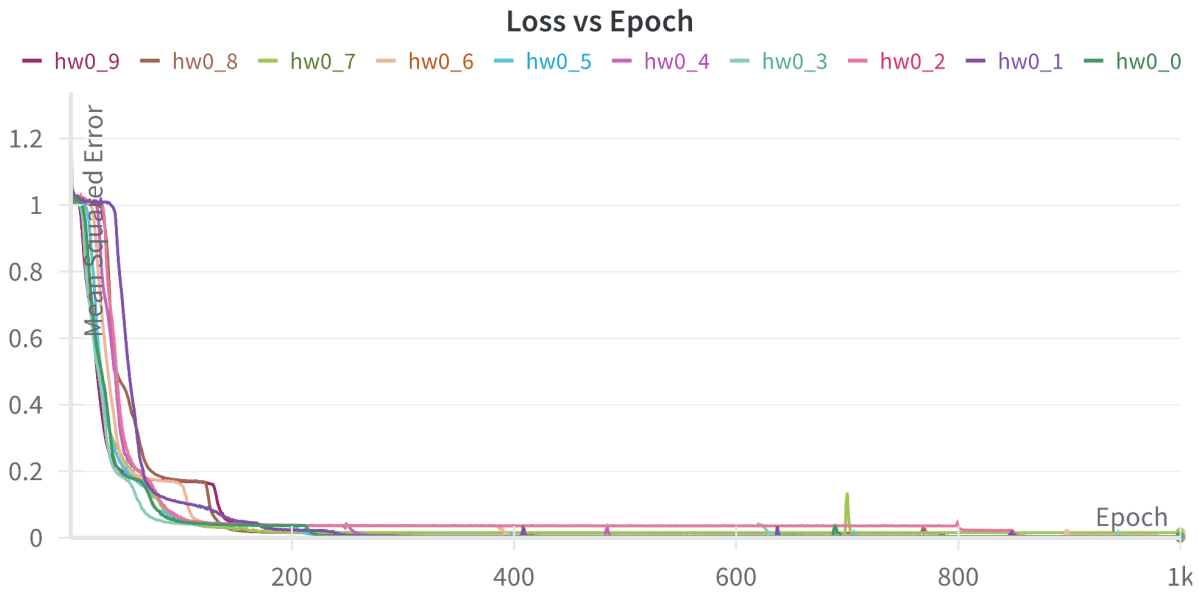
Brandon Michaud

C S 5043 Homework 0

Network Architecture

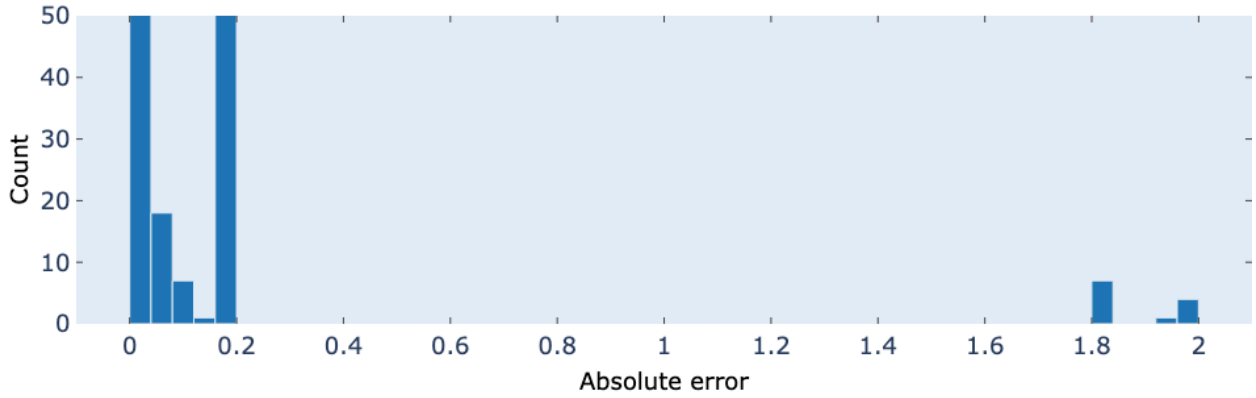
The final architecture I landed on has the input layer, then a hidden layer with 11 neurons, then a hidden layer with 5 neurons, and then the output layer. All layers are fully-connected and sequential. All neurons use a hyperbolic tangent activation function to match the data to have a range of -1 to 1. I used the Adam optimizer. I found that a learning rate of 0.005 was optimal for this architecture and that 1000 epochs was enough for the loss to converge.

