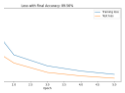
Training results – ViT Light

Dataset: MNIST

Hyperparams:

5 epochs, batch size 32

Adam optimizer lr 0.002



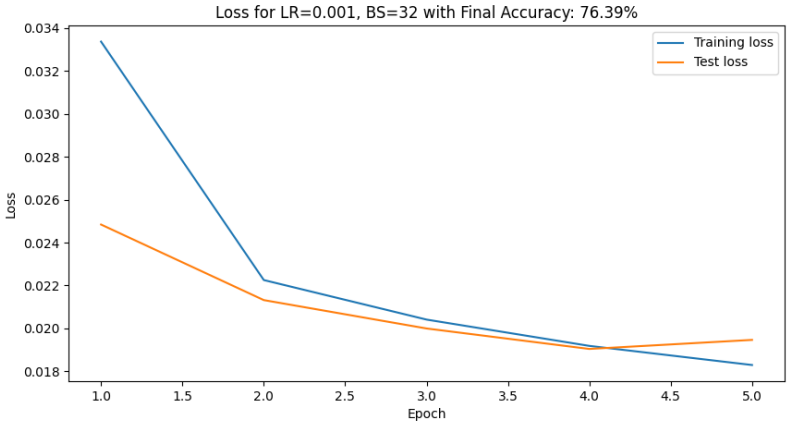
Dataset: Fashion MNIST

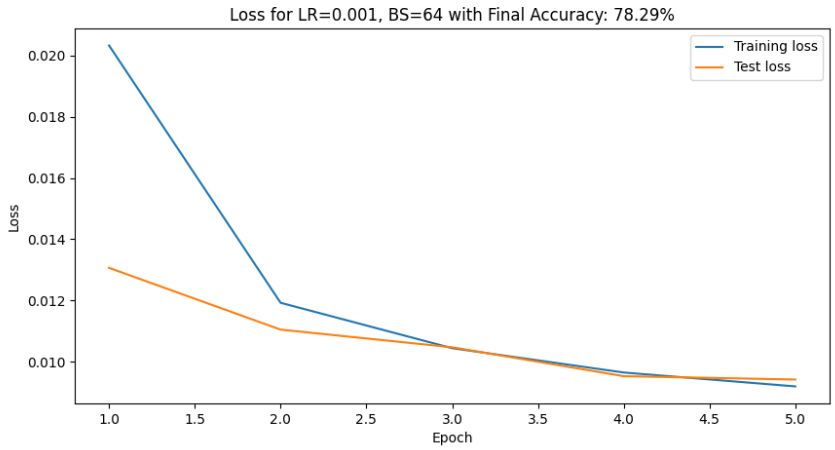
I trained different models with different hyperparameters:

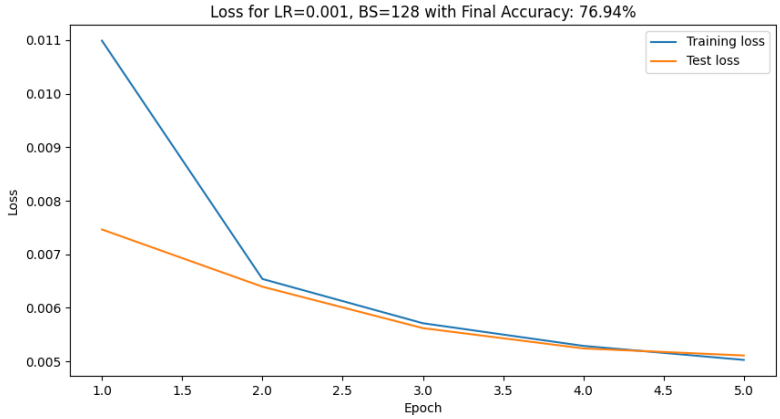
1. Learning rate [0.001, 0.01, 0.1]
2. Batch size [16, 32, 64]

