Adaptive Hyperparameter Search for Regularization in Neural Networks

By Devin Lu

Research Question

Traditionally, regularization constants and other hyperparameters are fixed for a model throughout training. Optimizing these hyperparameters is usually done through splitting out a portion of the training data into an evaluation set separate from the test data specifically for use in hyperparameter optimization. The model is then trained repeatedly using different hyperparameter settings and an optimal choice is made from performance on the dev set. The downside of this approach is that this is often very expensive, such as when training the model is expensive or when data volume is large enough that training examples cannot all be held in memory. These are issues that often arise with neural networks. Here, we examine the question of whether we can develop a policy that will

Problem Statement

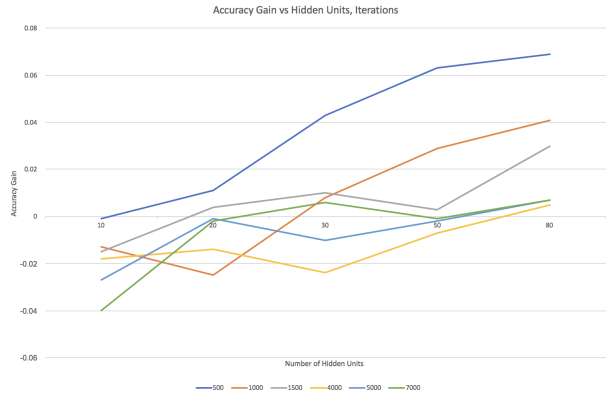
Can we achieve comparable or better validation accuracy from a regularization constant (or other hyperparameters) that vary during training than from fixed hyperparameters? Also, can we achieve comparable validation accuracy to training with the optimal hyperparameters in less iterations using a policy of adaptively changing hyperparameters?

Dataset

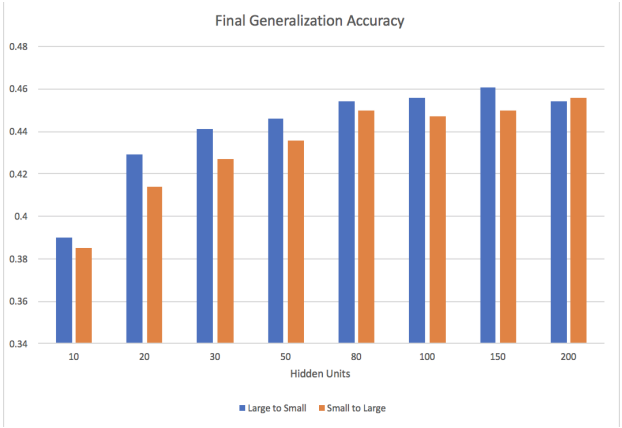
We used the CIFAR-10 dataset. This dataset is 60,000 images each of which is 32 x 32. Each image is colored, so with the three color channels each image consists of 32 x 32 x 3 = 3072 floating point values.
The dataset was split into a training set consisting of 50,000 images and a test set consisting of 10,000 images.
The total size of the dataset in memory is approximately 163MB.

NN Comparison

As a first test we implemented a feed forward neural network. We tried using a schedule that started with high regularization that gradually reduced. We compared this against using any of the individual regularization parameters in the schedule.



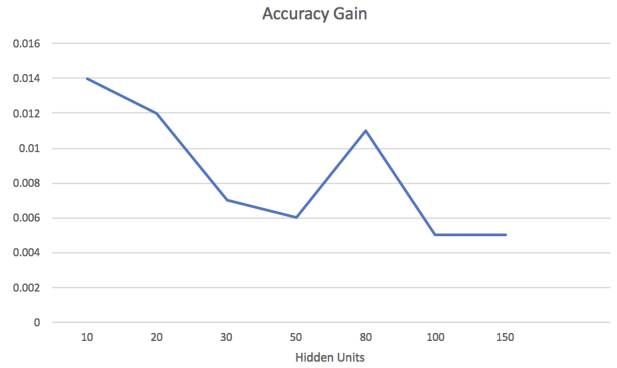
The above image shows the increase in accuracy from using a varying schedule over using the best fixed regularization parameter. To make the comparison fair, each comparison model was trained for as many iterations in total as during the varying schedule training.



We also observed that contrary to expectations, a schedule that started with low regularization and gradually increased performed worse than a schedule that started with high regularization and decreased.

Adaptive Scheduling, CNN Approach

We also experimented with a that chose its regularization updates adaptively. Here, we would intermittently check the training vs. validation accuracy, and it would adjust the regularization parameter up or down depending on whether this reading indicated overfitting.



We also implemented a three-layer convolutional neural network and applied these same techniques. We observed similar results.

Analysis

We see that non-constant regularization schedules can bring an accuracy gain to neural network based classifiers. We show this is not simply a matter of expanding the search space of regularization parameters, as it is possible to achieve accuracy even better than any of the individual regularization parameters searched trained for an equivalent amount of iterations.

In general, we see that the gains from this method decrease with the total number of iterations trained, suggesting it helps accelerate learning given a fixed training time, but with more training time a blind search starts closing the gap.