Mixture of Experts Using K-Means Compared Against Ensemble Learning

Adam Ali

Nicolas Perez

Intro

In this project, we examined the differences in performance of a mixture of experts and the ensemble method of machine learning on the fashion MNIST data set. The fashion MNIST data set consists of 70 000 black and white images, each one labeled as 1 of 10 different types of clothing articles and each image being 28 by 28 pixels in size.

Background

In machine learning the mixture of experts technique involves having multiple neural networks all trained on different portions of data which is divided using something called a gating function. In our project the gating function we used was a simple k-means clustering of the data. We have a neural network associated to each cluster, so that when a new piece of data is presented during the testing phase, we check first which cluster the data is closest to, which it is then passed onto the corresponding network to that cluster for classification and then gives a single prediction.

The ensemble method is a technique that involves combining the results of multiple different neural networks to get a single result. In our project we trained multiple neural networks, each on disjoint subsets of the training set for variance amongst the different networks and faster training. The idea behind the ensemble method is that having variance among different trained neural networks and then combining their results will yield a higher accuracy than any one of the trained neural networks on its own (this is something that has been proven to work before). During the testing phase, when a piece of data is fed into the ensemble it is fed into each one of the neural networks and then all of their predictions are summed together, and the class with the highest outputted value after the resulting summation is then used as the actual prediction of the ensembled network.

Problem statement

Talk strictly about what it is your are trying to solve. This can be one or two sentences. “What is the best neural network classification method out of either the averaging ensemble method on disjoint subsets of training data, mixture of experts with a k-means gating function or a single neural network trained on the entire data set?”

Results and Analysis

Implemented using tensorflow, ran experiments on computer with certain stats x y z. Show graphs. Compare with normal method. Show accuracy and time to train.

Conclusions

Key messages, and what I would do in the future.

For this specific domain, one method worked well, or both didn’t etc. Maybe one worked better but took more computation. More experimentation needed to prove a general case (one is strictly better). Summarize the whole paper.

TO DO: run both methods across a range of different number of neural networks. Start each method with 2, then 3, then 4 etc neural network models trained at once. Possibly they work both as well just with different numbers of neural networks.