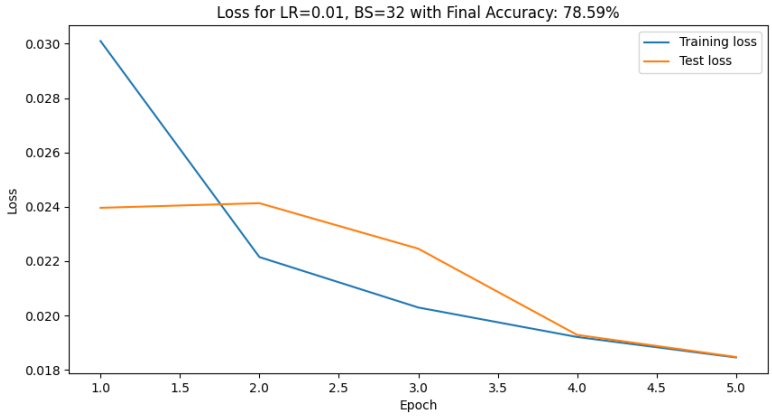
I plotted each of them and saved the best model. I only used 5 epochs because of computational resources I had.

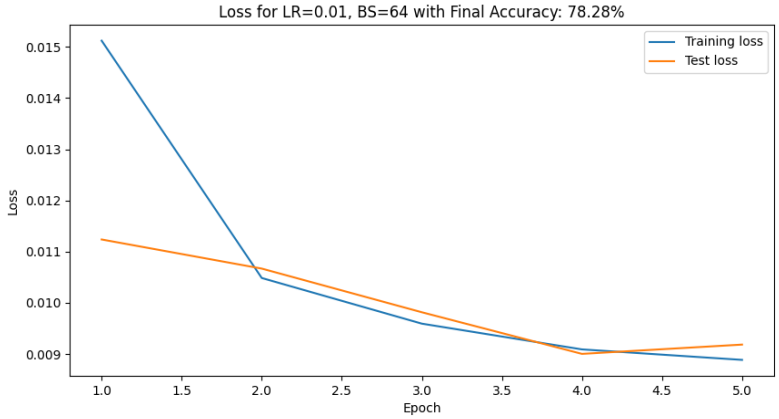
Results:

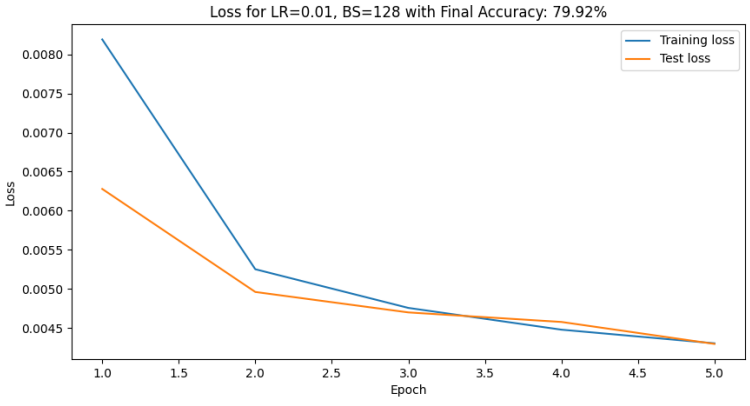


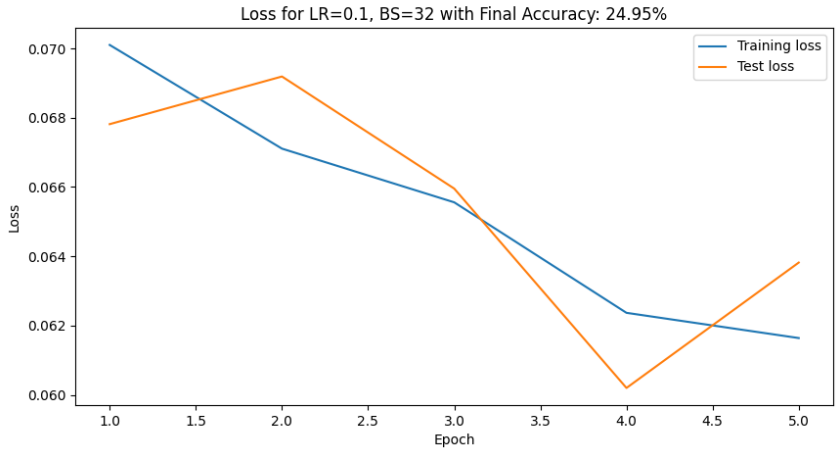


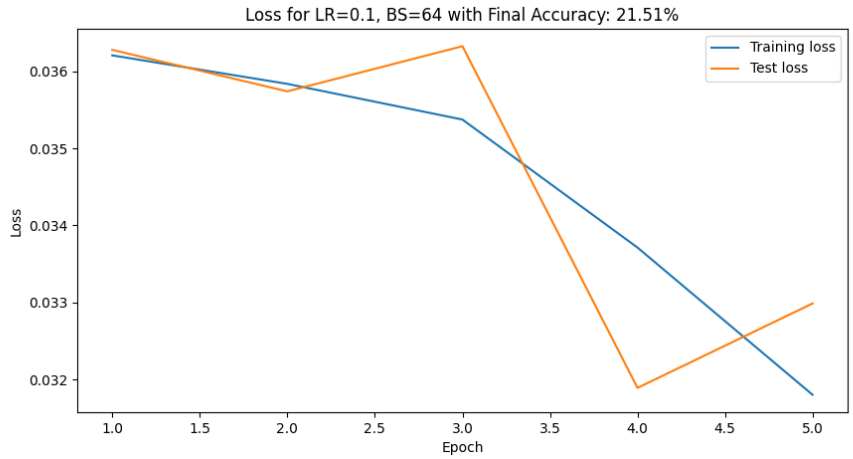

Loss for LR=0.001, BS=128 with Final Accuracy: 76.94%
Training loss
Test loss
Loss
Epoch

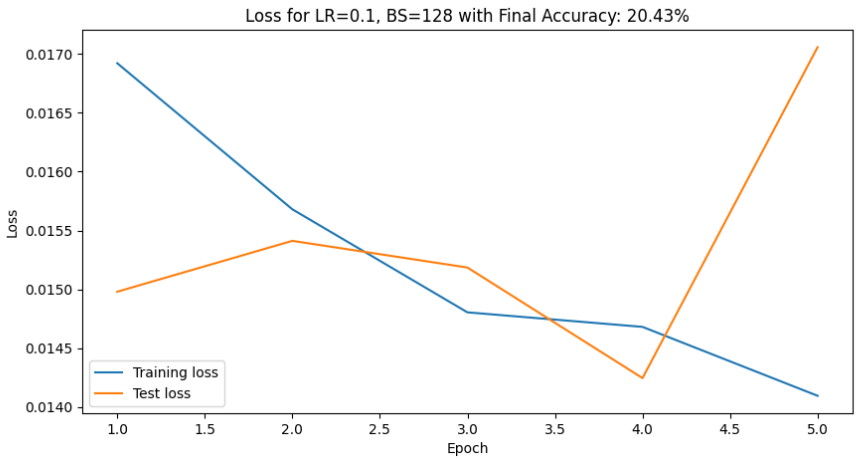

Loss for LR=0.01, BS=32 with Final Accuracy: 78.59%
Training loss
Test loss
Loss
Epoch


Loss for LR=0.01, BS=64 with Final Accuracy: 78.28%
Training loss
Test loss
Loss
Epoch


Loss for LR=0.01, BS=128 with Final Accuracy: 79.92%
Training loss
Test loss
Loss
Epoch







The best result on Fashion MNIST is 78.59 with batch size 32 and learning rate 0.01.

The results I achieve are low in comparison to ViT, but I believe that is because the structure we use is too simple. They achieve 99% accuracy on MNIST and 92% on Fashion MNIST, but they use more heads and more attention layer. They also use dropout for regularization:

Config	MNIST and FMNIST
Input Size	1 X 28 X 28
Patch Size	4
Sequence Length	7*7 = 49
Embedding Size	64
Parameters	210k
Num of Layers	6
Num of Heads	4
Forward Multiplier	2
Dropout	0.1

Original ViT-Base has 86 million, the ViT I use for comparison has 200k-800k. Our model has only 1994 trainable parameters.