Christopher Rico
MSDS 422
Assignment 6
11/2/19

Introduction

Within the realm of data science, there is a constant quest to optimize the tradeoff between model accuracy and speed. In a situation with boundless memory and compute speeds, accuracy would always win. However, that is almost never the case, so at times it becomes necessary to take a closer look at what a compromise between accuracy and training speed really means.

Computer vision, one of the most intensive topics in data science, tends to be notoriously resource-heavy, requiring vast amounts of input data, memory, and processor time, so image classification model optimization is of great interest. Deep neural networks are cutting-edge machine learning models that show great promise with complex prediction tasks, such as computer vision. However, they can be computationally intense, and often require rather intricate architecture tuning to strike a balance between accuracy and speed. In furtherance of optimizing this trade-off, several DNN modeling topologies were attempted on the MNIST dataset. MNIST, considered to be the “hello world” of image classification, is a well-recognized series of several thousand handwritten digits between 0-9. A DNN modeling technique was developed that is performant in classification results and training time.

Research Design and Statistical Methods

The MNIST dataset is a well-known dataset that consists of 70,000 28x28px images of hand-written digits between 0-9. Each record contains 785 variables: 784 pixel values and 1 label. The machine learning competition Kaggle has broken up these 70,000 images into a 42,000 image training set and a 28,000 image test set, which were the training and test datasets used to assess the DNN approaches. DNN topology -- the number of internal layers and nodes in each layer -- can have a powerful effect on the accuracy

and speed of training and prediction. To ascertain which topology yielded the best results, 4 DNNs, each with a different combination of internal layer number and nodes per layer, were compared to one another.

Programming Work

Model construction and evaluation was completed entirely in the cloud, using Google Colaboratory as a development environment and TensorFlow as a DNN modeling environment. TensorFlow is an industry standard neural net software built by Google.

To begin, four different DNN models were instantiaed. The topology combinations tested were combinations of 2 or 5 internal layers, and 10 or 20 nodes per layer. Each DNN was timed while being trained on the 42,000 image training set. Then, training set accuracy was reported and predictions on the 28,000 image test set were output to CSV files for submission and evaluation on Kaggle.

Results and Recommendation

Both training and test set accuracy results are fairly consistent across DNN topologies. Training time increases modestly as the complexity of the model increases, which is to be expected. The DNN with 5 internal layers and 20 nodes per layer yields a test set accuracy of 0.9506 with a training time of 102 seconds, which is outstanding.

Due to the high performance and acceptable training time, it is recommended to use a DNN with 5 internal layers and 20 nodes per layer for optical character recognition applications.

Number of Layers	Nodes per Layer	Processing Time	Training Set Accuracy	Test Set Accuracy
2	10	79.97	0.921809	0.9373
2	20	88.16	0.984833	0.9484
5	10	90.47	0.815167	0.9077
5	20	102.65	0.972143	0.9506

A shared version of the interactive Jupyter notebook used to run this analysis can be found at:
https://colab.research.google.com/drive/1k7xhjEH5Lp-W9uSe0W0H1JKkjXnkl73U

Kaggle profile with test scores can be found at:
https://www.kaggle.com/christophrico