Nathan Butler
MS Robotics
Oregon State University

Feb 9, 2025

DEEP LEARNING — Assignment 2

1 Feed Forward Neural Network - Performance and Tuning

In this assignment, I implemented a basic feed forward neural network for the CIFAR image dataset capable of achieving approximately 82% validation accuracy on the testing dataset. This network uses ReLU activation for the intermediate layers and sigmoid activation for a single binary output class. The following sections discuss experimental results for the following hyperparameters: batch size, learning rate, hidden layer units, and number of hidden layers.

Unless otherwise stated, the networks were trained with 3 hidden layers of dimension 128, batch size 256, and step size 0.001. All results use a momentum value of 0.8 with weight decay 0.00001.

1.1 Batch Size

Figs. 1, 2, and 3 display the training and testing results for training batch sizes 64, 128, and 256, respectively.

These results show a couple of trends. First, smaller batch sizes yield greater training iterations. This is evident, as batch size 64 reached 40,000 iterations while batch size 256 reached only 10,000. However, as a result, backpropagation occurs much more frequently for smaller batches. We see the negative impact of this, as batch sizes 64 and 128 appear to overfit the training data by the end of the training cycle, ultimately seeing high validation loss.

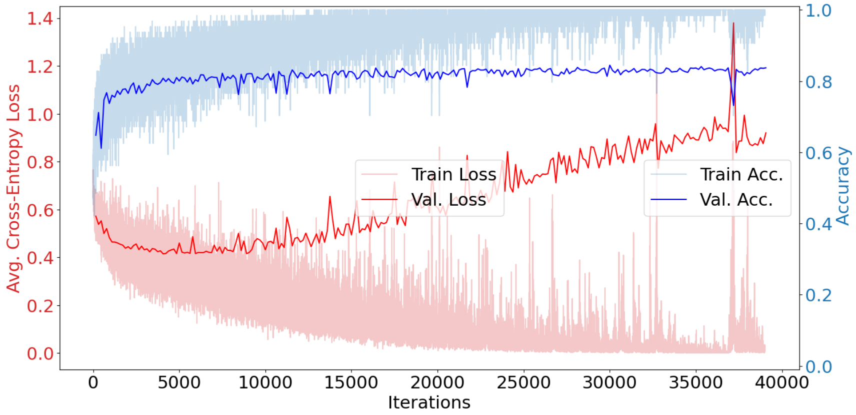

Figure 1: Training and testing results for batch size 64

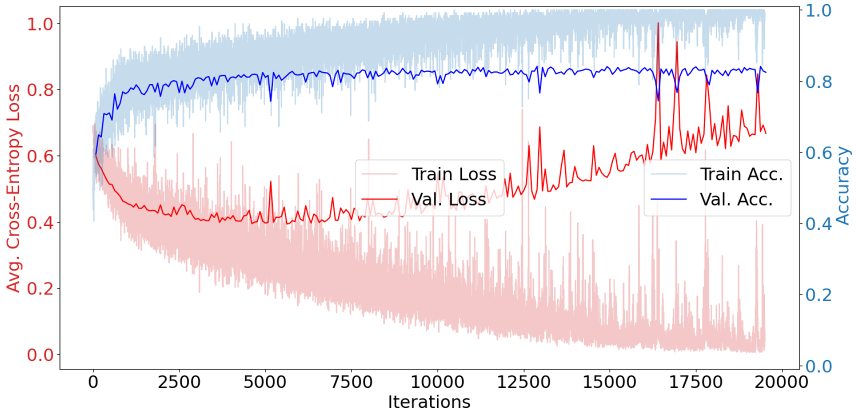

Figure 2: Training and testing results for batch size 128

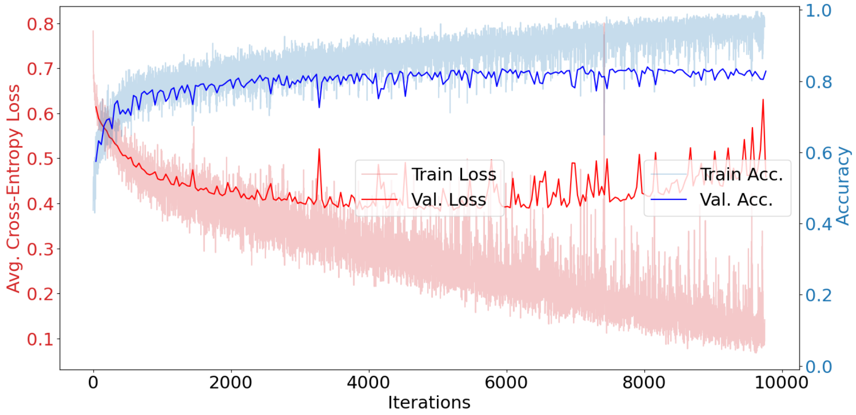

Figure 3: Training and testing results for batch size 256

1.2 Learning Rate

Figs. 4, 5, and 6 display the training and testing results for learning rates 0.01, 0.001, and 0.0001, respectively.

These results serve to emphasize the impact that learning rate can have on the "smoothness" of training. By limiting the size of the gradient update, step size 0.0001 achieves gradual and controlled training performance. However, if training time was limited then a faster method would be desirable.

On the other end of the spectrum, step size 0.01 rapidly converges to accurate predictions, but sees unstable loss behavior. We can observe especially unstable behavior when loss reaches 0.0.

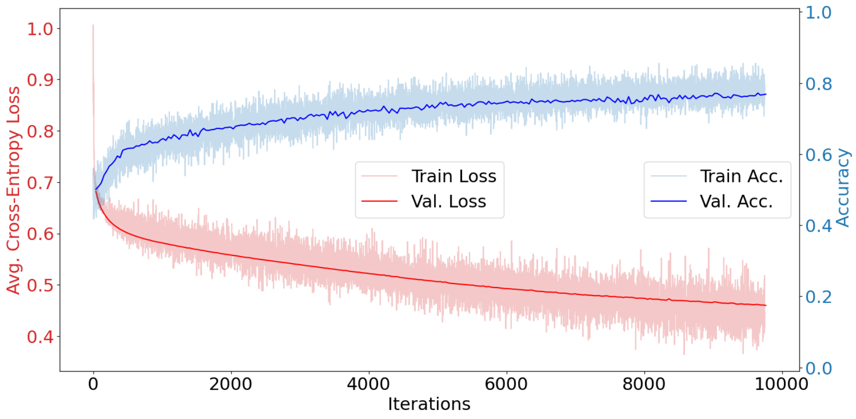

Figure 4: Training and testing results for learning rate 0.0001

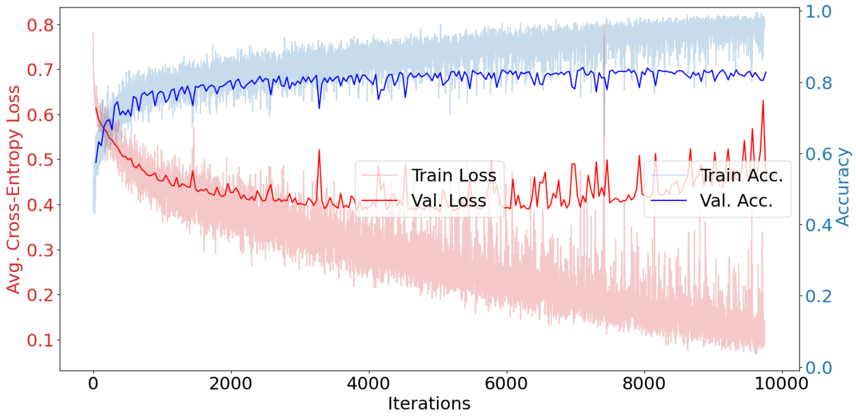

Figure 5: Training and testing results for learning rate 0.001

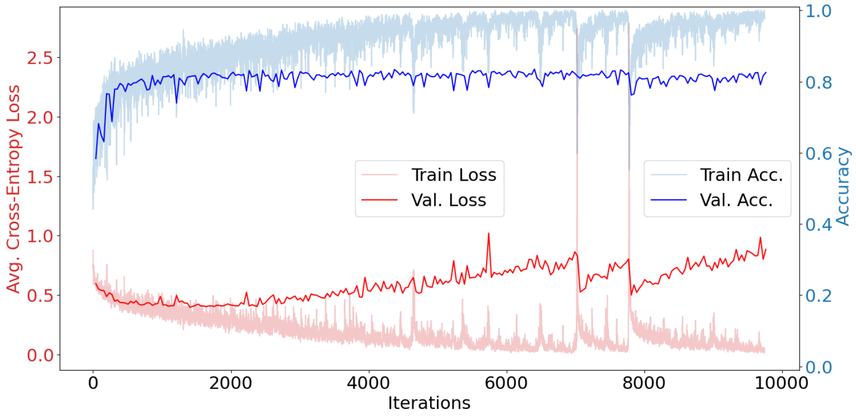

Figure 6: Training and testing results for learning rate 0.01

1.3 Hidden Layer Width

Figs. 7, 8, and 9 display the training and testing results for hidden layer dimensions 64, 128, and 256, respectively.

