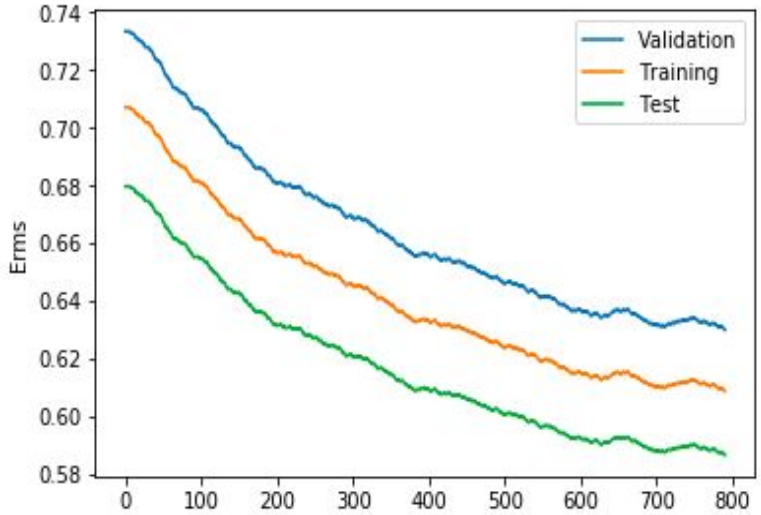
Training phi shape: (1266, 18)
Validation phi shape: (158, 18)
testing phi shape: (158, 18)

Ideal Case



Reduced cluster size

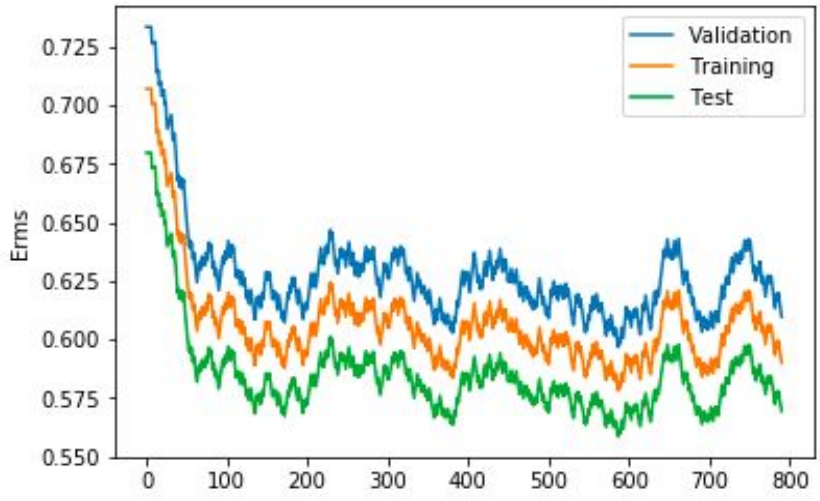
Cluster : 2



Increased learning rate

As expected with higher learning rate we can observe more noise in the erms graph.

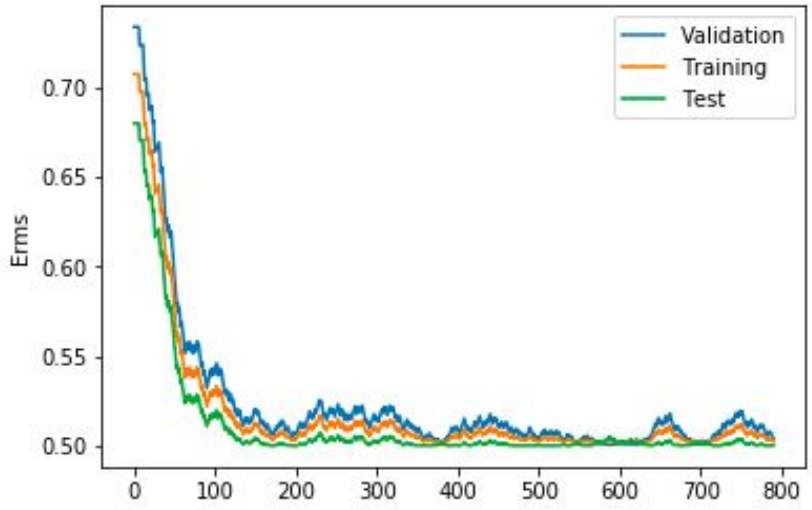
Learning Rate: 0.01



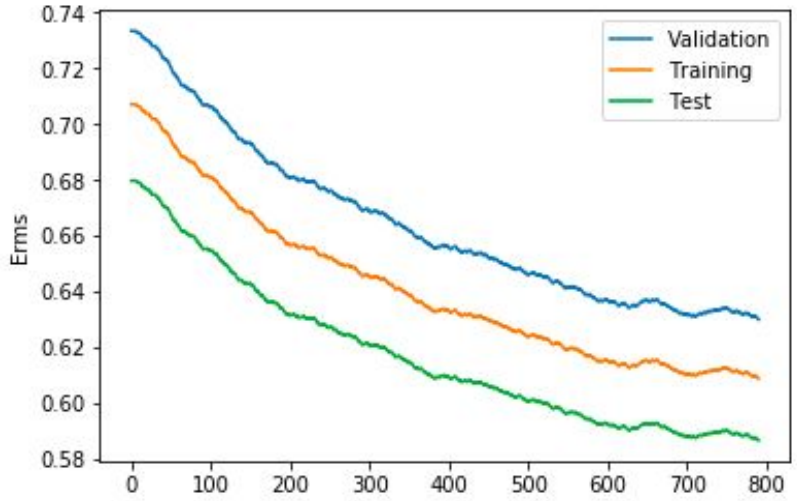
b) Human Observed Dataset with feature subtraction

