COMP 4107
NEURAL NETWORKS

MIXTURE OF EXPERTS USING K-MEANS AGAINST ENSEMBLE LEARNING

ADAM ALI
101004367

NICOLAS PEREZ
100978917

Table of Contents

1. Introduction

For the following project, we have compared the performance and training times of two types of machine learning methods on the fashion MNIST dataset. Both the mixture of experts and the ensemble methods acted on the multilayer perceptron model of a neural network. We trained the neural networks for 1 and 3 epochs and compared the accuracies and training time for both methods. For the control, we included the basic M.L.P network to compare training times and performance.

2. Background

The experiment was done on the fashion MNIST dataset. This dataset consists of 70,000 black and white images which are 28 by 28 pixels in size. Each image is labeled as 1 of 10 different types of clothing such as trousers, dresses, or coats (Figure 1). Using k-fold cross-validation, the set can be divided into 63 000 images for the training set and 7 000 images for the test set. The fashion MNIST data was used due to our observation that the regular MNIST is relatively easy for basic feed-forward networks to classify, and many neural network exercises use the regular MNIST dataset.

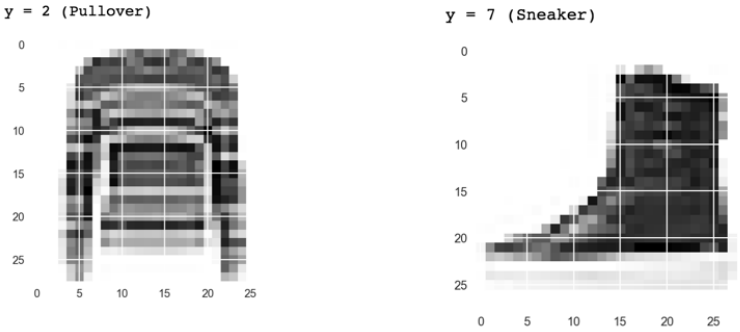

Figure 1. Two images from the fashion MNIST dataset, with labels “Pullover” and “Sneaker” . [1]

2.1. Mixture of experts

The Mixture of Experts technique was developed in the early 1990s to deal with extremely large datasets. In machine learning the technique involves having multiple neural networks all trained on different portions of the data [2]. A process called specialization has each neural network improve performance on a subset of the whole dataset. This requires the data to be separated into regimes and given to different models during training. The idea is to identify the model which is doing better than the others on a particular subset of the

data and have it focus on predicting the right answers on the subset while ignoring others, which leads to specialization. In order to match each model to a regime of the dataset, we need to cluster the data using a gating function [3]. We use a simple k-means clustering method on the data and associate a neural network with a specific cluster. When new input is to be given to the whole model, the cluster closest to the input is computed and the associated model is responsible for the classification. Ultimately the mixture of experts technique uses the dataset in a unique way, in that if the dataset is not large enough, it will not make good use of the data due to the partitioning that occurs over multiple models. However, as the dataset grows, the mixture of experts method allows efficient use of all the data.

2.2. Ensemble method

Ensemble methods for machine learning employ the idea that multiple neural networks can be combined to give a result that is better than any single one. We have multiple predictors that may be the same type of learner or different types of learner. In our experiment we used the same type of multilayer perceptron with a sum rule from all the predictors in the ensemble for the final classification. The dataset is partitioned into disjoint subsets to create variance among the predictors and increase training efficiency. When the ensemble neural network receives input, it runs every predictor on data. The output of all the neural networks in the ensemble are combined and the sum rule is used to make the final classification of the ensemble network [4].

3. Problem statement

Will one of the mixture of experts, ensemble or single model method for neural network machine learning far surpass the others in accurately classifying a large dataset while having similar training times?

4. Methodology

Each neural network across all experiments are composed of 784 neurons for the input layer, 625 for the first hidden layer, 300 for the second hidden layer and 10 for the output. Each hidden layer uses the sigmoid activation function. Softmax is used as the cost function along with back-propagation with a learning rate of 0.5. Training consisted of mini-batching with size 128. If the data for a model does not divide evenly by 128, a smaller batch with the remaining data is added to the input to ensure each model trained on all its data. While keeping the above parameters constant, both mixture of experts and the ensemble method runs with the number of models in the neural network increasing from 2 to 10. One of the goals is to see if there

are different optimal values for the number of models trained with each method when classifying fashion MNIST images. The training time and accuracy for each method are recorded, with 1 epoch to train and 3 epochs to train. Time to train is calculated by measuring the call to the full back-propagation algorithm, summed up for all batches across all epochs and k-folds for all models for a single method. A k-fold cross-validation with k=10 is used for all experimentation so that a confidence interval can be calculated using the results from each fold. The training time of the preprocessed k-means clustering for the mixture of experts is recorded separately, due to it only needing to be computed once. On top of this, the time to train and accuracy for a single multi layer perceptron on all the training data is also recorded.

5. Results and analysis

The performance of the networks was initially tested for only 1 epoch. After seeing a single MLP performing better than all other methods, the performance for after 3 epochs was tested to see if some of the methods converged more slowly than the single MLP (Figure 2).