The width of the hidden layers appears to impact the overall network stability during training. For the smaller dimension 64, training validation loss is jittery. By contrasting this with the relatively stable performance observed with 256 hidden dimension nodes, we observe that greater node counts yields more stable behavior.

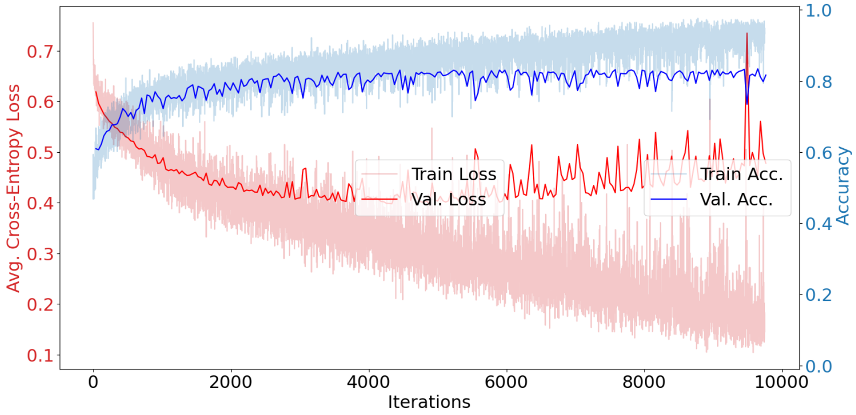

Figure 7: Training and testing results for hidden layer dimension 64

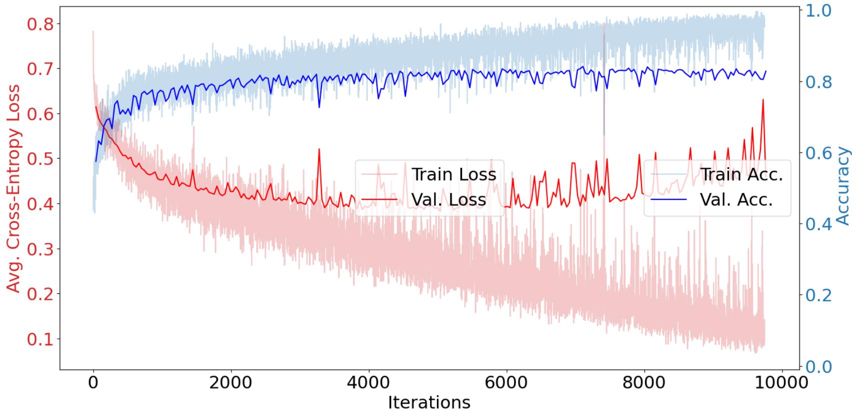

Figure 8: Training and testing results for hidden layer dimension 128

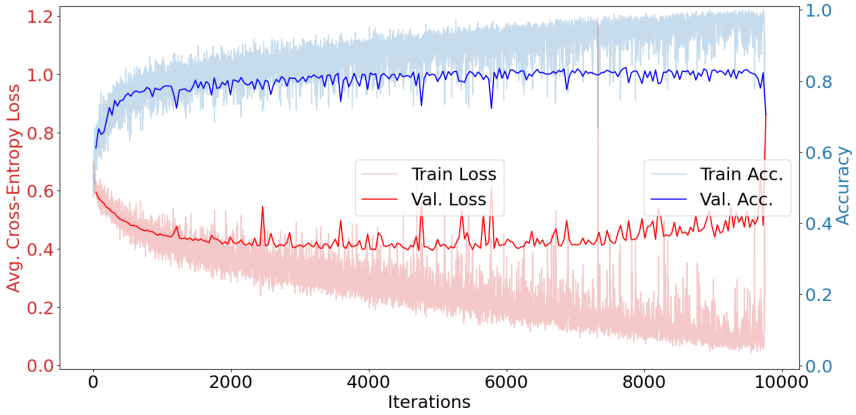

Figure 9: Training and testing results for hidden layer dimension 256

1.4 Number of Hidden Layers

Figs. 10, 11, and 12 display the training and testing results for 1, 2, and 3 hidden layers, respectively.