Training phi shape: (1266, 9)
Validation phi shape: (158, 9)
Testing phi shape: (158, 9)

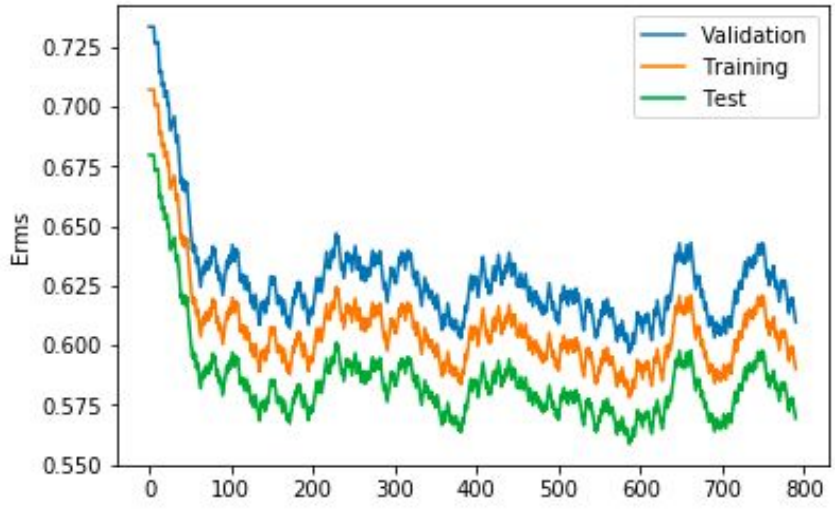
Ideal:



Cluster: 2

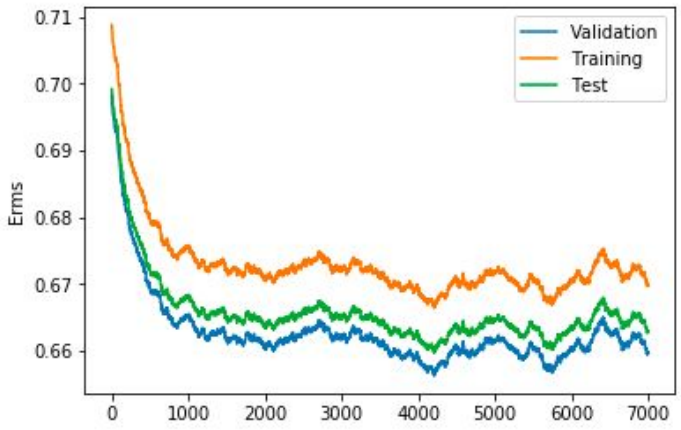


Learning rate: 0.01

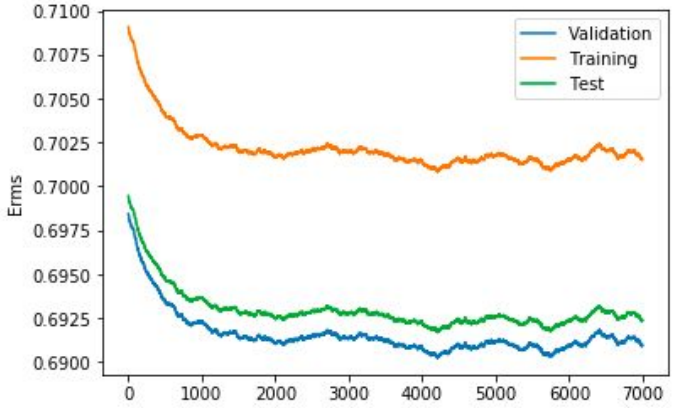


c) GSC Dataset with feature concatenation

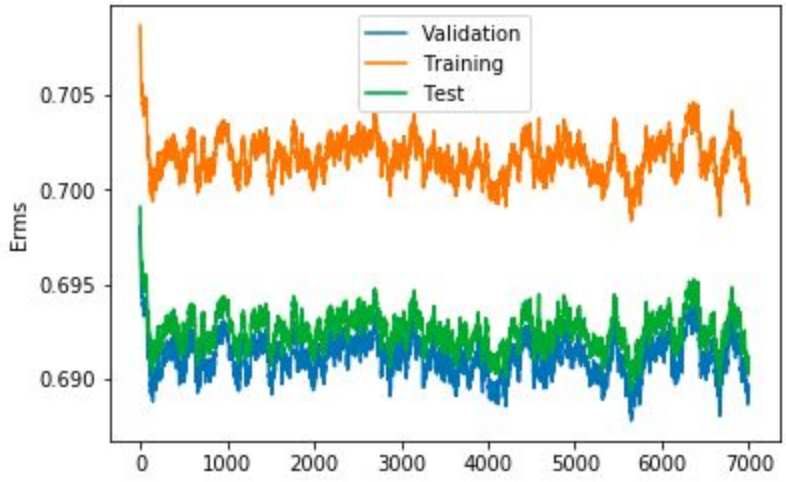
Training phi shape: (11200, 498)
Validation phi shape: (1400, 498)
Testing phi shape: (1400, 498)



Number of cluster 2



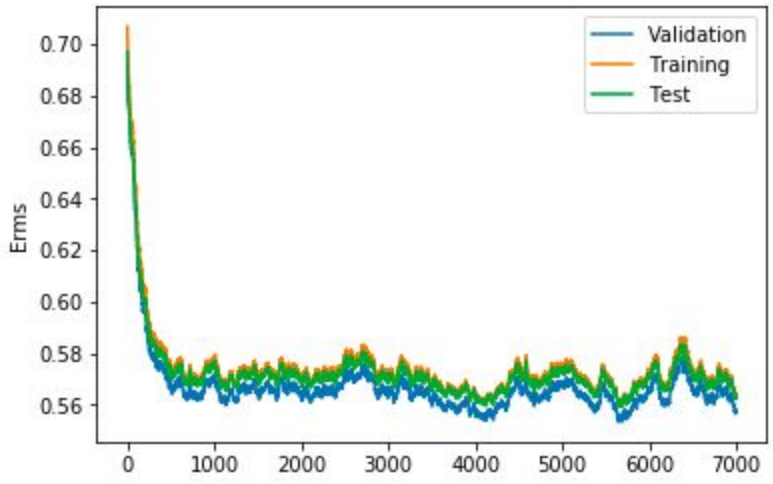
Learning rate 0.01



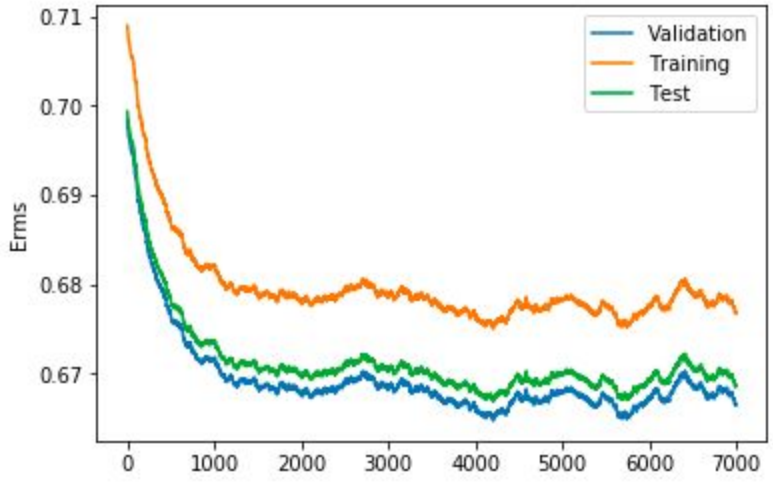
d) GSC Dataset with feature subtraction

Training phi shape: (11200, 498)
Validation phi shape: (1400, 498)
Testing phi shape: (1400, 498)