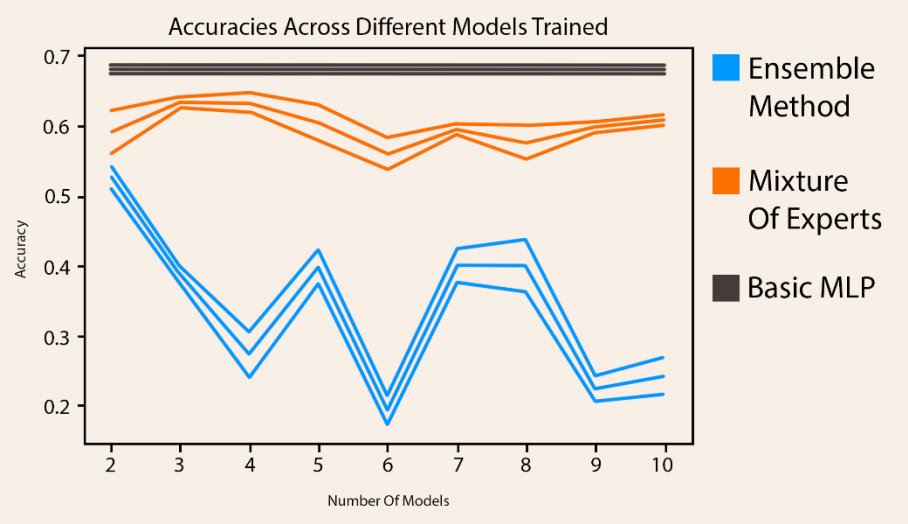

Figure 2. Confidence interval showing the classification accuracy of each method after 1 epoch using k-fold cross-validation

The single MLP continued to improve while the other 2 techniques did not improve enough to show that training past 3 epochs would yield one technique outperforming a single MLP (Figure 3). The fashion MNIST dataset was most likely not large enough to take advantage of combining variance amongst models or having specialized models. The resulting disjoint datasets with only 2 models was enough to degrade the performance of both the mixture of experts and ensemble networks. The size of the dataset was not

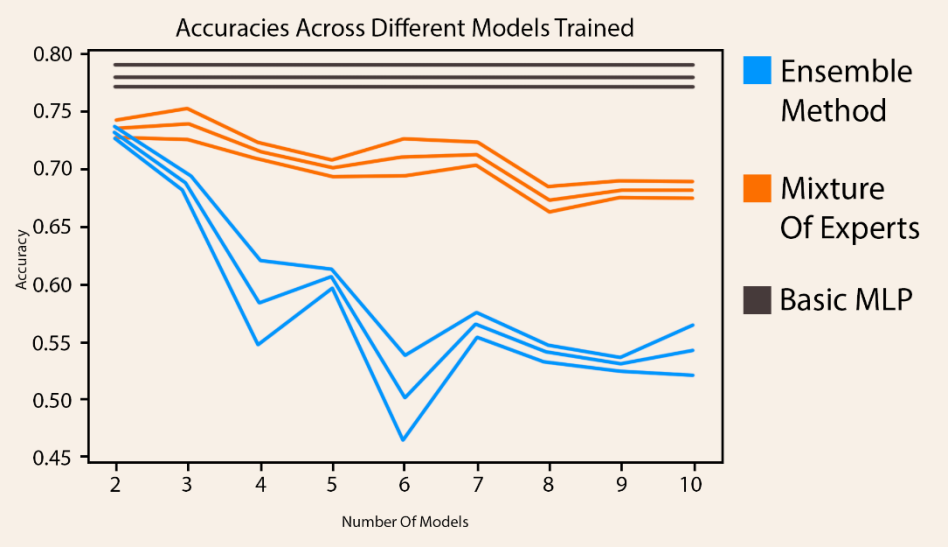

Figure 3. Confidence interval showing the classification accuracy of each method after 3 epochs using k-fold cross-validation

large enough to overcome the obstacle of not having as much training data to fine tune the parameters of each neural network model. Mixture of experts with three models trained is shown to be similar in accuracy to the basic MLP. The time to train increases largely as more models are trained for both the mixture of experts and ensemble method, which is not what was initially expected. Since k-means only needs to be computed once for a dataset, it was included separately in the training times (Figure 4).

