Sanjay Ananthanarayan

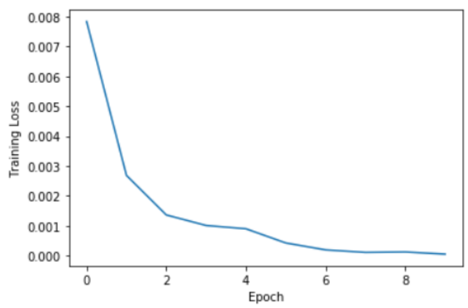
14 December 2019

Machine Learning

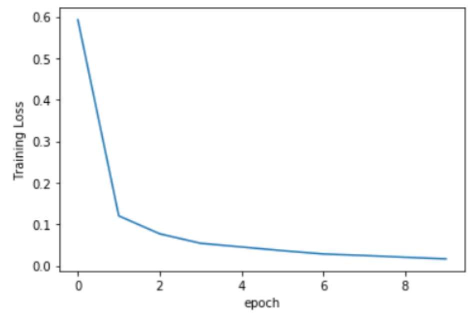
Project Part 2

Part 1:

Different Network



Baseline Network

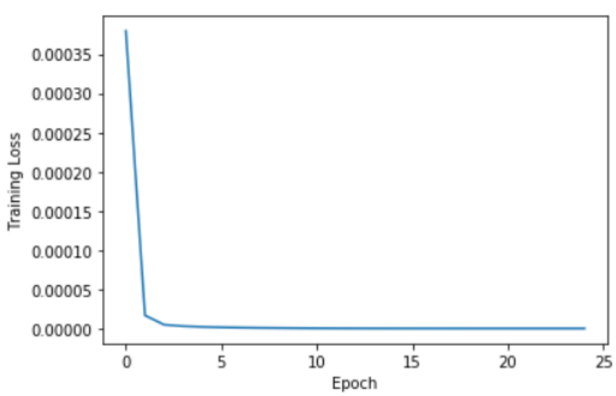


In both figures above, the Adam optimizer was used for both networks. Adam was chosen as it learns the fastest and its good learning rate is between 0.005 and 0.001, so the learning rate used was 0.001 to compare the performance. With the different network, the accuracy was 98.6%, while the accuracy was 98.45% for the baseline network. The different network performed better because training loss improved by reaching zero a lot faster, resulting in a better classification accuracy.

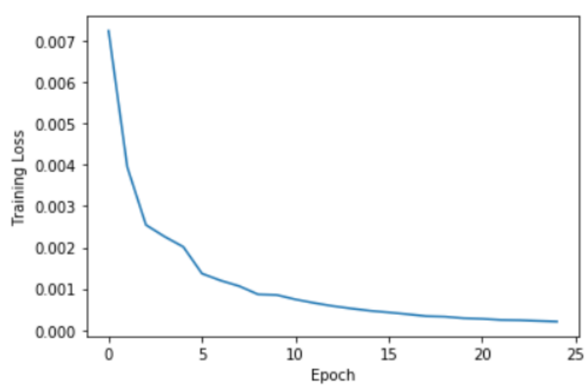
Part 2:

1. Learning rate and Optimizer: Adam or SGD

Best Learning rate for Adam = 0.001

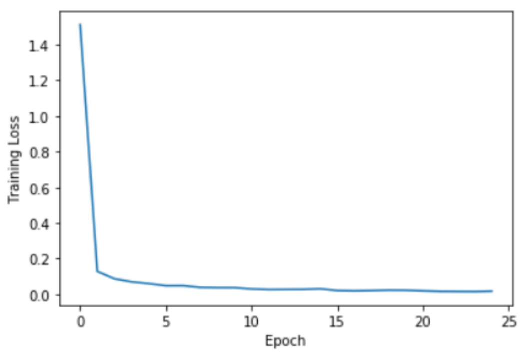


Best Learning rate for SGD = 0.12

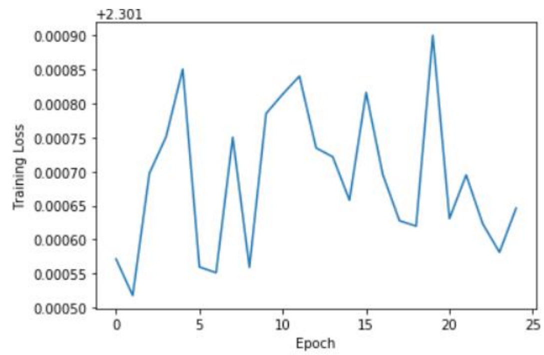


For the Adams optimizer, the best learning rate was 0.001 as this was where training loss converged at zero the fastest, resulting in the classification accuracy being 98.98%. For the SGD optimizer, the best learning rate was 0.12 as this was where training loss converged at zero the fastest, resulting in a classification accuracy of 98.53%. The Adams optimizer is better than the SGD because the network learned faster, resulting in a better classification accuracy.

