Execution-Time Performance of Deep Learning Networks on CPU, GPU and TPU Runtime Environments

Summary

A performance review of execution times on Google Colab, for five deep learning network examples, was conducted on CPU, GPU and TPU runtime environments using the MNIST dataset. The networks were 1) a multi-layer perceptron (MLP) network, 2) a convolutional neural network (CNN), 3) a recurrent neural network (RNN), 4) a long short-term memory network (LSTM), and 5) an autoencoder.

General findings

Training times (Table 1) for all five network exemplars were significantly better on the GPU runtime environment than on Google Colab's CPU environment. Of the networks, the CNN had the greatest performance improvement on GPUs than CPUs only, with a speedup of over 33 times (3332%). This was followed by the LSTM, which had a speedup of over 22 times (2257%), while speed ups for the autoencoder, MLP and RNN were 1464%, 697% and 229% respectively.

Execution time performance for model testing was also significantly better on GPUs than CPUs, for the exemplars. Speedups for the LSTM, CNN, RNN, autoencoder and MLP where 1113%, 915%, 601%, 326%, and 177% respectively.

The TPU runtime environment performed worse than the CPU environment, on training times for the autoencoder, RNN and CNN. Performance time declines were most significant for the autoencoder (-10%). TPU training times were nevertheless significantly better for the LSTM (+9%), and marginally better for the MLP (+1%), than on CPU runtime. All model exemplars performed worse on model evaluation times, on TPUs than on CPUs.

Discussion

To leverage advantages of using TPUs, optimizations could have been applied to the code used for the performance evaluations. Nevertheless, no customizations were made to the code used, for a head-to-head comparison in the environments. The network code examples were simply run under the three runtime environment options by changing the relevant Colab notebook settings.

Appendix:

Table 1: Summary of CPU, GPU, TPU Performance

Deep Learning Network	Training Execution Time On:			Testing Execution Time On:			Training Time Speedup As Percentage of Performance on CPU			Testing Time Speedup As Percentage of Performance on CPU		
	CPU	GPU	TPU	CPU	GPU	TPU	CPU	GPU	TPU	CPU	GPU	TPU
MLP	71.622	8.983	70.666	0.929	0.335	0.954	0%	697%	1%	0%	177%	-3%
CNN	1259.83	36.711	1271.57	5.41	0.533	5.626	0%	3332%	-1%	0%	915%	-4%
RNN	125.111	38.01	126.034	6.573	0.938	6.952	0%	229%	-1%	0%	601%	-5%
LSTM	398.724	16.914	364.476	18.504	1.525	21.61	0%	2257%	9%	0%	1113%	-14%
Autoencoder	127.795	8.172	142.383	5.327	1.25	5.935	0%	1464%	-10%	0%	326%	-10%