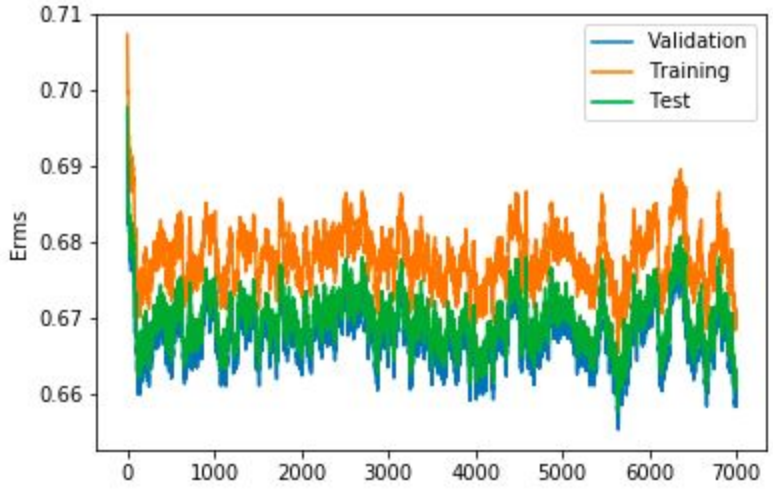
Ideal



Number of cluster 2



Learning rate: 0.01

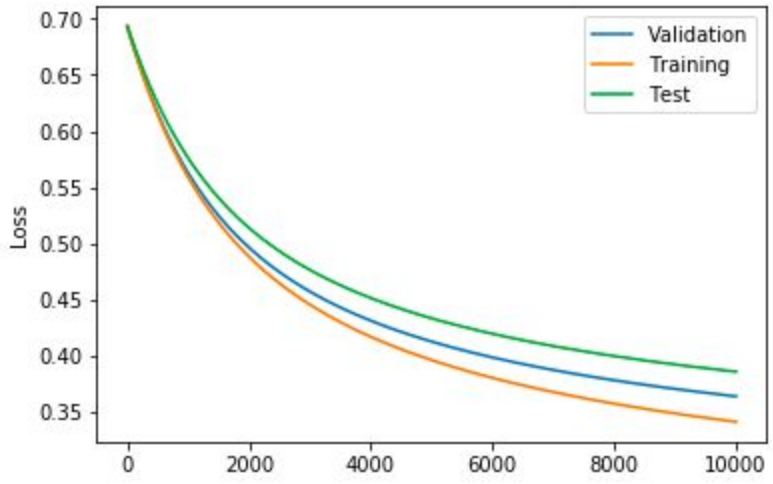


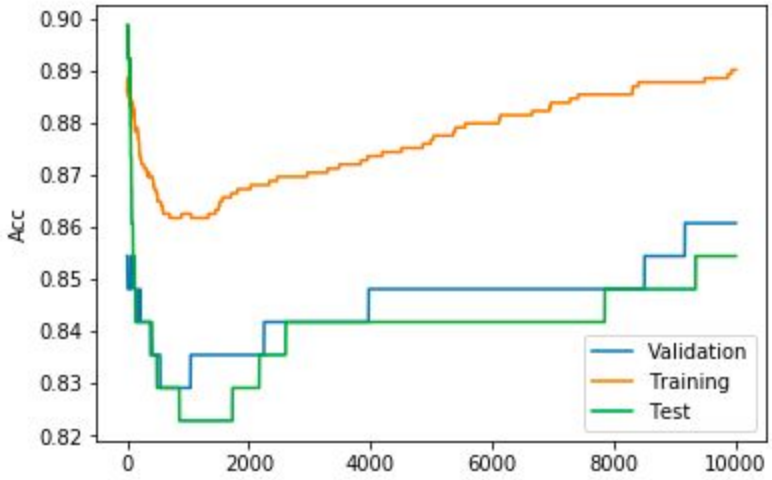
Logistic Regression

Obtained best results using the following configuration:

Learning rate: 0.001
Epoch: 10000

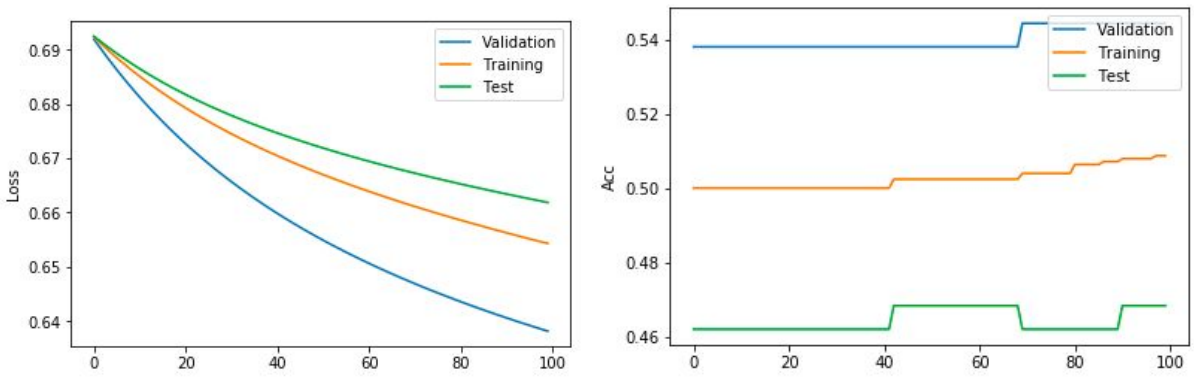
a) Human Observed Dataset with feature concatenation



