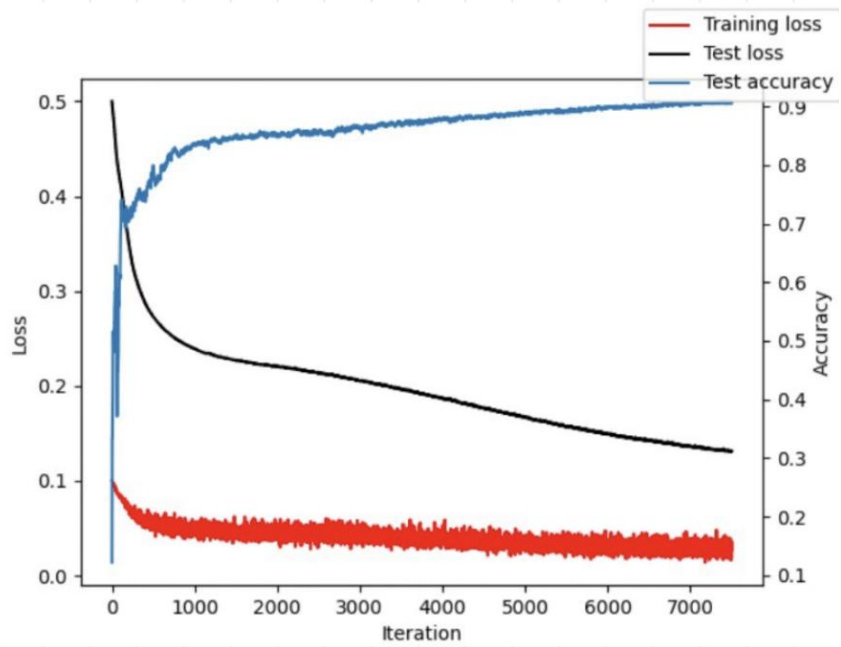
Q1 – Accuracy, Loss and Labeling



The test accuracy was a little over 90% per epoch with:

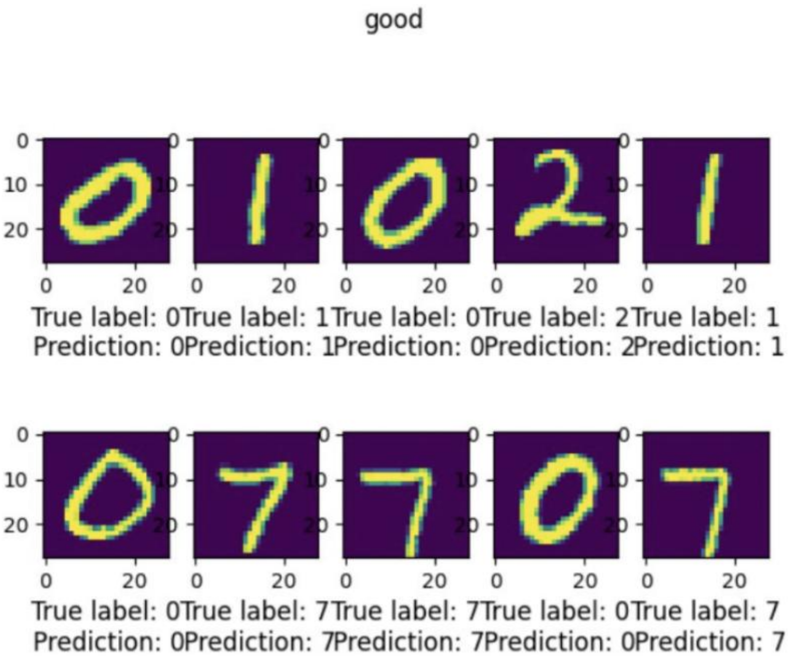
Epoch 0 : 0.904

Epoch 1: 0.917

Epoch 2: 0.918

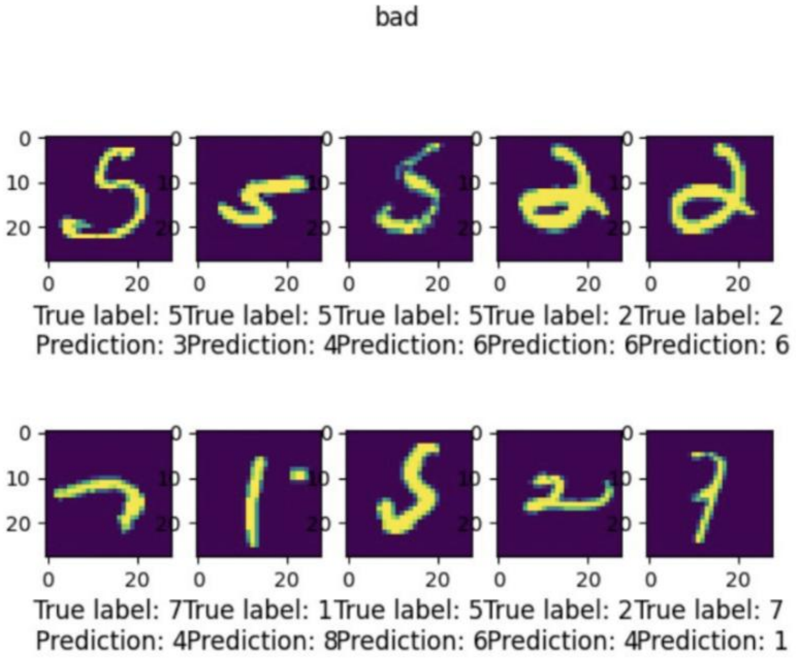
Epoch 3: 0.922

Regarding the labelling we can see:



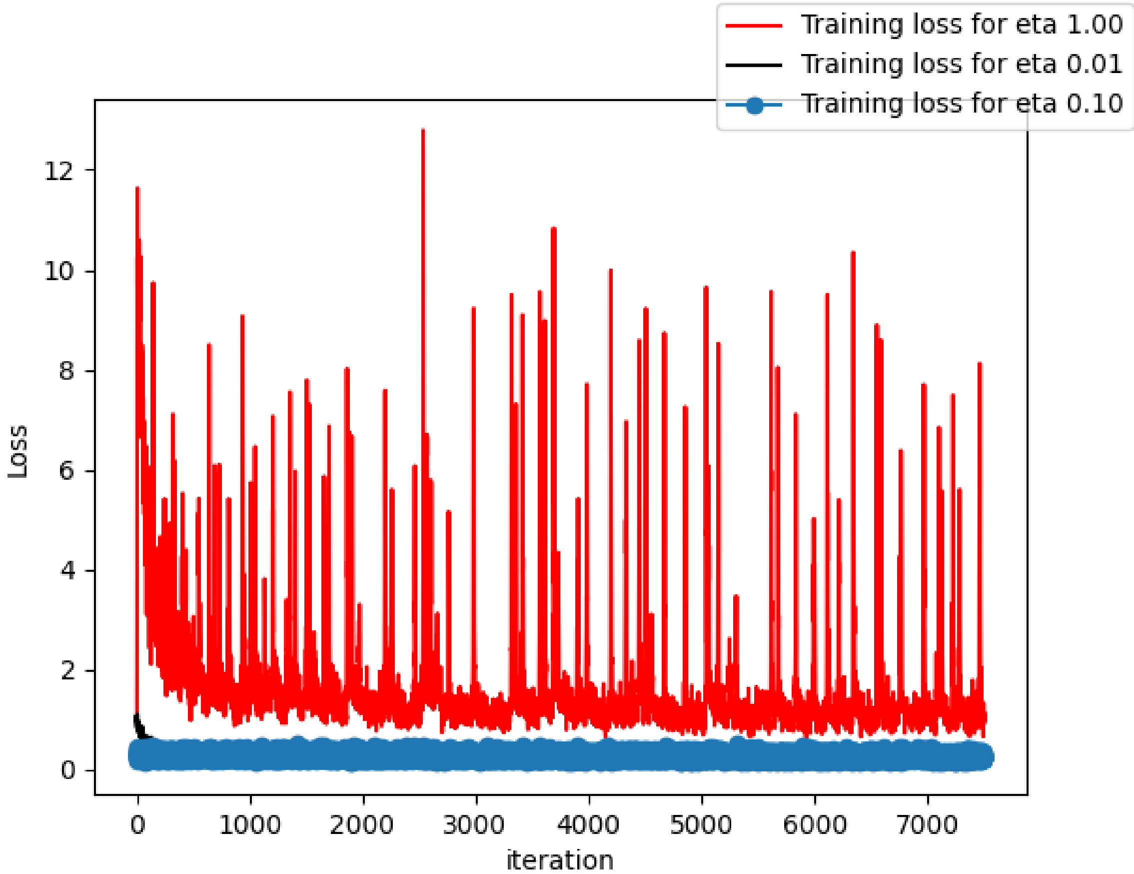
In the first case we observe the images easier to predict and its evident that the handwriting in these images is very neat and readable and so the algorithm was able to easily predict the true label. Lines and circles in each number are deterministic, nothing interrupts them.