Multi-Layer Perceptron (MLP) Example using MNIST Dataset

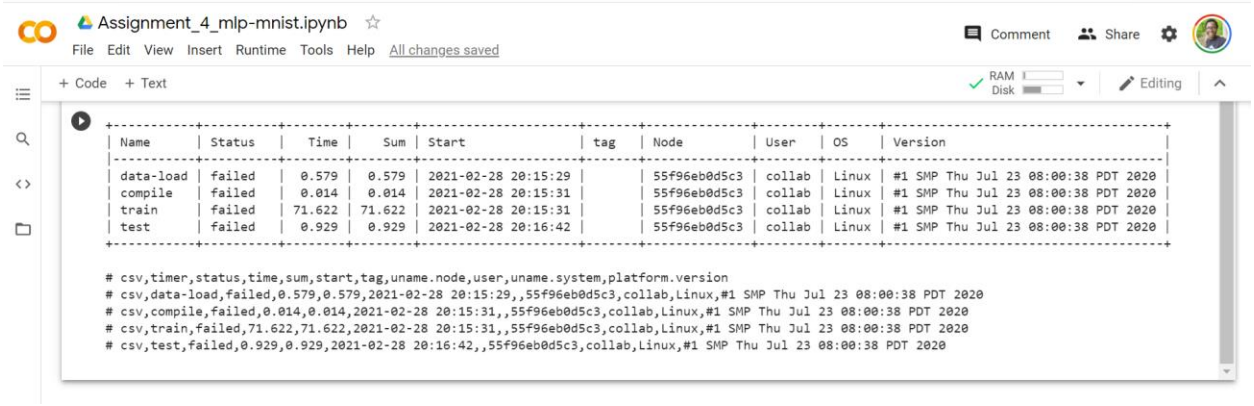

Figure 1: MLP using CPUs only

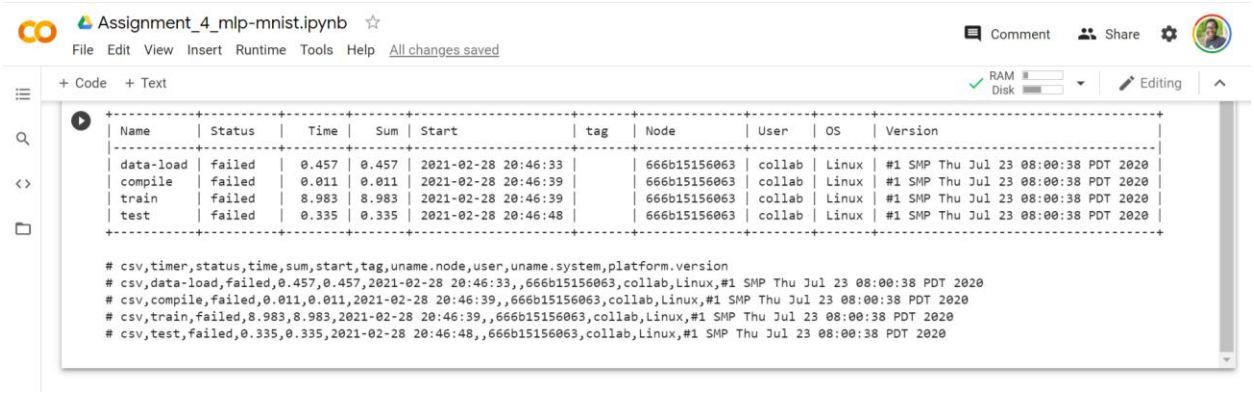

Figure 2: MLP using GPUs

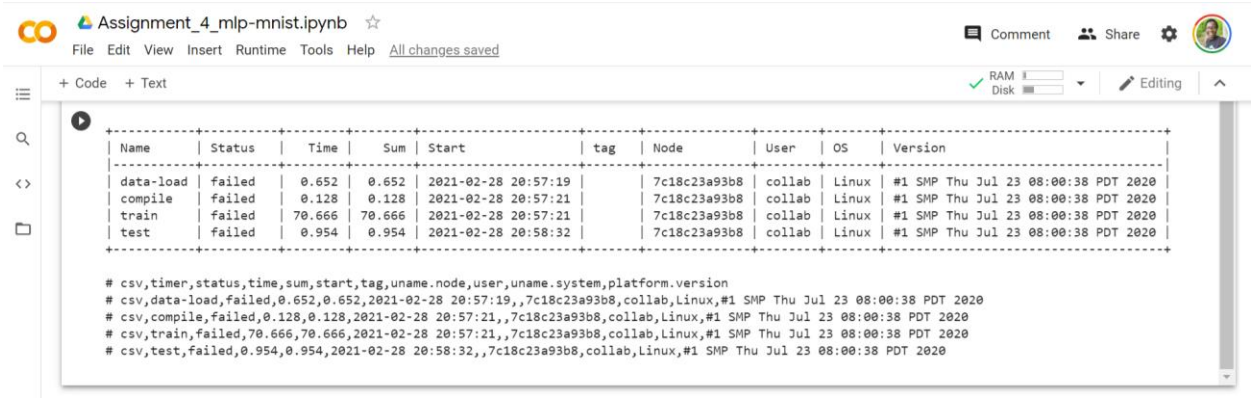

Figure 3: MLP using TPUs

Convolutional Neural Networks (CNN) Example using MNIST Dataset