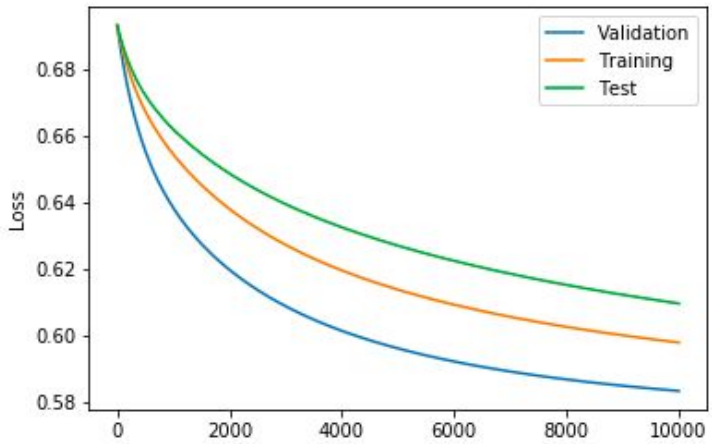


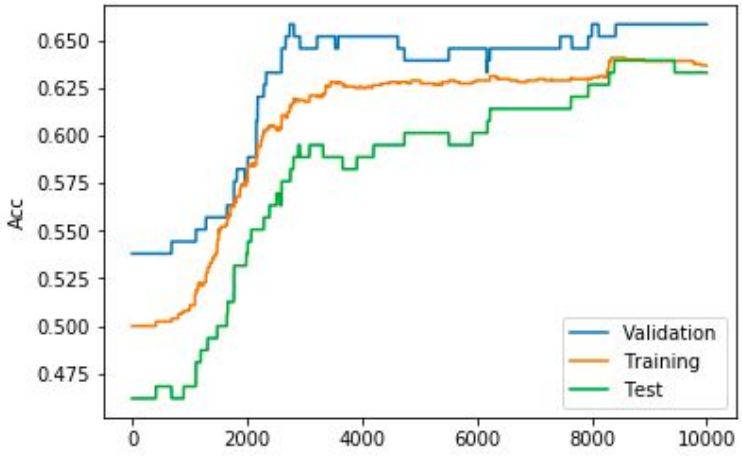
Lower Epoch and higher learning rate:

Lower epoch and high learning rate significantly fact the accuracy of was model , as one can observe that the accuracy is also constant over the period.

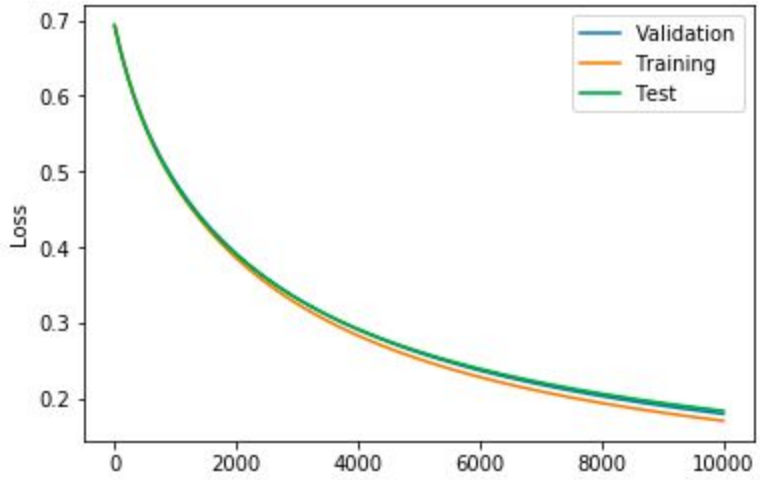


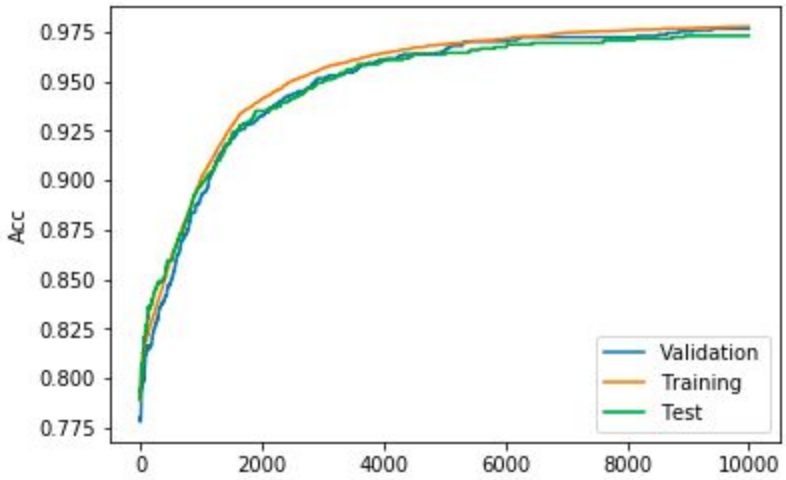
b) Human Observed Dataset with feature subtraction



