American Sign Language Recognition using CNNs and Data Parallelism in PyTorch

Nikunj Rajesh Lad (001422467)
{lad.n}@husky.neu.edu

1. Introduction

1.1 Background

We are privileged as humans to talk and converse and communicate with each other. But what about those people who can't talk or speak? Such people find it hard to convey their ideas. In view of that, what if a system exists which can help in identifying hand signs of people and can categorize them from live stream of video like a webcam or a camera. This project intends to make use of Convolutional Neural Networks to identify and classify hand signs into corresponding classes – like say letter A, B, etc. We also make use of 3 more categories of signs, which are custom made – space, delete and nothing. Therefore, in total we have 29 categories or signs for classification.

Traditionally sign detection was done using SURF and SIFT features using single core processors and vanilla image processing [1] techniques. They were robust back in the time about 4-5 years ago around 2012-2016 era when image classification was just evolving. The use of GPUs and Convolutional Neural Networks [2] based algorithms were evolving gradually during this period and that spurred a lot of attention of researchers and deep learning practitioners to implement traditionally executed image processing algorithms in a faster manner. Running models on GPU's eventually became the norm when the devices manufactured by Nvidia helped in fast parallel processing of tensors (smallest element defined to hold a data, a container). When Alex Kirzhevsky [3] first implemented 'AlexNet', an architecture which was executed on 2 GTX 580 Nvidia GPU's each with 3GB memory, it opened gates to infinite possibilities in the field of deep learning. More recently, Model Parallelism [5] and Data Parallelism [4] techniques are introduced which helped in speeding up the training process of large deep learning models and reducing computational time.

In this work, we intend to explore Data Parallelism techniques and use Convolutional Neural Networks for recognizing hand signs and classifying them after training them on GPUs.

1.2 Objectives

A major objective of this project is to solve a real-world communication problem between humans who can speak and those who cannot speak. There are about 0.7 to 0.8 million people in the world who are mute/deaf and how can we make their lives easier. How can we reduce human dependency for every person who is deaf/mute? Computer Vision can help here. By training images on a huge set of images of hand signs we can develop a system to intelligently help us recognize and classify hand signs in various ambient conditions using state-of-the art Convolutional Neural Network architectures.

Another major objective or rather the central objective of this project is to exploit the power of deep learning and high-performance computing platforms to train a massive dataset of 87029 images of hand signs and develop a model to help identify such images. How does GPU's help in speeding up computational tasks? How does multiple GPUs and Data Parallel strategies help us save runtime? How does batch size have an effect on total memory consumption? What is the relation of number of GPUs and batch size of training? These questions are very important and needs to be answered. While it is imperative to develop state-of-the art software and algorithms, but none of those techniques work efficiently if the underlying hardware it is implemented on is not utilized intelligently. Understanding the nature and behavior of these high-performance computing platforms is essential for successful development of an optimal solution to such large-scale real-world problems.

2. Hardware Specifications

Cluster: Discovery High-Performance Computing Cluster
Reservation: CSYE7374_GPU
Reservation memory: 95 GB
Time reserved: 10 hrs.

GPU
Model: Nvidia V100-SXM2
GPU count: 0 – 4
GPU Memory: 32.4805 GB / GPU
Clock: 1.290Ghz with max boost of 1.530GHz

CPU
Model: Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz
Sockets: 2
Cores/socket: 14
Threads/core: 1
CPUs: 28

Command for entering reservation:

```
srun -p reservation --reservation=CSYE7374_GPU --gres=gpu:v100-sxm2:4 --pty --mem=95000 --time=10:00:00 /bin/bash
```

Command for getting GPU information

```
nvidia-smi -q -d CLOCK & watch -n 0.1 nvidia-smi
```

Command for getting CPU information

```
lscpu
```

3. Dataset Specifications

Dataset size: 87029 images. 26 alphabets and 3 special characters
Each image is 200x200x3 pixels in resolution. The image in resized to 150x150 resolution for faster processing and since information is local to central region and has less complexity.

Training data size: 76760 images
Validation data size: 8528 images
Testing data size: 1741 images