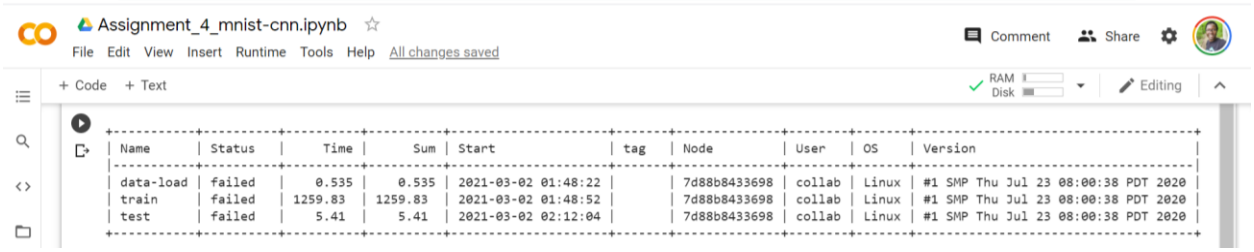

Figure 4: CNN using CPUs only

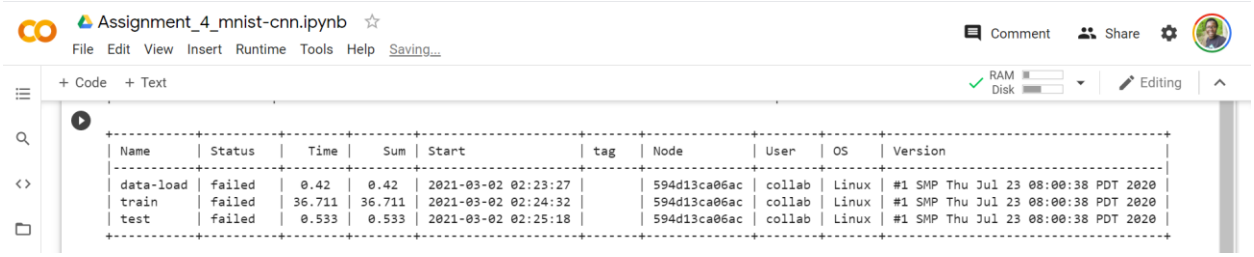

Figure 5: CNN using GPUs

Figure 6: CNN using TPUs

Recurrent Neural Networks (RNN) Example using MNIST Dataset

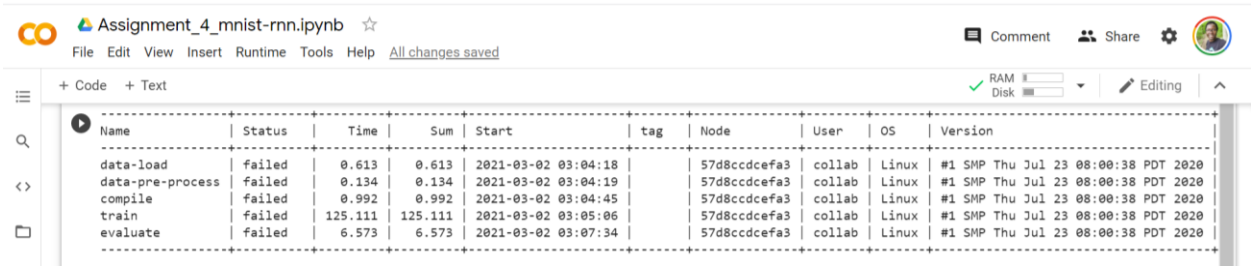

Figure 7: RNN using CPUs only