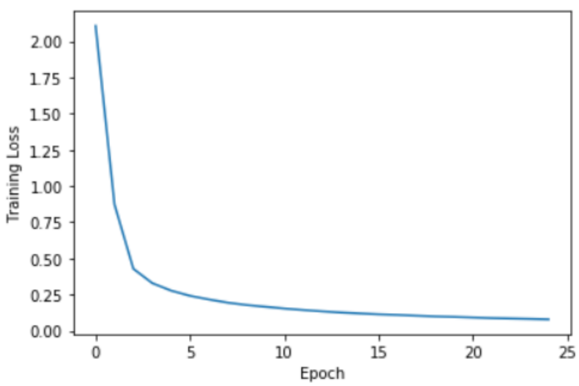
Learning rate for Adam = 10*0.001 = 0.01



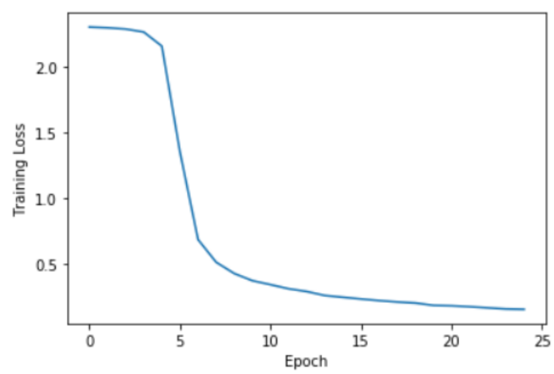
Learning rate for SGD = 10*0.12 = 1.2



Learning rate for Adam = 0.1*0.001 = 0.0001



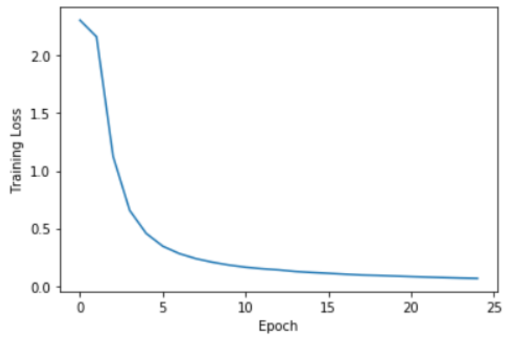
Learning rate for SGD = 0.1*0.12 = 0.012



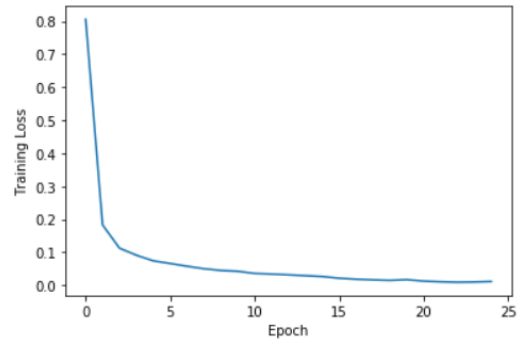
The learning rates of 0.01 and 0.0001 for Adams and 0.012 for SGD all converged at zero, while the learning rate of 1.2 for SGD does not converge. This means the network applied was good for those learning rates, but since SGD of 1.2 is greater than 1, multiple saddle points are created instead of network converging at zero. All 3 learning rates for the Adams optimizers converge to zero a lot faster than the SGD optimizer since Adam learns the fastest.

2. Activation Functions

Sigmoid Activation (LR = 0.001)



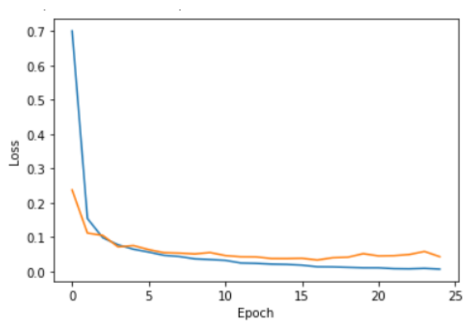
ReLu Activation (LR = 0.001)



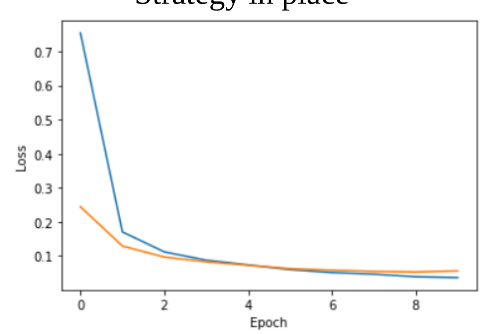
From both graphs, a converge can be seen but the sigmoid graph shows a slower convergence to zero. Also, the classification accuracy for sigmoid is 97.59%, while its 98.89% for ReLu. Due to the slow convergence in sigmoid activation, the ReLu activation function is commonly used these days in all deep learning models as it converges sooner, fixes & avoids the vanishing gradient issue, and is more accurate in training the network.

3. Early Stopping Strategy

No strategy



Strategy in place



Orange line – validation loss, blue line – training loss

The early stop strategy had a classification accuracy of 98.92%, while the classification accuracy for no strategy was 98.78%. Implementing the early stop strategy reduces overfitting to occur on the dataset which in return improves the accuracy. As the accuracy shows, the early stop strategy has very little effect on this network if you have a really good learning rate.

4. Data Augmentation

The classification accuracy with augmentation was 99.19%, while the accuracy with no augmentation was 99.01%. As the accuracy shows, augmenting the training sets improves the training of the network. Improving the model by augmenting the transformation angle by 30 degrees improved the accuracy by 0.18%, meaning augmenting the angle even slightly can improve accuracy but not as much as the learning rate.

5. Network depth vs network width

With a 5-layer network, the classification accuracy was 98.72%, while the accuracy was 98.92% for a 3-layer network. This means that increasing the number of layers in the network does not mean accuracy will increase. Having a deep neural network will improve accuracy only when there is a lot of data in the network.

<u>Part 3:</u>

```
Accuracy of the network on the 10000 test images: 65.820 %
```