Next, we wanted to observe bad predictions:



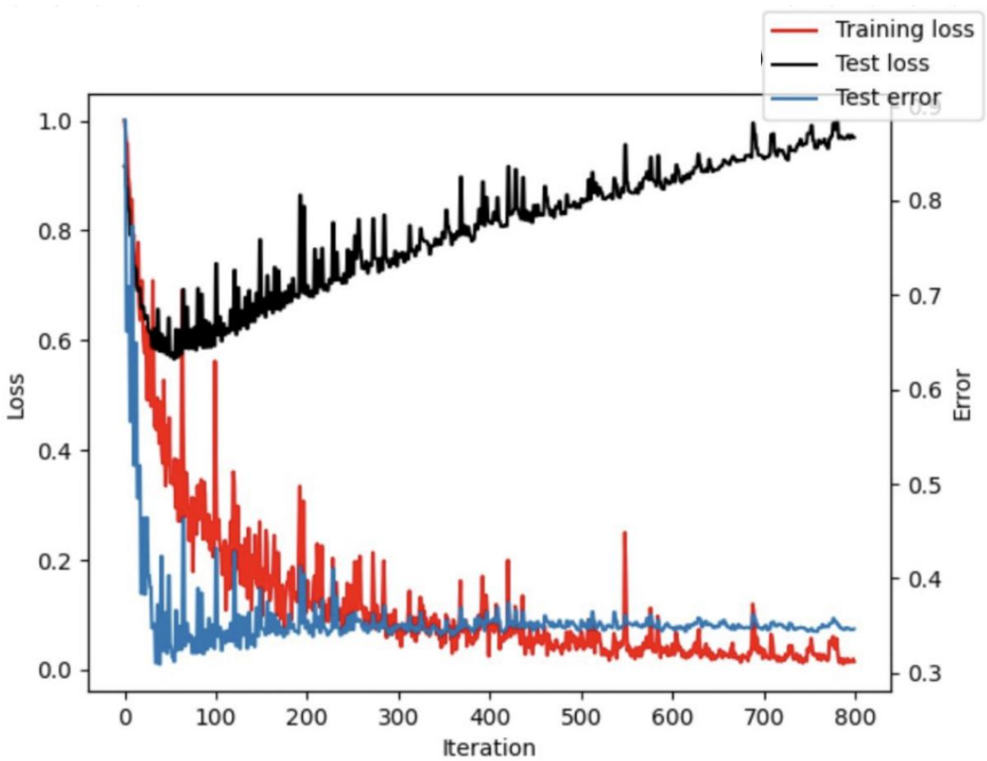
Unlike the previous "good" handwritten numbers, here everything is a bit smudge and more difficult to understand both to people and the algorithm.

Q2 – Train loss per ETA



This graph portrays the differences in training loss per ETA. It's evident that the training loss for ETA = 0.1 is the smallest, followed by 0.01 and 1. This makes sense since if ETA is too large then jumps might be too big and it therefore might never converge to the global minimum. If ETA is too small, out jumps might be stuck on a local minimum instead so it wouldn't converge to the true global minimum. To conclude, an average ETA (not too big or too small) is preferable and would have the smallest loss.