These results also indicate that higher parameter counts can lead to more stable training performance, as indicated by the width of the training loss band. However, training stability does not always lead to improved validation accuracy, as all results achieve approximately 80% accuracy on the testing data. Rather, the larger networks may just overfit on the training data.

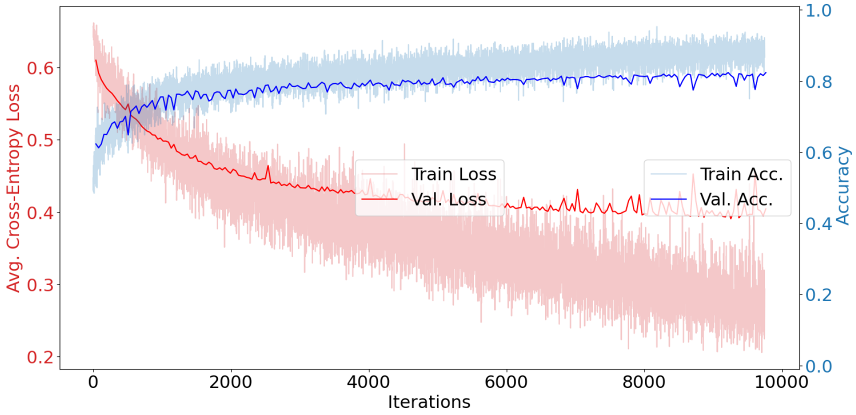

Figure 10: Training and testing results for 1 hidden layer

Submitted by Nathan Butler on Feb 9, 2025.

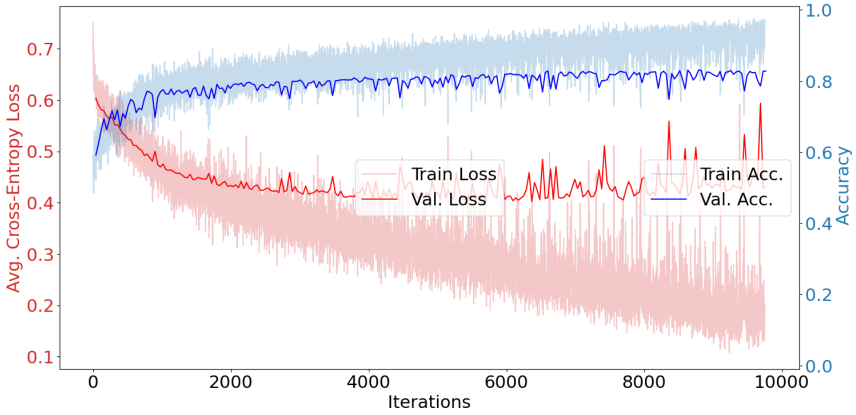

Figure 11: Training and testing results for 2 hidden layers

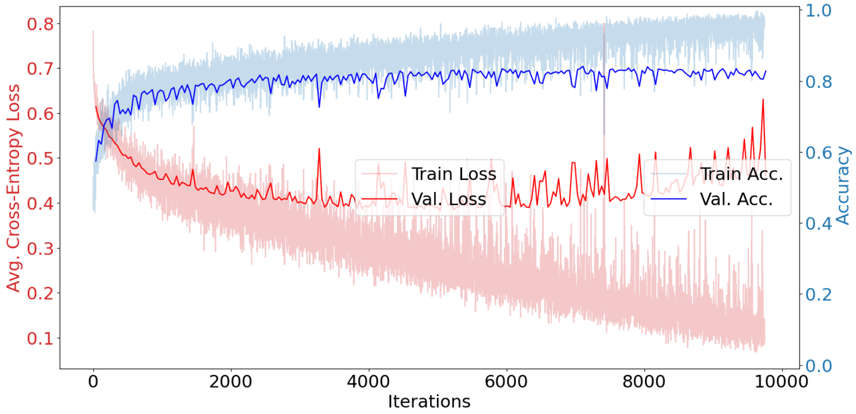

Figure 12: Training and testing results for 3 hidden layers