Loss vs Epoch



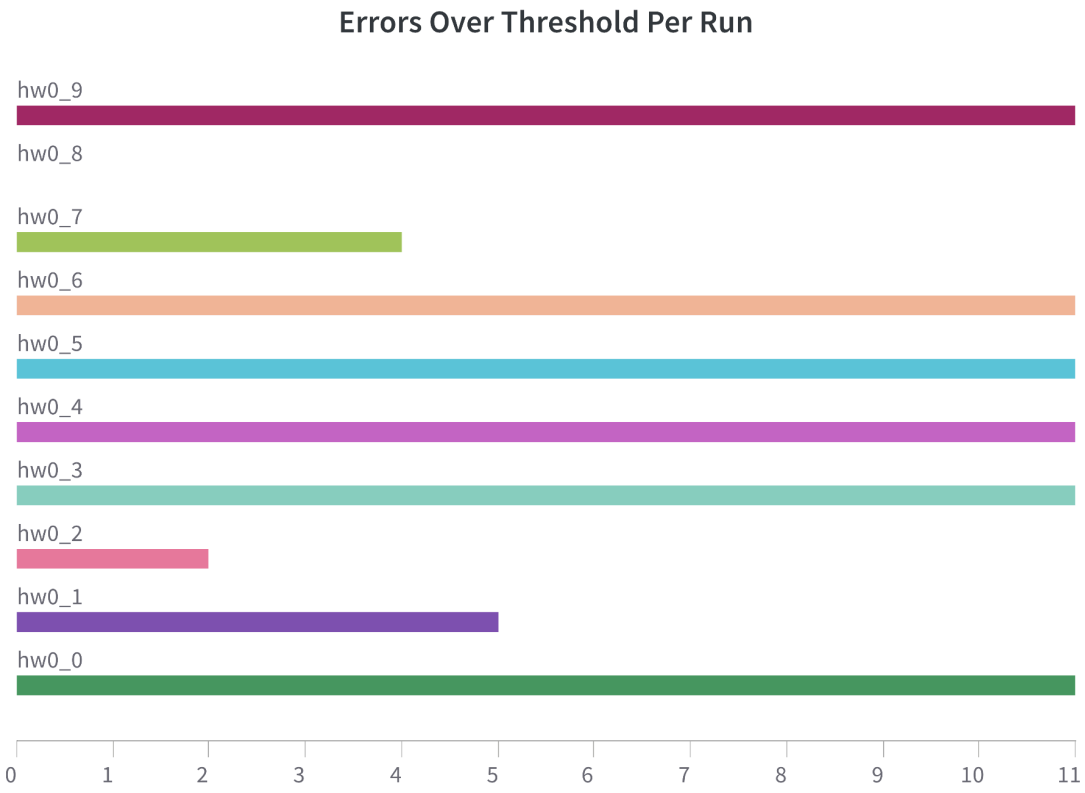
The loss used was mean squared error. Some runs plateau, then start reducing the loss shortly later, before plateauing again, while others plateau only once. It depends on the random initialization of the model and what the gradient is. After about 200 epochs every run stops greatly decreasing the mean squared error, but some do make minor improvements. Some runs have spikes where the loss greatly increases and then decreases back to the level it was before in short succession. This is likely a result of the Adam optimizer escaping the minimum temporarily due to momentum or some distant gradient. Every run converges to about the same small value for loss.

Errors Across All Runs



The errors are concentrated at the two extremes. The vast majority of errors are close to zero and start trailing off before spiking again at around 0.2. On top of this, there is a small number of errors that are very large at over 1.8. This indicates the model fits most data almost perfectly, but some specific points very poorly.

Errors Over Threshold



The x-axis is the number of absolute errors over 0.1. Each run only has a handful of errors over the threshold. None exceed 11. Run 8 was the only run that had no errors over the threshold. The

number is dependent on the random initialization of the model and how that lends itself to a gradient which the optimizer can traverse.

Sum of Errors

The x-axis is the sum of absolute errors for all predictions in the run. It varies greatly across runs and is again dependent on the random initialization of the model. All runs had a sum less than 9 but greater than 1. Even runs with the same number of errors over the threshold vary greatly in sum of errors.

Maximum Error

The x-axis is the maximum absolute error. There is not as much variation in this chart. That is because most runs have a maximum error close to the theoretical maximum for my architecture, which is 2 (1 when target is -1 or -1 when target is 1). When paired with the sum of errors, this indicates that the model is getting a very small subset of examples very wrong and the rest mostly correct. Notably, run 8, which had no errors over the threshold, has the lowest maximum error by far. The optimizer was able to find an almost perfect solution in this case.