CS3450  
Kevin Paganini  
3/7/2023  
Lab 1: Predicting Runtimes  
  
  
  
**Predict Relative Runtime from GPU Specs**

The NVIDIA V100 has a nominal double-precision throughput of 16.4 single-precision (float) TFLOPS. (floating point operations per second). The T4 has a throughput of 8.1 single-precision. In your report, record what ratio you expect between the times these two systems will take to train a neural network. Briefly explain your reasoning. Assume that only one GPU is used on each node. I don’t know of an authoritative source defining what counts as a FLOP, but this Stack Overflow answer has the ring of truth for me (Dr. Yoder).  
  
FLOPs amount of time it takes to do one addition and multiplication.   
  
**Answer:** Since the V100 has double the throughput as the T4, I would expect it to be able to train roughly doubly as fast.

**Experimentally Determine the Runtime from Running One Epoch****T4 Node two CPUs:   
Job 107186: 24.4 minutes**Text

Description automatically generated

**T4 Node 16 CPUs:  
Job 107482: 512 seconds**Text

Description automatically generated

**T4 Node with GPU and eight CPUs  
Job 107456: 86 seconds**Text

Description automatically generated **Running on a DGX Node with a GPU (and eight CPUs)  
Job 107474: 36.8 seconds**Text

Description automatically generated

**Predict the Runtime for 20 epochs**T4 Node two CPUs: 8.1 hours  
  
T4 Node 16 CPUs: 2.85 hours  
  
T4 Node with GPU (and eight CPUs): 28.67 minutes  
  
Running on a DGX Node with a GPU (and eight CPUs): 12.3 minutes **Experimentally Determine the Runtime from Running 20 Epochs**

**T4 Node two CPUs: 8.8 hours  
Job 107507:**Text

Description automatically generated **T4 Node 16 CPUs: 3.25 hours  
Job 107643:**Text

Description automatically generated

**T4 Node with GPU (and eight CPUs): 25.75 minutes  
Job 107522:**Text

Description automatically generated **Running on a DGX Node with a GPU (and eight CPUs): 10.8 minutes  
Job 107524:**Text

Description automatically generated

**Full table of runtimes:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epochs | 2 CPUs | 16 CPUs | T4 + 8 CPUs | V100 + 8 CPUs |
| 1 | 24.4 minutes | 8.5 minutes | 86 seconds | 36.8 seconds |
| 20 (predicted) | 8.1 hours | 2.85 hours | 28.67 minutes | 12.3 minutes |
| 20 (actual) | 8.8 hours (26.4 minutes per epoch) | 3.25 hours (9.75 minutes per epoch) | 25.75 minutes (1.3 minutes per epoch) | 10.8 minutes (32.65 seconds per epoch) |

**Discussion**Include a concluding discussion in your report. How closely did your predictions match the reality? Explain any differences. Also, be sure to discuss how the relative speed between the DGX and the T4 compared with your predictions.  
  
I predicted that the V100 would roughly be double as fast as the T4. The actual ratio experimentally found is T4 / V100 = 86 seconds / 36.8 seconds = 2.34, (25.75 / 10.8 = 2.38). So, the experimentally found relative training time is actually slower (in terms of the T4) than what was predicted. This is most likely due to other processes running on the teaching node whilst I was running my experiment. If all things were held equal for both the T4 and V100, I believe that the ratio of training times would be closer to 2.   
  
Both the training sessions that only used CPU’s took longer than I anticipated. Both training sessions using GPU’s took shorter than expected. I think part of the reason why the CPU jobs took longer than expected is because both of them ran for longer, so they probably started running into traffic from other users. The GPU jobs ran for a lot shorter so they had a lot less chance to run into network traffic. Even though the estimates weren’t exact they would all be very good to use when trying to provide a ball park estimate for a training session, with a certain number of epochs.