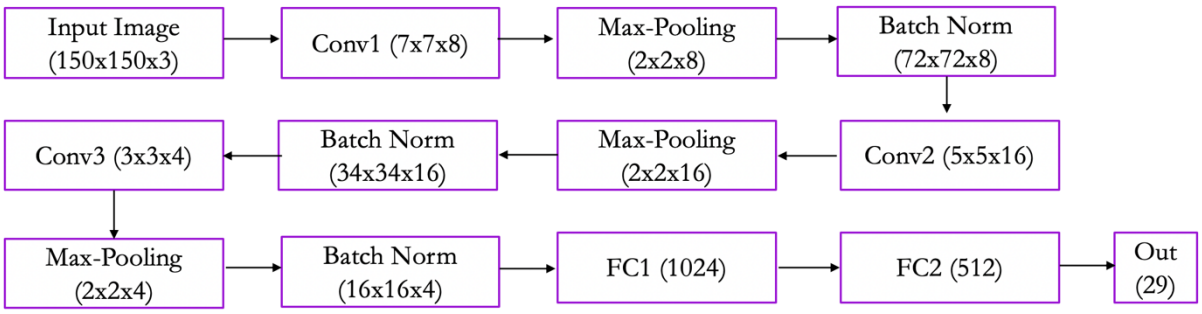
4. Methodology

4.1 Architecture

The below architecture was derived from PyTorch. It gives detailed breakdown of individual filters and image size as it traverses along the network.

```
Net (
  (conv1): Conv2d (3, 8, kernel_size= (7, 7), stride= (1, 1))
  (conv2): Conv2d (8, 16, kernel_size= (5, 5), stride= (1, 1))
  (conv3): Conv2d (16, 4, kernel_size= (3, 3), stride= (1, 1))
  (fc1): Linear (in_features=1024, out_features= 512, bias=True)
  (fc2): Linear (in_features=512, out_features= 29, bias=True)
  (norm4): BatchNorm2d (4, eps=1e-05, momentum=0.1, affine=True,
track_running_stats=True)
  (norm8): BatchNorm2d (8, eps=1e-05, momentum=0.1, affine=True,
track_running_stats=True)
  (norm16): BatchNorm2d (16, eps=1e-05, momentum= 0.1, affine=True,
track_running_stats=True)
  (pool): MaxPool2d (kernel_size=(2, 2), stride=(2, 2), padding=0,
dilation=1, ceil_mode= False)
  (dropout): Dropout (p=0.2, inplace=False)
)
```

4.2 Block Diagram



4.3 Calculation

Given an image it is very important to wisely construct an architecture. Since our images are in 200x200x3 shape and we are resizing it to 150x150x3 so as to reduce computation, it is extremely important to architect the network so as to avoid gradient explosion or disappearance.

We use the formula:

Output = ((input – kernel_size + 2 * padding) / strides) + 1

Convolution layer 1 (conv1):
Input dimensions: 150x150 with depth 3
Kernel size: 7x7
# of filters: 8
Padding: 0
Strides = 1

Output after 1st convolution: ((150 – 7 + 2 * 0) / 1) + 1 = **144x144x8**

Max-Pooling:
Input dimensions: 144x144 with depth of 8
Kernel size: 2x2
Strides: 2

Output after Max-Pooling: **72x72x8**

Convolution layer 2 (conv2):
Input dimensions: 72x72 with depth 8
Kernel size: 5x5
# of filters: 16
Padding: 0
Strides = 1

Output after 1st convolution: ((72 – 5 + 2 * 0) / 1) + 1 = **68x68x16**

Max-Pooling:
Input dimensions: 68x68 with depth of 16
Kernel size: 2x2
Strides: 2

Output after Max-Pooling: **34x34x16**

Convolution layer 3 (conv3):
Input dimensions: 34x34 with depth 16
Kernel size: 3x3
# of filters: 4
Padding: 0
Strides = 1

Output after 1st convolution: ((34 – 3 + 2 * 0) / 1) + 1 = **32x32x4**

Max-Pooling:
Input dimensions: 32x32 with depth of 4
Kernel size: 2x2
Strides: 2

Output after Max-Pooling: **16x16x4**

Flatten:
16x16x4 = **1x1024**

Fully Connected layer 1 (fc1):
Input dimensions: 1024
Output dimensions: **512**

Fully Connected layer 2 (fc2):
Input dimensions: 512
Output dimensions: **29 (equivalent to output classes dimensions)**

Total Layers: 5-layer architecture with Batch Normalization after every layer.

Why did we decide on a custom architecture?

1. Firstly because, state-of-the-art models require 227x227 or 224x224 dimensional images as inputs.
2. They are very deep, and our dataset is not so complex so as to have such deep architectures
3. They can lead to overfitting and may cause vanishing gradient problems.

5. Analysis

We aim to exploit the second objective of our project in the analysis section.

1. What is the time consumed by GPUs on 128, 256, 512 batch size of images?
2. What is the memory consumption of the GPUs using the above batch sizes?
3. How do we have a reasonable tradeoff in performance time and batch size by selecting an optimal number of GPU and data batches.

Following are the experiments performed on various batch sizes. For every case, the following are constant:

1. Training dataset consists of 76760 images,
2. Validation dataset consists of 8528 images,
3. Testing dataset consists of 1741 images.
4. Number of epochs: 25 (since changing them will cause run time variations). So fixing the epochs is essential for generalizing our results and observations.
5. Learning rate: 0.005 (this will affect how fast or slow network converges and since early stopping is disabled, hardly will affect our analysis, since we will run for the entire 25 epochs).
6. Optimizer: Stochastic Gradient Descent (based on empirical results)
7. Activation: ReLU (Rectified Linear Units) are used since they lead to faster computation since reduces run time by ~25%.
8. Dropout factor is 20%. This is again kept constant for generalization and as we know the dropout helps in regularization and prevents overfitting our models.

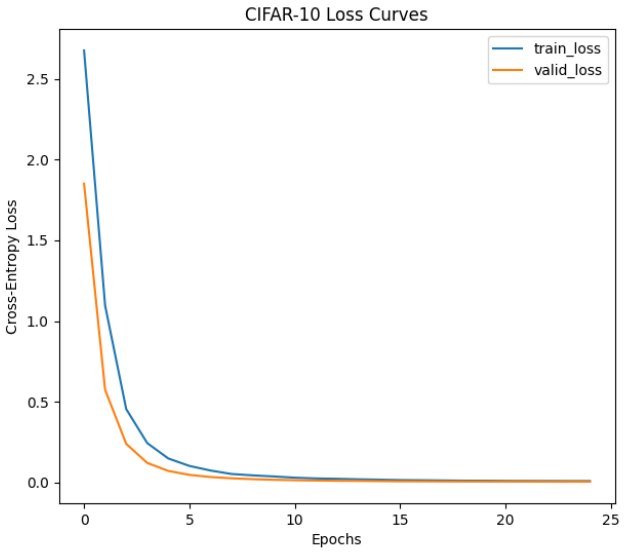
We define the following 4 alternate hypothesis which we intend to prove