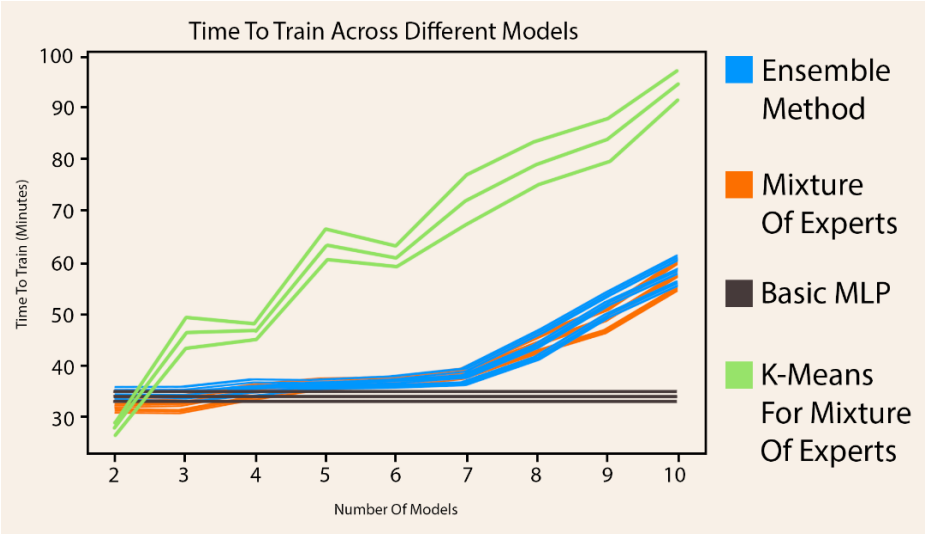

Figure 4. Confidence interval showing the time to train for each method after 3 epochs using k-fold cross-validation. The time to compute k-means clustering is included in this graph only.

5.1. Challenges and solutions

The time to train results were not expected. The goal was initially to have the training times be similar across all the methods described (Figure 5). No

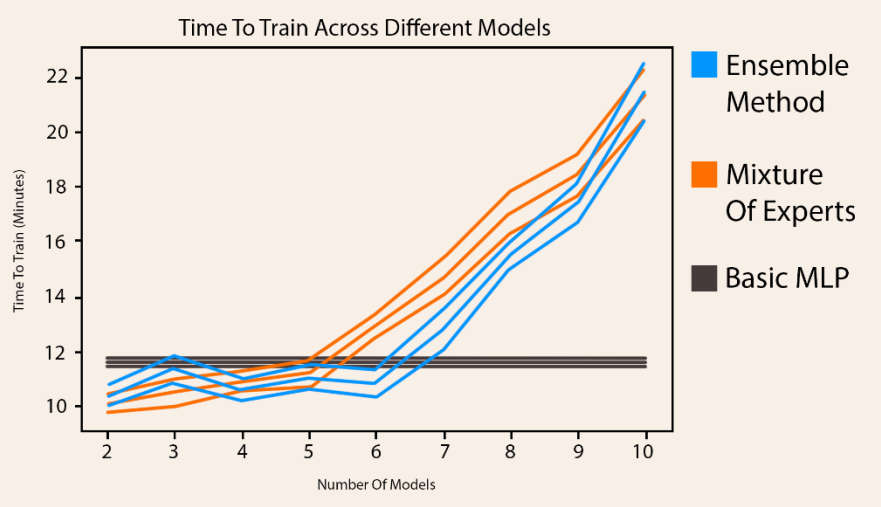

Figure 5. Confidence interval showing the time to train of each method after 1 epoch using k-fold cross-validation

matter the number of models used to train, the same amount of data was used as input in each epoch for the mixture of experts, ensemble and standard method. The data partitioned across the models was disjoint which together equaled the entire dataset. Tensorflow was used for the design of all the models. It is speculated that underlying optimization which Tensorflow performs is being hindered when multiple models are created and deleted frequently, since we only had one model in main memory at a time while performing experiments. Another reason for why the time to train was increased was due to more models partitioning data. There were more partial batches less than 128 in size used for a single epoch, thus increasing the total calls to the back-propagation algorithm. This alone was not a significant enough factor to account for the increase in training time. More investigation is required. Another challenge in testing was partitioning data properly. A library that facilitated training multiple models would be very beneficial for further testing and applications in the tech industry.

6. Conclusion

The mixture of experts and ensemble method machine learning techniques were tested on the fashion MNIST dataset for both 1 and 3 epochs, recording training time and accuracy. The results show that with both the methods described, a single multi layer perceptron trained clearly surpassed in

performing accurate classifications of the fashion MNIST dataset. Further investigation on the time required to train is required. More experimentation is required to show whether the simple mixture of experts and simple ensemble method is preferable in a general case with very large datasets. A larger dataset is recommended for further experimentation. Simulating a situation where large amounts of training data is rapidly available to the model after it has already been trained may show that training an additional model when new data is presented could be far superior in terms of training time and accuracy as a result. For further experimentation, a combination of the mixture of experts and the ensemble method could prove to be more effective. One idea is to have multiple networks trained via the ensemble method but to multiply one network's predictions by some scalar if it is assessed as having a dataset closer to the trained data computed with some gated function.

7. References

[1]"Classifying clothes using Tensorflow (Fashion MNIST)", *Medium*, 2018. [Online]. Available: https://medium.com/tensorist/classifying-fashion-articles-using-tensorflow-fashion-mnist-f22e8a04728a. [Accessed: 16- Dec- 2018].

[2]G. Hinton, *Mixtures of Experts [Neural Networks for Machine Learning]*. 2016.

[3]M. Hauskrecht, *Ensamble methods. Mixtures of experts*. university of pittsburgh, 2004.

[4]R. Polikar, "Ensemble learning", *scholarpedia*, 2009. [Online]. Available: http://www.scholarpedia.org/article/Ensemble_learning. [Accessed: 16- Dec- 2018].