WebNN Server CPU

Undergraduate Students: Ronald Cotton < *ronald.cotton@wsu.edu* > - Aaron Goin < *aaron.goin@wsu.edu* >
Faculty: Xinghui Zhao, Ph.D. < *x.zhao@wsu.edu* > ***DSR Lab*** - Shawn Welter < *shawn.welter@wsu.edu* > ***IT Systems Specialist***

Background and Solutions

Neural Network

Develop a Deep Learning framework that provides both Dense Neural Networks (DNN) and Convolutional Neural Networks (CNN.)

Distributed

Distributed Tensorflow is comprised of workers, a special worker named the chief, and parameter server(s) that hold weights, biases, and other tensorflow variables. The entire dataset is processed in small batches in an synchronous fashion via the chief (worker 0). As gradients return to the chief, the server receives stale gradients from seven steps back. *tf.train.SyncReplicasOptimizer* prevents this adverse effect by averaging all gradients and applies this to all affected variables in the parameter server(s) before the other workers continue.

Environment

Docker containers deploy a predictable and repeatable environment on local and cloud environments. Docker packs an application's required libraries and dependencies into a script which can be built to execute large software environments. This reduces the build time and the host can externally control the container.

Compatible

Docker allows Containers to run on multiple operating systems. Docker also provides methods to isolated the Containers into a few cores and minimal memory. WebNN Javascript runs on many browsers and uses OpenCL.