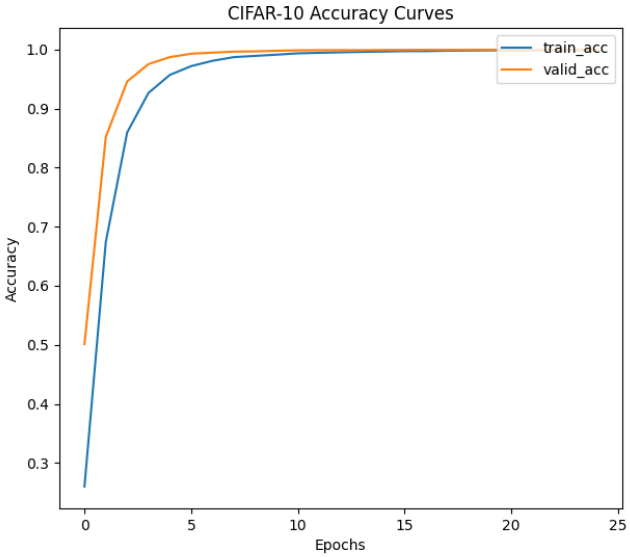
1. **Hypothesis 1**: We can recognize hand signs from images or stream of images using state of the art techniques.
2. **Hypothesis 2:** Using more GPU's always speeds up computation and saves time.
3. **Hypothesis 3:** Larger batch size means better model approximation.
4. **Hypothesis 4:** Keeping number of GPUs constant, there exists a linear relationship between number of batches and time taken. In short, as number of batches increases, time decreases.

5.1 Experiment 1

Batch Size: 128
# of training batches: 600
# of validation batches: 67
 # testing batches: 14

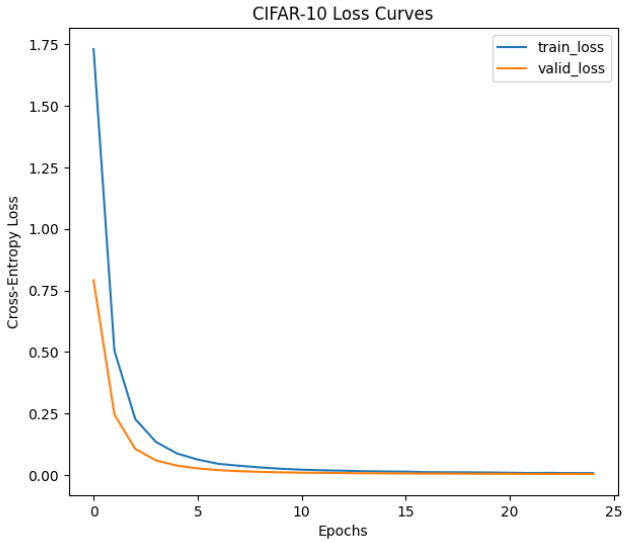
1. **With # of GPUs: 1 and Data Parallelism: No**

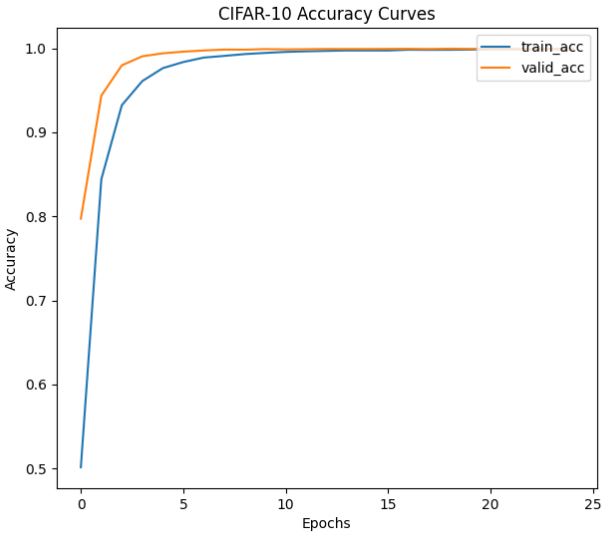





```
Sat Apr 11 23:39:39 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   48C    P0   125W / 300W |   2870MiB / 32480MiB |     27%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
| N/A   37C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
| N/A   37C    P0    44W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   40C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    313996      C   python                                     2859MiB |
+-----------------------------------------------------------------------------+
```

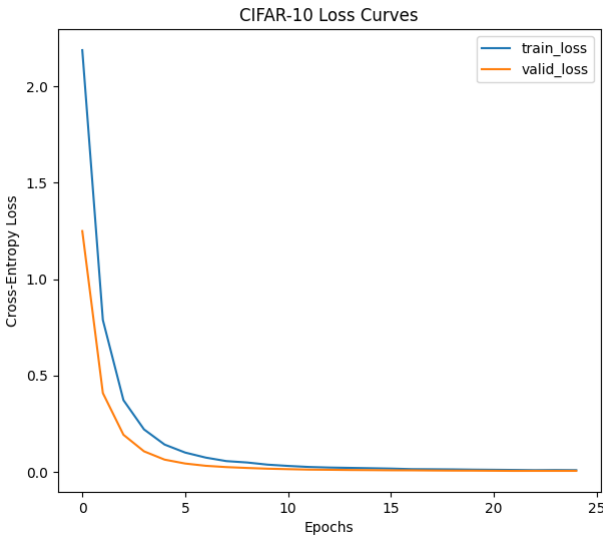
2. With # of GPUs: 2 and Data Parallelism: Yes