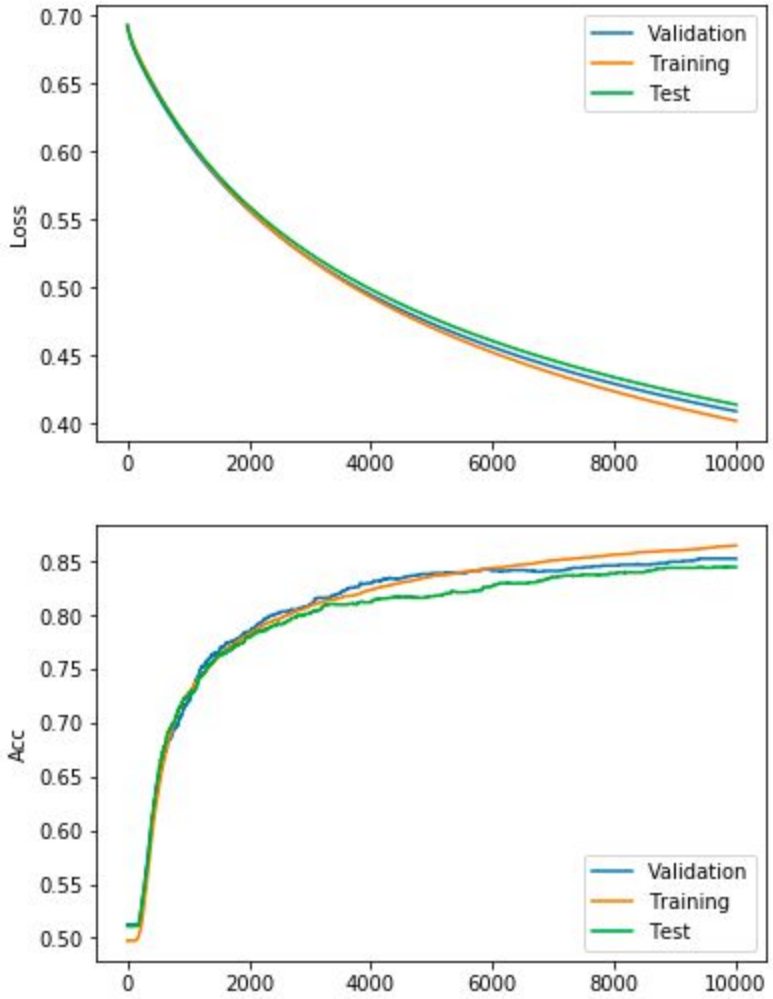


c) GSC Dataset with feature concatenation

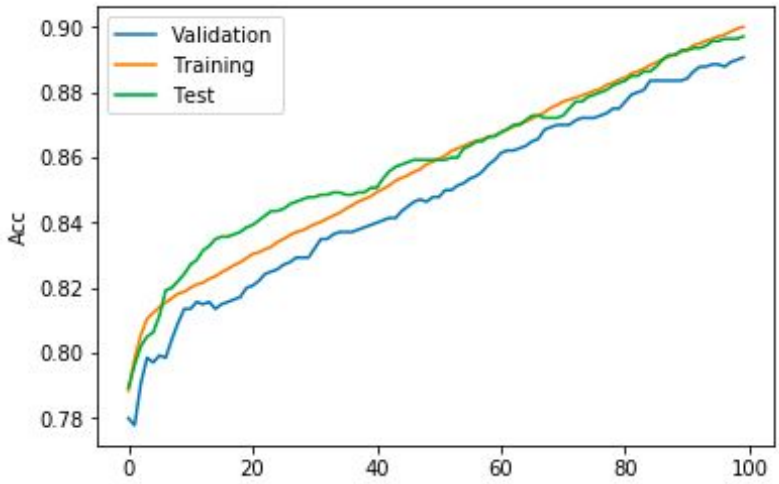




d) GSC Dataset with feature subtraction



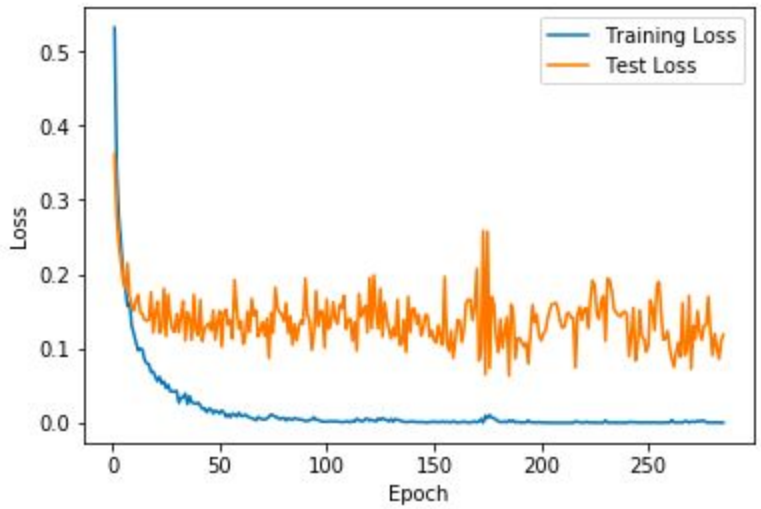
Lower epoch and high learning rate:

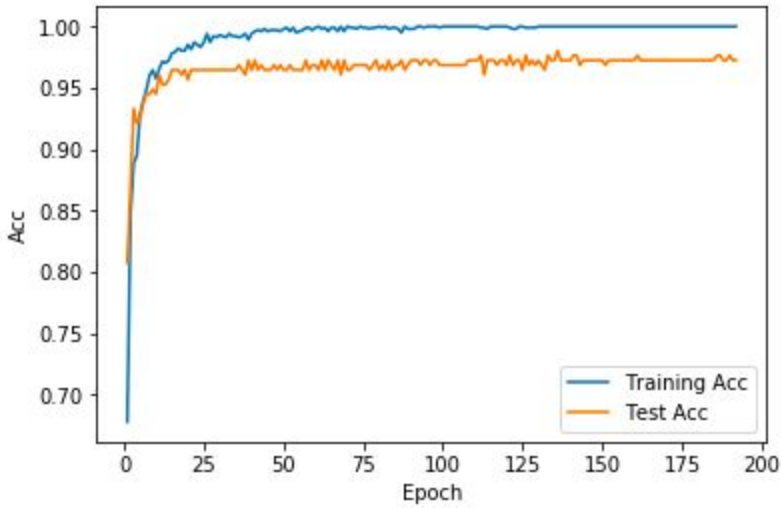


Neural Networks [Bonus]