Q3 – Train vs Test (100 samples)

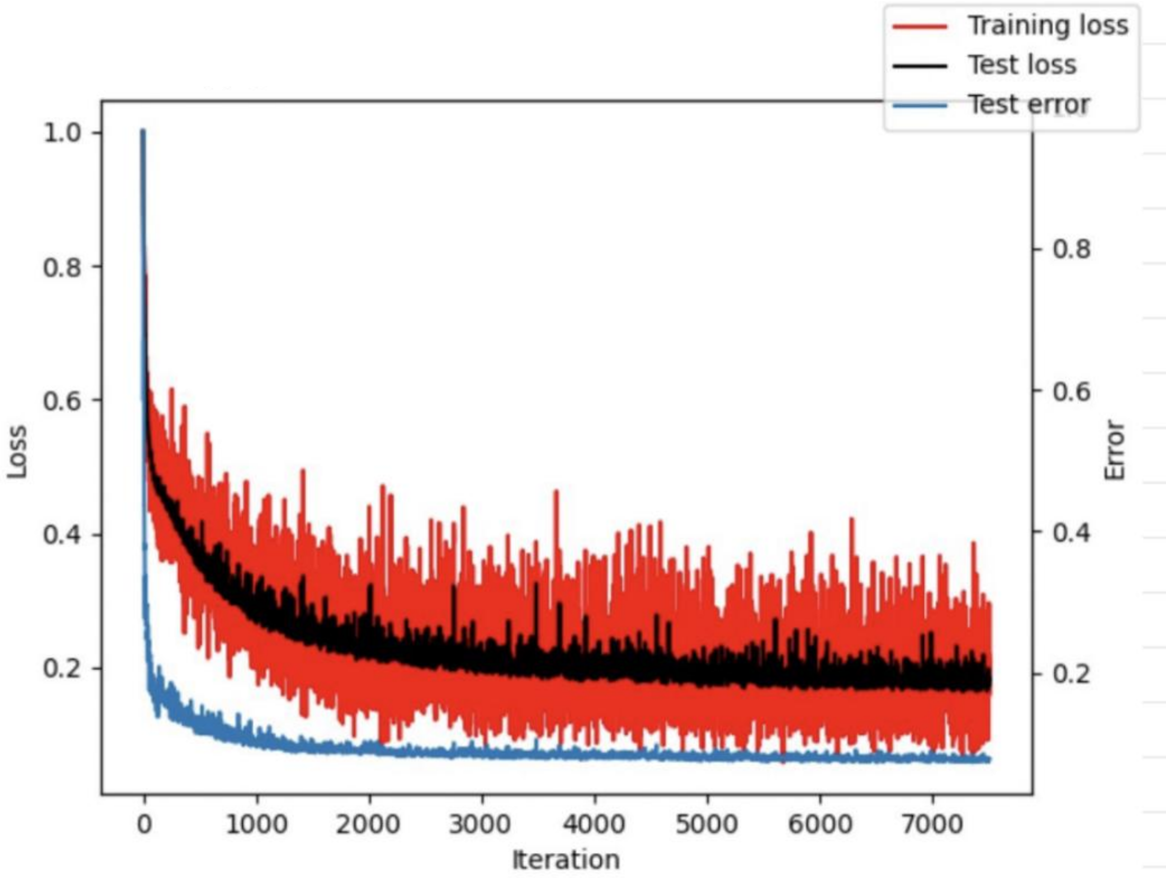


In this example we have 200 epochs, over 10 samples of the 10 digits.

It seems clear that 10 samples per digit isn't enough to capture the variability in images of handwritten digits. It's better than nothing since without training we have 10% accuracy meaning 90% error. But still its not perfect and the error is about 35% meaning 65% accuracy. Yet as observed earlier, we can achieve 90% accuracy.

It's evident that somewhere around 90 iterations there is a sharp increase in test loss simultaneously with a sharp decrease in train loss which means we are overfitting the results to the training data, so the loss gets smaller since we are improving but it's too specific than the test loss gets bigger and we are getting away from the true labels. To conclude, it would be good to stop early (before 90 iterations) to obtain the best predictions.

Q4 – Train vs Test (all data)



In this case, we ran the whole training set (not just a small pre-defined set) and observed a shift in test error which decreased significantly to about 7% (meaning 93% accuracy). Additionally, we observe a significant decrease in test loss which appears in line with the changes in train loss. That is expected (as explained earlier) since here we have much more data to learn from (train on). Unlike in the previous sections, here there is a decrease in both train and test loss which means no overfitting and therefore no need to stop iterations early on.