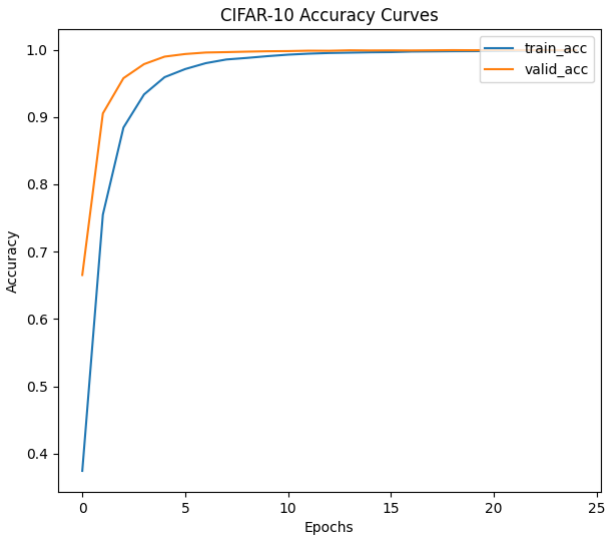





```
Sat Apr 11 22:41:53 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   45C    P0    70W / 300W |   1580MiB / 32480MiB |     23%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
99N/A   41C    P0    60W / 300W |   1542M0B / 32480MiB |     19%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
5 N/A   37C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   40C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    260240      C   python                                     1569MiB |
|    1    260240      C   python                                     1531MiB |
+-----------------------------------------------------------------------------+
```

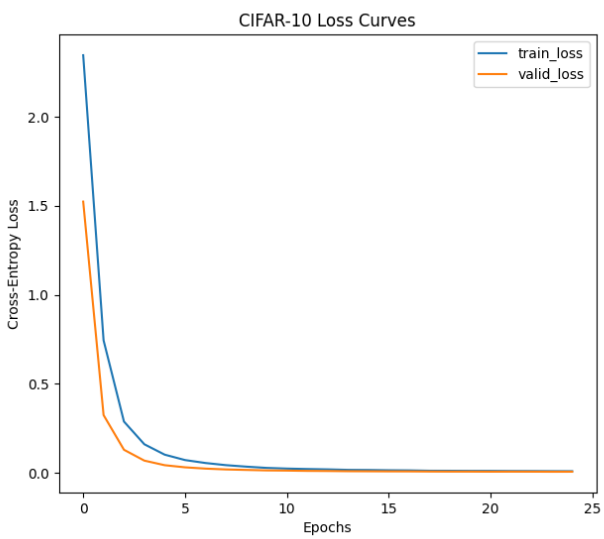
3. With # of GPUs: 3 and Data Parallelism: Yes

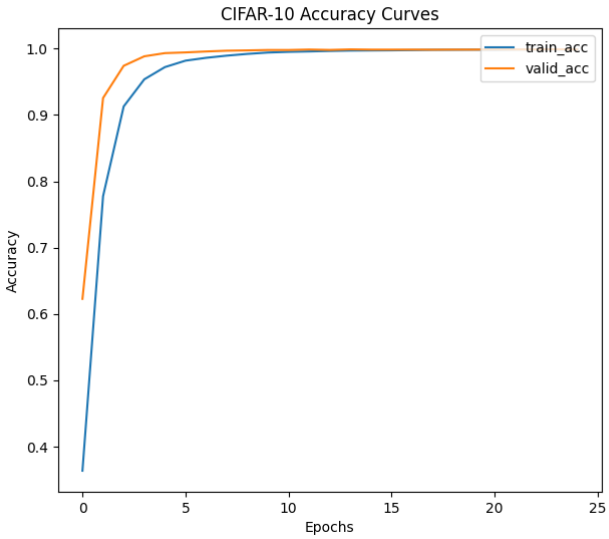





```
Sat Apr 11 20:29:47 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   43C    P0    61W / 300W |   1430MiB / 32480MiB |     18%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
12N/A   40C    P0    59W / 300W |   139 7iB / 32480MiB |     14%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
 7N/A   40C    P0    59W / 300W |   138114B / 32480MiB |     14%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   40C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    101071      C   python                                     1419MiB |
|    1    101071      C   python                                     1381MiB |
|    2    101071      C   python                                     1373MiB |
+-----------------------------------------------------------------------------+
```