Num epochs = 700
Model batch size = 128
Batch size = 32
Early patience = 100
Drop out = 0.20
First dense layer nodes = 256
Second dense layer nodes = 256
Third dense layer nodes = 2
Activation = softmax
Learning rate = 0.001

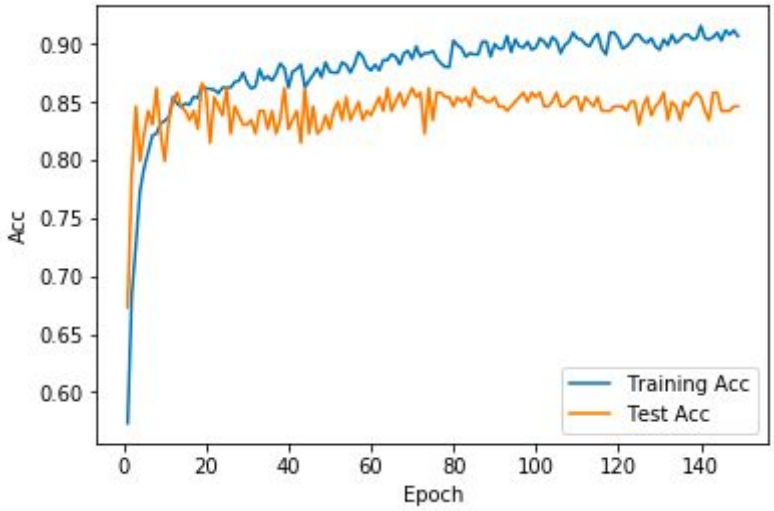
a) Human Observed Dataset with feature concatenation

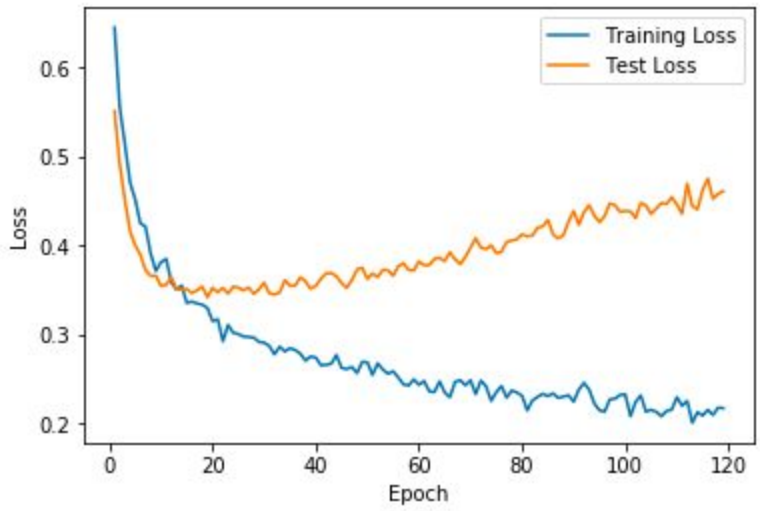




Errors: 3 Correct :155
Testing Accuracy: 98.10126582278481

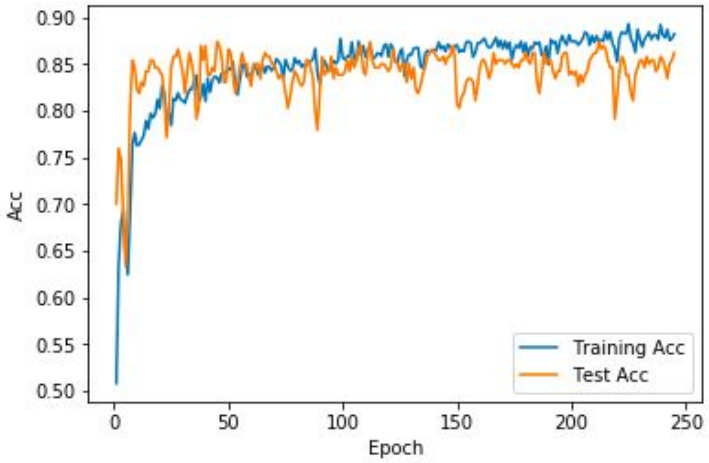
b) Human Observed Dataset with feature subtraction





