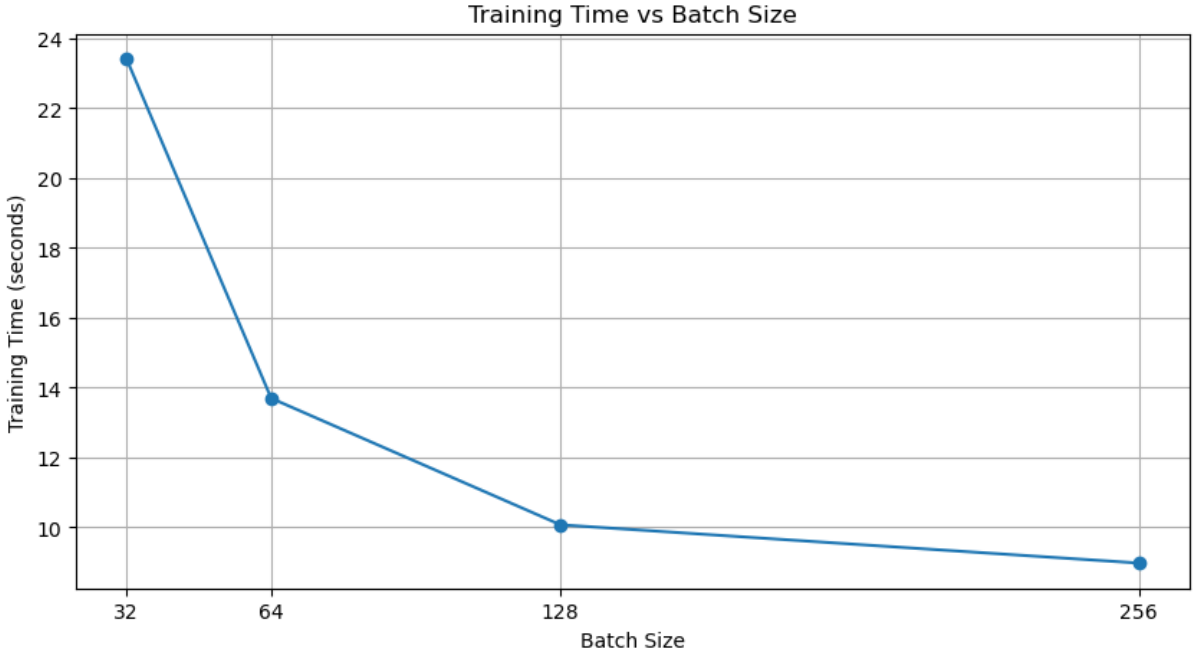
CDS 451: Designing High Performant Systems for AI
Fall 2023

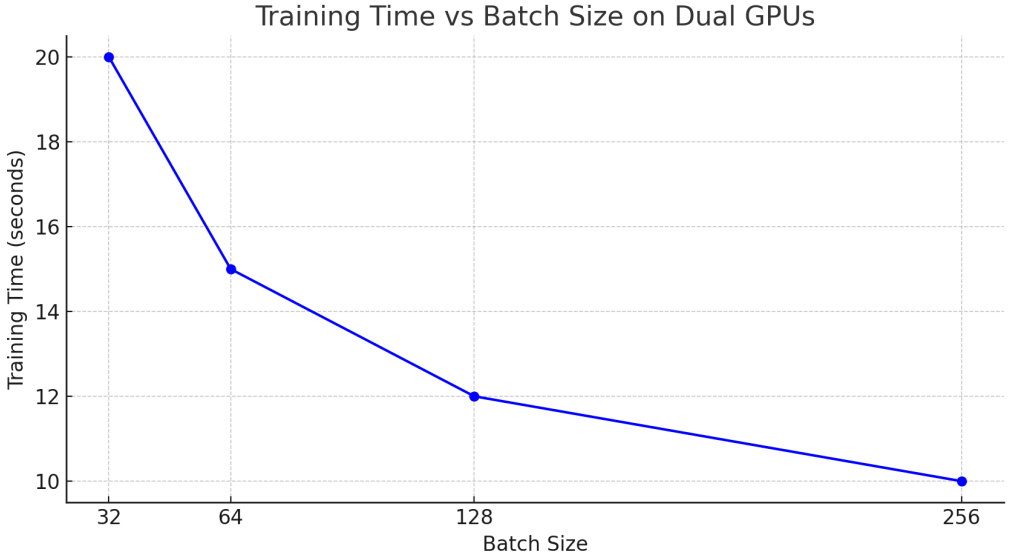
Programming Assignment 3

Ruilin Jin (rxj420@case.edu)

Problem a Please see the python code for implementation.
Problem b Here I used 32, 64, 128, and 256 as my batch sizes.
Problem c Here's the plot:



Problem d Here's the plot:



Problem e GPUs I used:

```
(base) ulab@ulab-2rtx:~$ nvidia-smi
Tue Nov 21 14:35:35 2023
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.113.01             Driver Version: 535.113.01   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA TITAN RTX               Off | 00000000:01:00.0  On |                  N/A |
| 41%   28C    P8              2W / 280W |    239MiB / 24576MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA TITAN RTX               Off | 00000000:02:00.0 Off |                  N/A |
| 40%   26C    P8             15W / 280W |    174MiB / 24576MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    0   N/A  N/A       930      G   /usr/lib/xorg/Xorg                          137MiB |
|    0   N/A  N/A      1360      G   /usr/bin/gnome-shell                         26MiB |
|    0   N/A  N/A     84988      G   /opt/teamviewer/tv_bin/TeamViewer             4MiB |
|    0   N/A  N/A     85131      G   ...0980027,16165953840484018868,262144       36MiB |
|    0   N/A  N/A     86615      G   ...sion,SpareRendererForSitePerProcess       26MiB |
|    1   N/A  N/A       930      G   /usr/lib/xorg/Xorg                            4MiB |
```

Problem f

Batch size directly affects training speed. With a larger batch size, more data is processed at once, leading to faster epochs because the GPU can parallelize the workload efficiently. However, this is up to a point where the GPU memory is maximized without being overloaded. Beyond that point, too large a batch size can slow down training due to memory swapping or may not fit into GPU memory at all.

The trend between training on 1 GPU versus 2 GPUs generally shows that training with 2 GPUs is faster due to parallel processing. However, the speedup is not always linear due to overhead from communication and synchronization between GPUs.

Problem g

With 2 GPUs, the training speed observed is faster compared to using just 1 GPU. The reasons include increased parallelism, as both GPUs can process separate batches of data simultaneously, effectively doubling the computational throughput, assuming there's efficient scaling and the neural network model is large enough to benefit from the extra GPU. However, for very small models or batch sizes, the overhead of managing multiple GPUs can outweigh the speed benefits.

Reference

1. Keskar, Nitish Shirish, et al. "On large-batch training for deep learning: Generalization gap and sharp minima." arXiv preprint arXiv:1609.04836 (2016).
2. Shen, Xiaoqi, et al. "Effects of patchwise sampling strategy to three-dimensional convolutional neural network-based Alzheimer's disease classification." Brain Sciences 13.2 (2023): 254.
3. https://medium.com/mini-distill/effect-of-batch-size-on-training-dynamics-21c14f7a716e
4. https://www.sciencedirect.com/science/article/pii/S2405959519303455