4. With # of GPUs: 4 and Data Parallelism: Yes





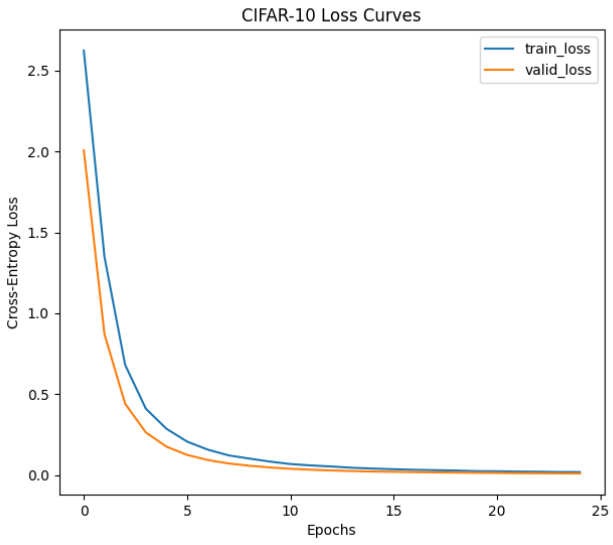

```
Sat Apr 11 21:37:43 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   42C    P0    61W / 300W |   1366MiB / 32480MiB |      9%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
12N/A   39C    P0    59W / 300W |   132 6iB / 32480MiB |      7%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
10N/A   38C    P0    57W / 300W |   1328MiB / 32480MiB |      5%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
10N/A   42C    P0    58W / 300W |   1328MiB / 32480MiB |      5%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    173642      C   python                                     1355MiB |
|    1    173642      C   python                                     1317MiB |
|    2    173642      C   python                                     1317MiB |
|    3    173642      C   python                                     1317MiB |
+-----------------------------------------------------------------------------+
```

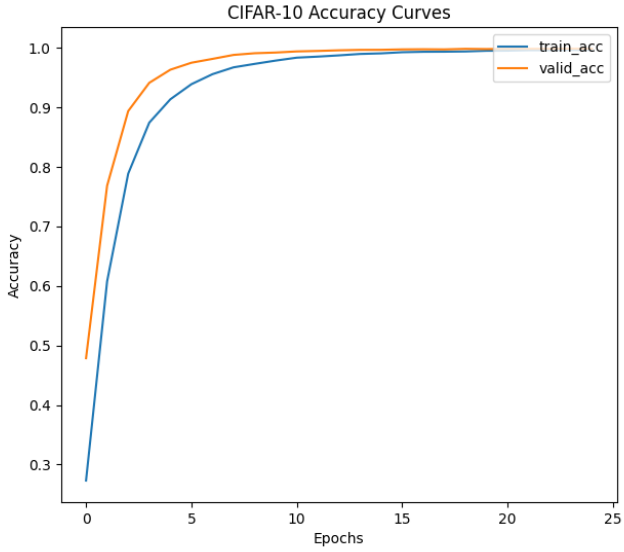
Experiment 1	GPU 0 (GB)	GPU 1 (GB)	GPU 2 (GB)	GPU 3 (GB)	Exec. Time (secs)	Data Parallel	Train Acc	Valid Acc	Test Acc
Using 1GPU	2.859	NA	NA	NA	2500.82	No	99.87	99.91	99.88
Using 2GPU	1.569	1.531	NA	NA	2306.56	Yes	99.89	99.95	100.0
Using 3GPU	1.419	1.381	1.373	NA	2296.49	Yes	99.84	99.90	99.88
Using 4GPU	1.355	1.317	1.317	1.317	2335.14	Yes	99.86	99.88	99.94

5.2 Experiment 2

Batch Size: 256
# of training batches: 300
# of validation batches: 34
 # testing batches: 7

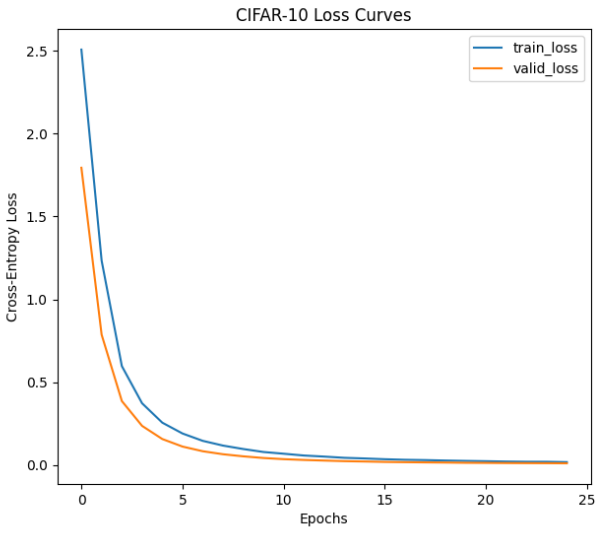
1. **With # of GPUs: 1 and Data Parallelism: No**

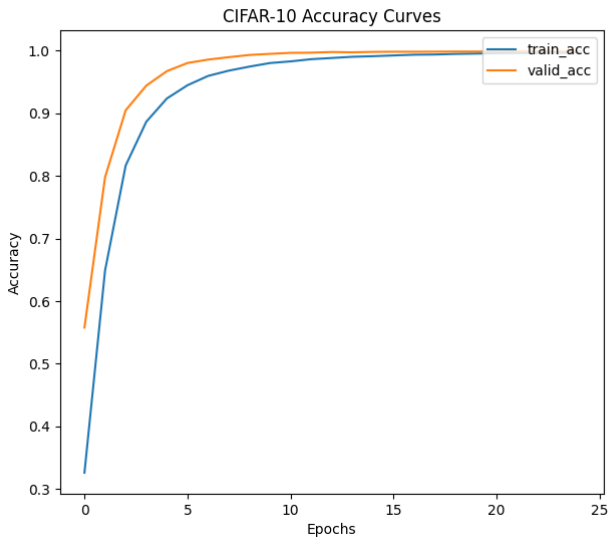




```
Sun Apr 12 10:00:09 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   45C    P0   121W / 300W |   4126MiB / 32480MiB |     16%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
| N/A   37C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
| N/A   36C    P0    44W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   40C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    406045      C   python                                     4115MiB |
+-----------------------------------------------------------------------------+
```