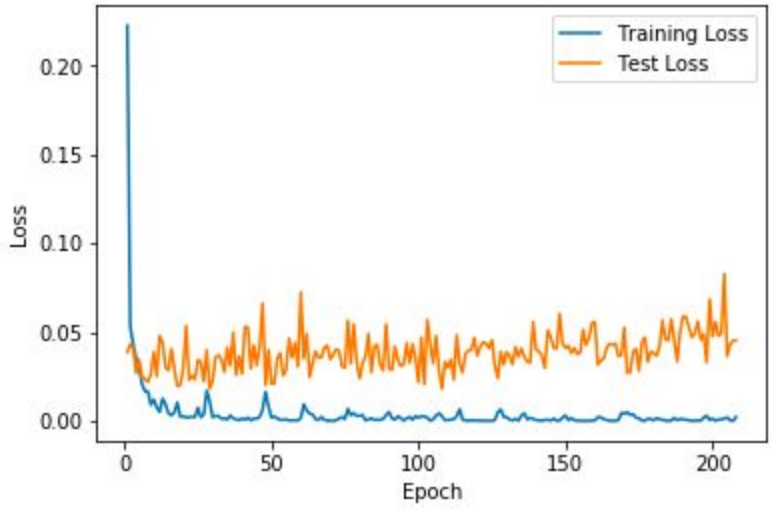
Errors: 22 Correct :136
Testing Accuracy: 86.07594936708861

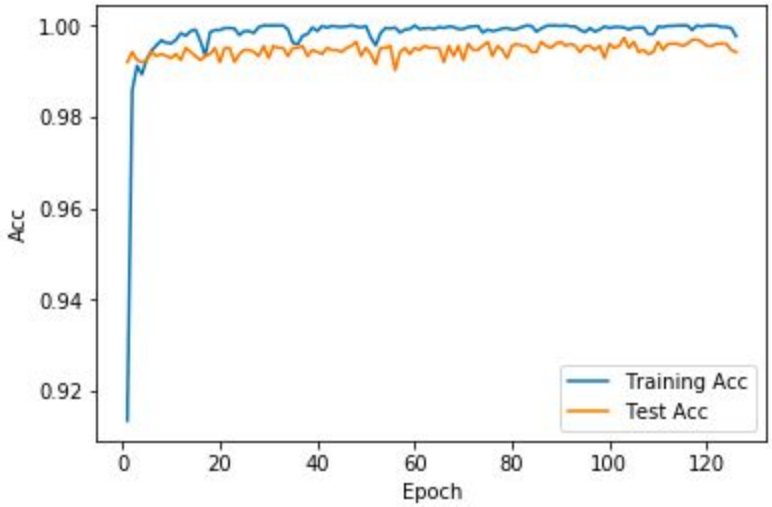
Higher Model batch size = 500, performs better than lesser batch, this is a expected behaviour has it can consume more data at once, but the processing time is much higher.



Errors: 18 Correct :140
Testing Accuracy: 88.60759493670885

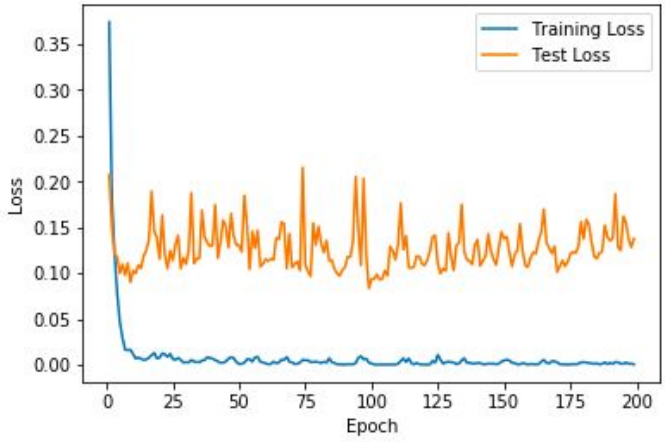
c) GSC Dataset with feature concatenation

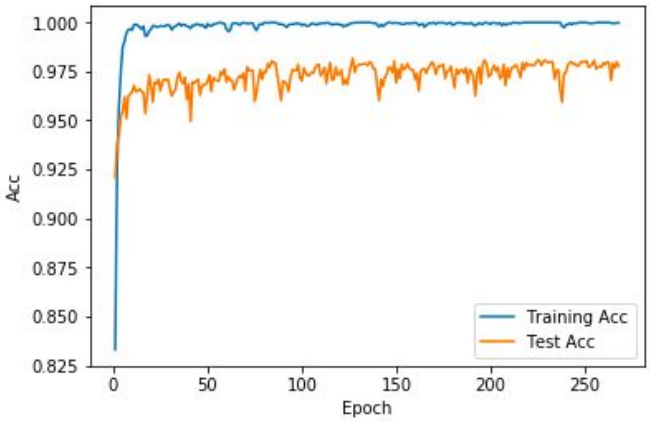




Errors: 15 Correct :1385
Testing Accuracy: 98.92857142857143

d) GSC Dataset with feature subtraction





Errors: 43 Correct :1357
Testing Accuracy: 96.92857142857143