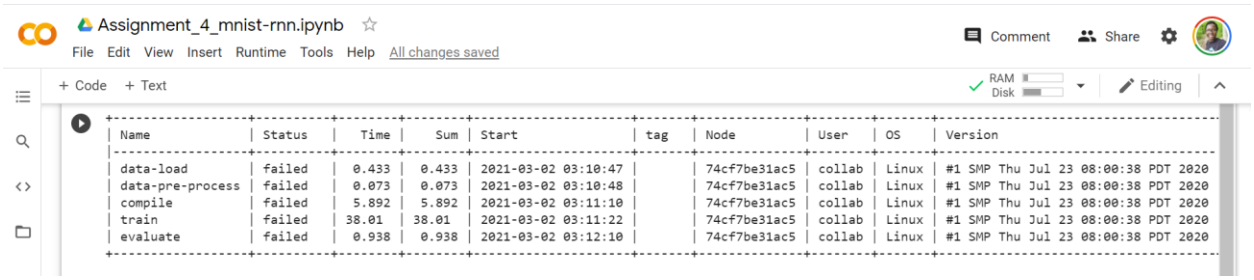

Figure 8: RNN using GPUs

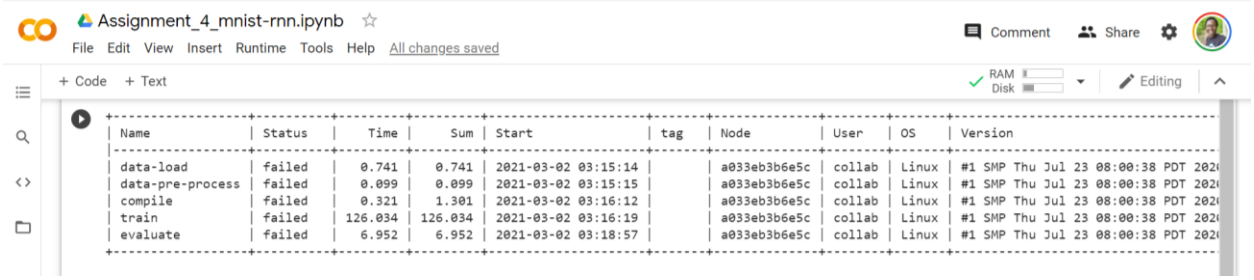

Figure 9: RNN using TPUs

Long Short-Term Memory (LSTM) Example using MNIST Dataset

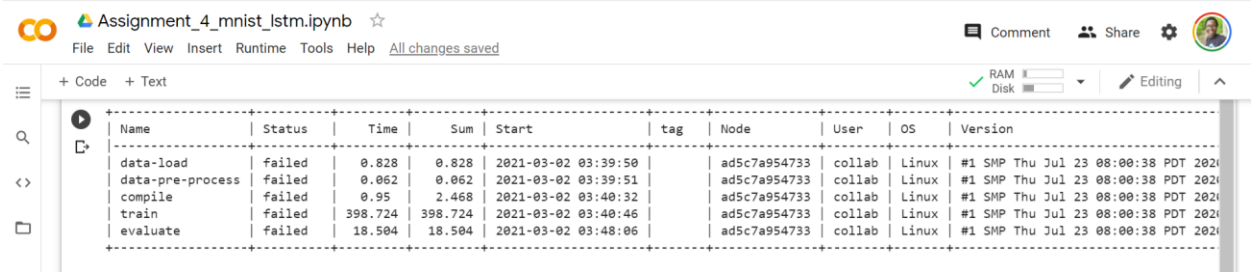

Figure 10: LSTM using CPUs only

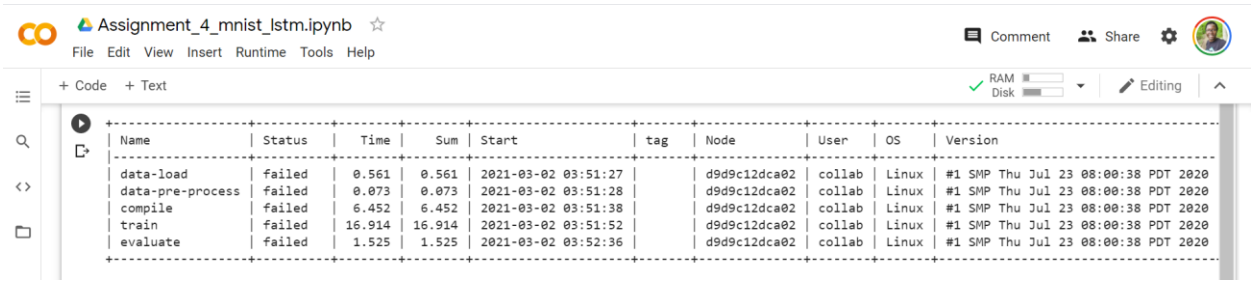