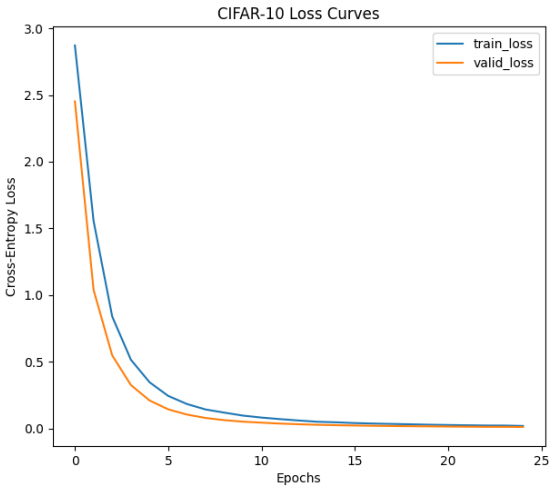
2. With # of GPUs: 2 and Data Parallelism: Yes

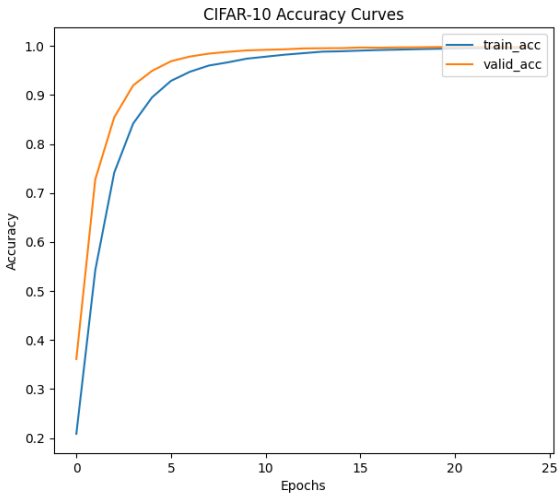




```
Sun Apr 12 11:03:28 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   44C    P0    60W / 300W |   2932MiB / 32480MiB |     28%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
246/A   40C    P0    57W / 300W |   286 923 / 32480MiB |     24%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
| N/A   36C    P0    44W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   39C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    427563      C   python                                     2921MiB |
|    1    427563      C   python                                     2851MiB |
+-----------------------------------------------------------------------------+
```

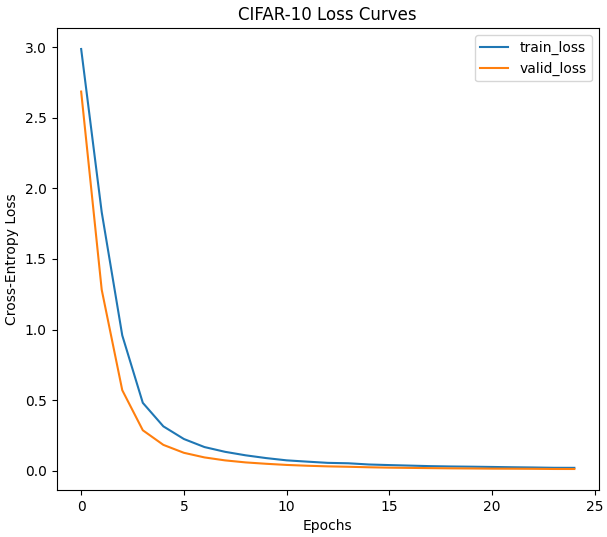
3. With # of GPUs: 3 and Data Parallelism: Yes

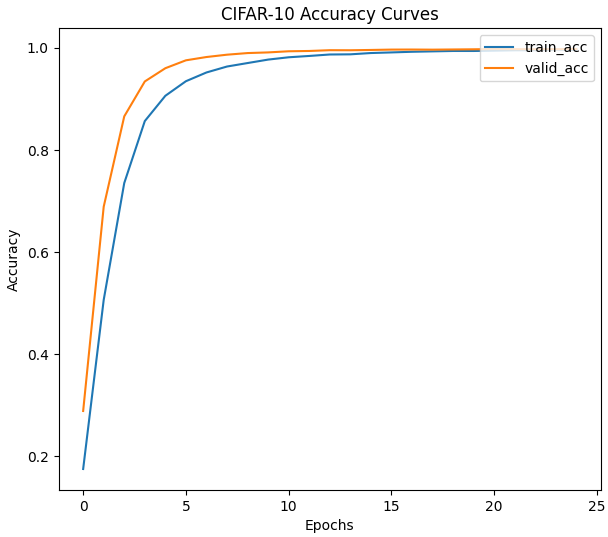




```
Sun Apr 12 12:22:22 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   46C    P0    67W / 300W |   1820MiB / 32480MiB |     14%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
14N/A   41C    P0    59W / 300W |   17506iB / 32480MiB |      9%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
31N/A   40C    P0    58W / 300W |   1730MiB / 32480MiB |     14%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   41C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0      9650      C   python                                     1809MiB |
|    1      9650      C   python                                     1739MiB |
|    2      9650      C   python                                     1719MiB |
+-----------------------------------------------------------------------------+
```