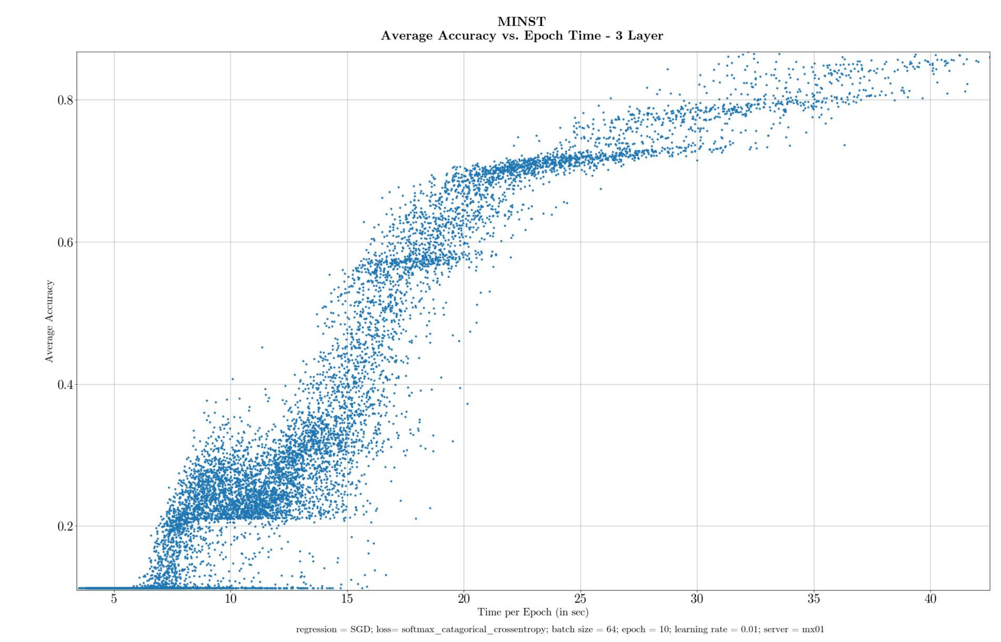
Software Implemented

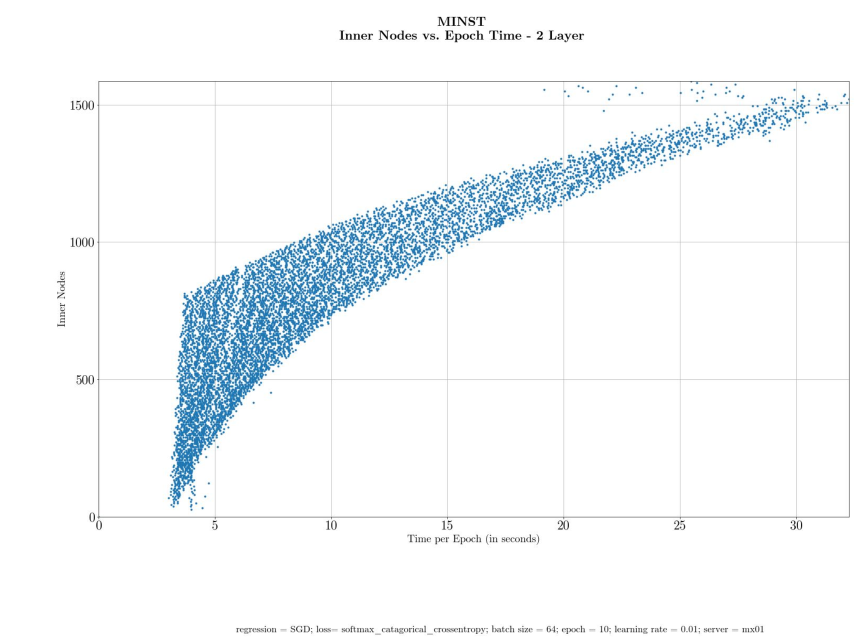
Keras with a Tensorflow Backend

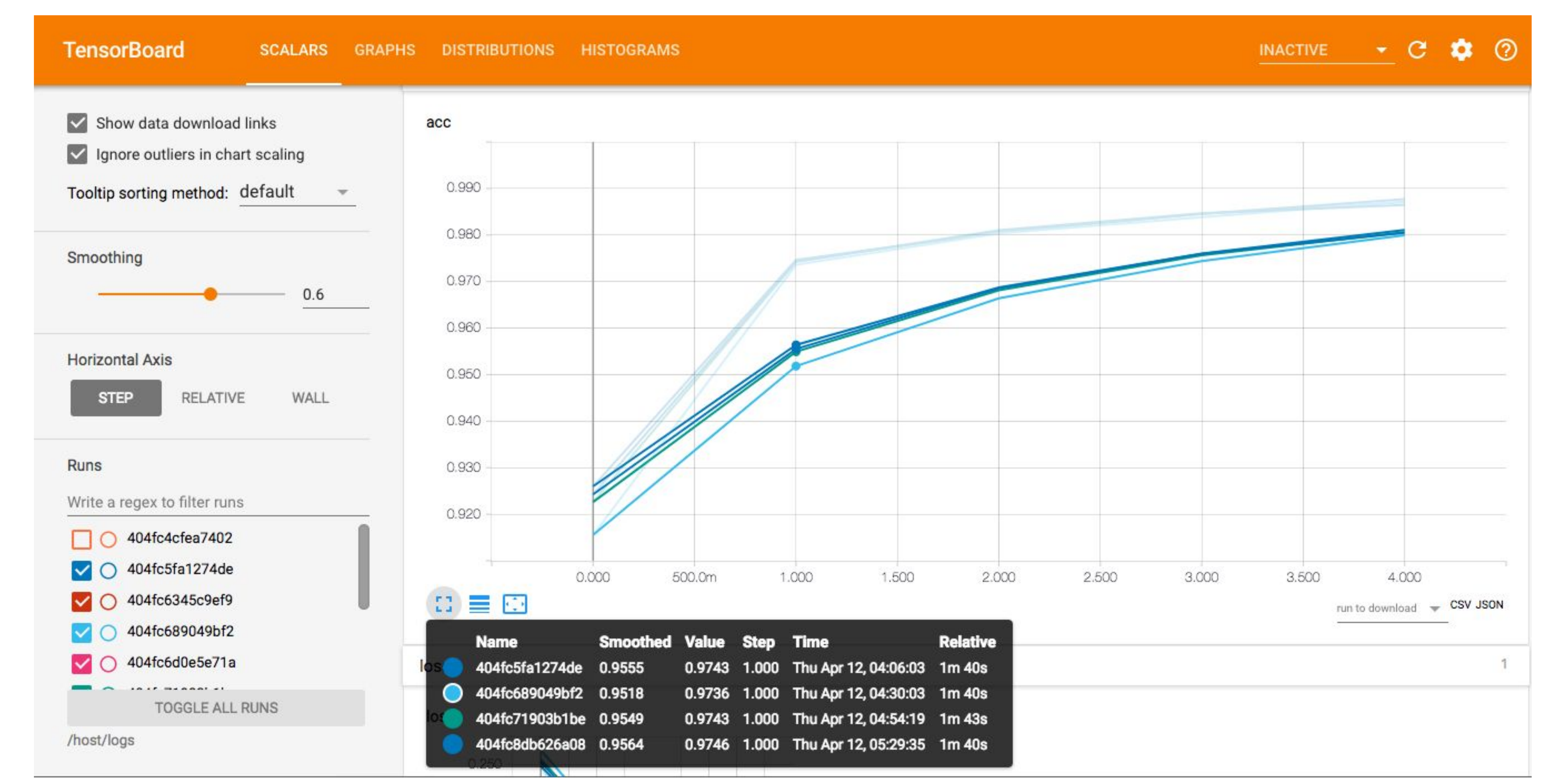
Docker - Ubuntu Docker Container

Results

Hyperparameter Grid Search over a DNN using Tanh activation and a softmax output layer with the MNIST dataset.







Several runs visualizing accuracy on Tensorboard remotely.

Conclusions

After reviewing initial tests conducted by Aaron Goin for WebNN Tensorflow.js, I believe Neural Network processing should only be completed on GPU assets.

ZVC Whitening as well as other Image processing before batching is costly operation on CPU, even when processed in the background.

Further Research

Uber's Horovod utilizes a MPI model requires effort (coding) to distribute neural network with promising results.

WASHINGTON STATE UNIVERSITY

WSU Vancouver: School of Engineering and Computer Science - 14204 NE Salmon Creek Ave, Vancouver, WA 98686