Figure 11: LSTM using GPUs

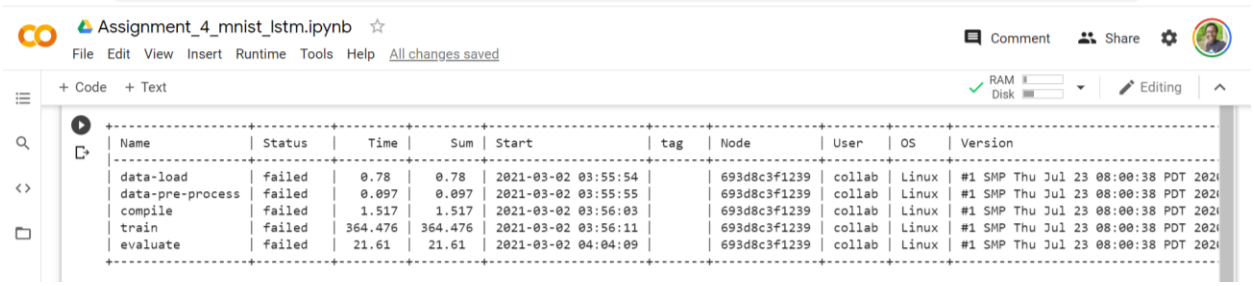

Figure 12: LSTM using TPUs

Autoencoder Example using MNIST Dataset

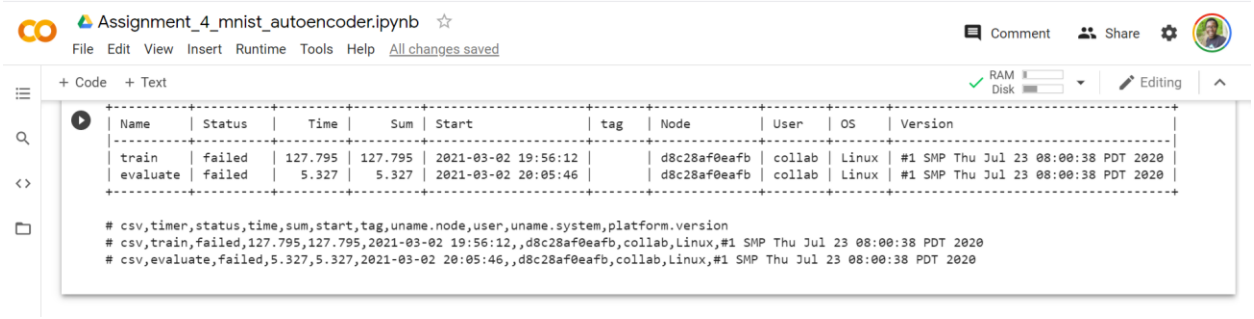

Figure 13: Autoencoder using CPUs only

Figure 14: Autoencoder using GPUs

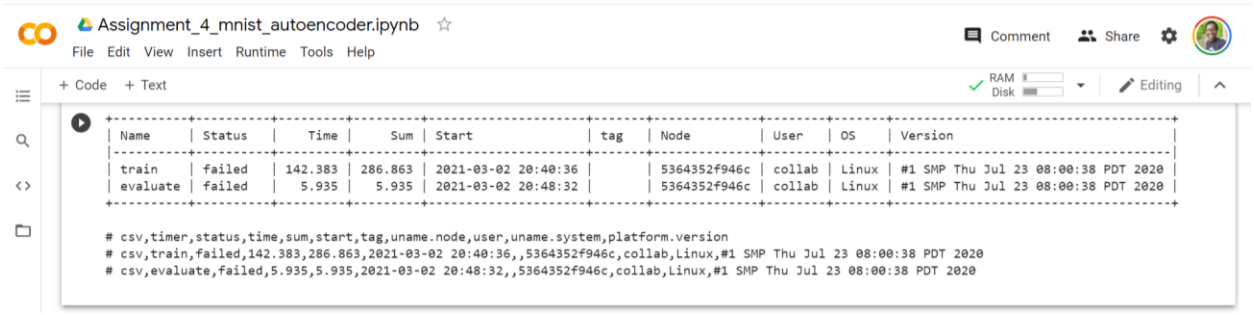

Figure 15: Autoencoder using TPUs