4. With # of GPUs: 4 and Data Parallelism: Yes





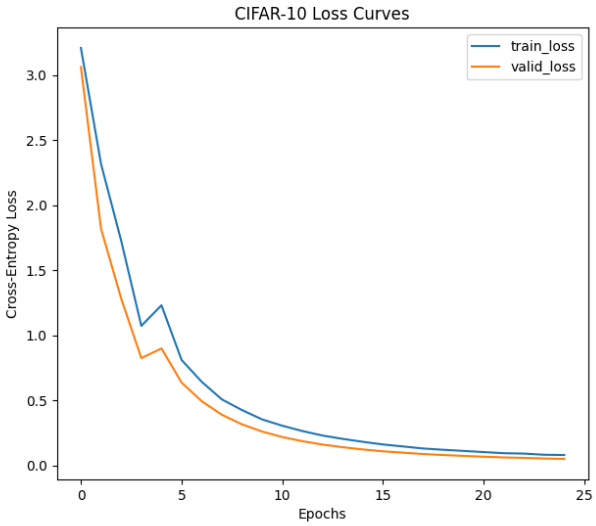
```
Sun Apr 12 13:17:32 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   44C    P0    61W / 300W |   1660MiB / 32480MiB |     14%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
 1N/A   40C    P0    57W / 300W |   1590MiB / 32480MiB |     10%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
10N/A   40C    P0    59W / 300W |   1590MiB / 32480MiB |     10%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
10N/A   44C    P0    60W / 300W |   1590MiB / 32480MiB |     10%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0     57479      C   python                                     1649MiB |
|    1     57479      C   python                                     1579MiB |
|    2     57479      C   python                                     1579MiB |
|    3     57479      C   python                                     1579MiB |
+-----------------------------------------------------------------------------+
```

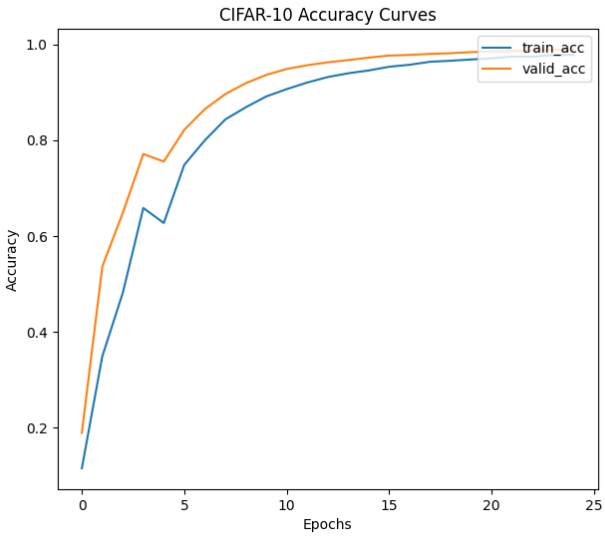
Experiment 2	GPU 0 (GB)	GPU 1 (GB)	GPU 2 (GB)	GPU 3 (GB)	Exec. Time (secs)	Data Parallel	Train Acc	Valid Acc	Test Acc
Using 1GPU	2.859	NA	NA	NA	2461.46	No	99.66	99.87	99.88
Using 2GPU	2.921	2.851	NA	NA	2360.52	Yes	99.71	99.88	99.71
Using 3GPU	1.809	1.739	1.719	NA	2028.72	Yes	99.64	99.82	99.82
Using 4GPU	1.649	1.579	1.579	1.579	2350.78	Yes	99.86	99.88	99.94

5.3 Experiment 3

Batch Size: 512
# of training batches: 150
# of validation batches: 17
 # testing batches: 4

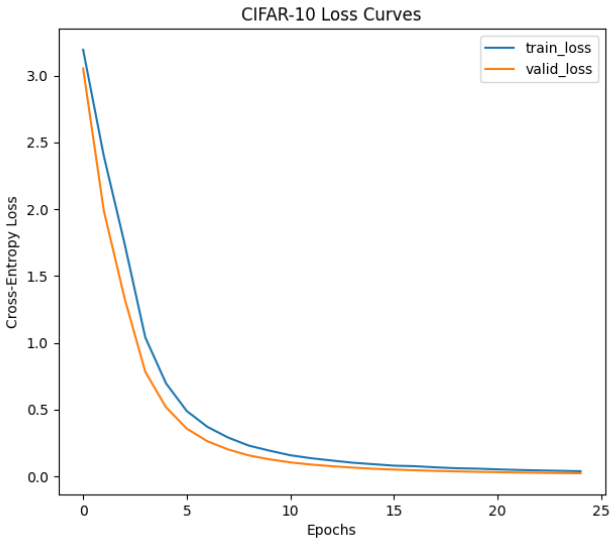
1. With # of GPUs: 1 and Data Parallelism: No

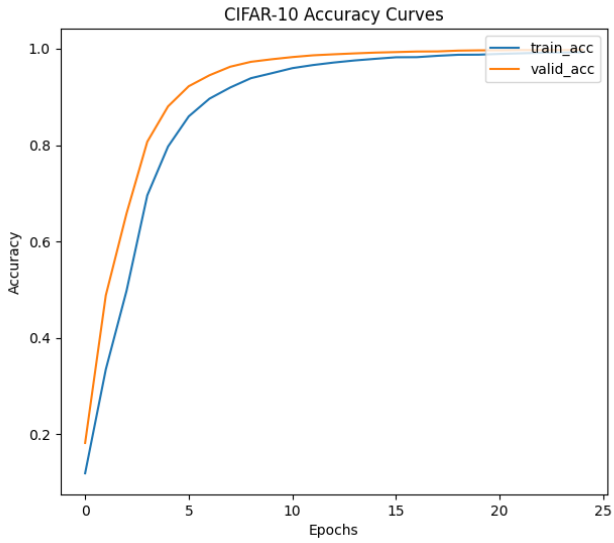





```
Sun Apr 12 17:38:13 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   47C    P0    73W / 300W |   7414MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
| N/A   37C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
| N/A   36C    P0    44W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   40C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    214145      C   python                                     7403MiB |
+-----------------------------------------------------------------------------+
```

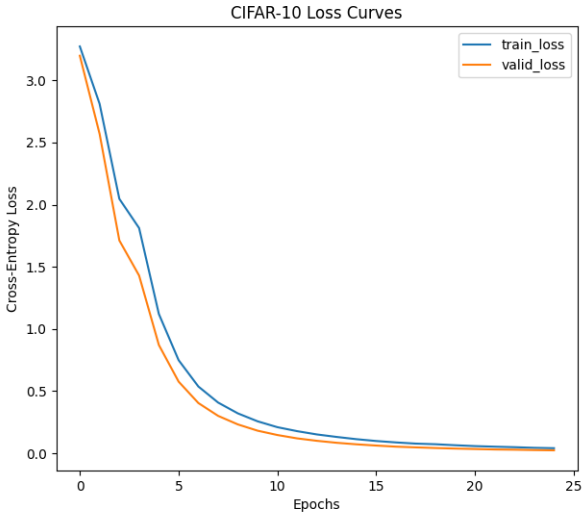
2. With # of GPUs: 2 and Data Parallelism: Yes

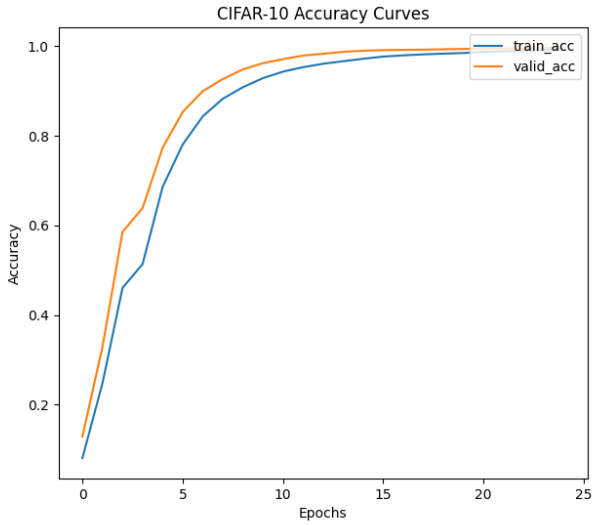





```
Sun Apr 12 16:56:17 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   48C    P0   105W / 300W |   3260MiB / 32480MiB |     38%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
47N/A   43C    P0    70W / 300W |   312417B / 32480MiB |     21%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
| N/A   36C    P0    44W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   39C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    187854      C   python                                     3249MiB |
|    1    187854      C   python                                     3113MiB |
+-----------------------------------------------------------------------------+
```

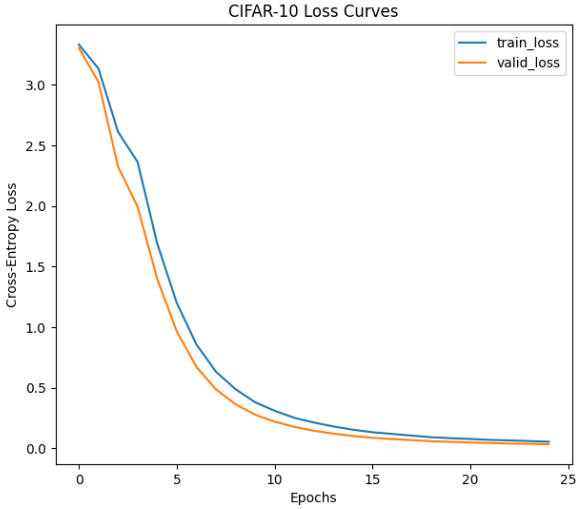
3. With # of GPUs: 2 and Data Parallelism: Yes

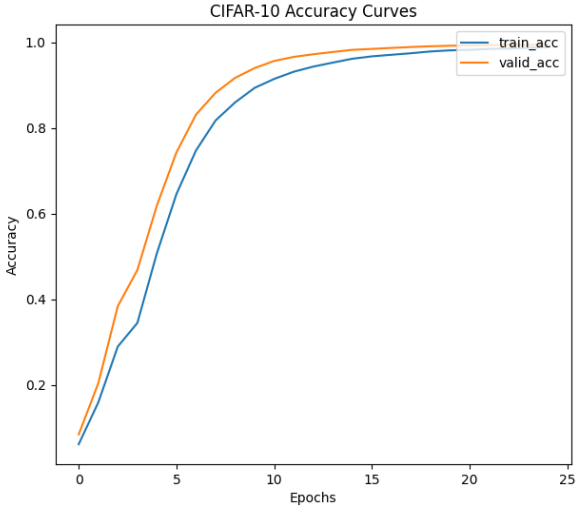





```
Sun Apr 12 15:49:47 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   46C    P0    73W / 300W |   3644MiB / 32480MiB |     40%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
 0N/A   40C    P0    57W / 300W |   35087iB / 32480MiB |      1%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
120/A   40C    P0    58W / 300W |   3488M 0 / 32480MiB |     35%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
| N/A   40C    P0    45W / 300W |     11MiB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    157744      C   python                                     3633MiB |
|    1    157744      C   python                                     3497MiB |
|    2    157744      C   python                                     3477MiB |
+-----------------------------------------------------------------------------+
```

4. With # of GPUs: 2 and Data Parallelism: Yes



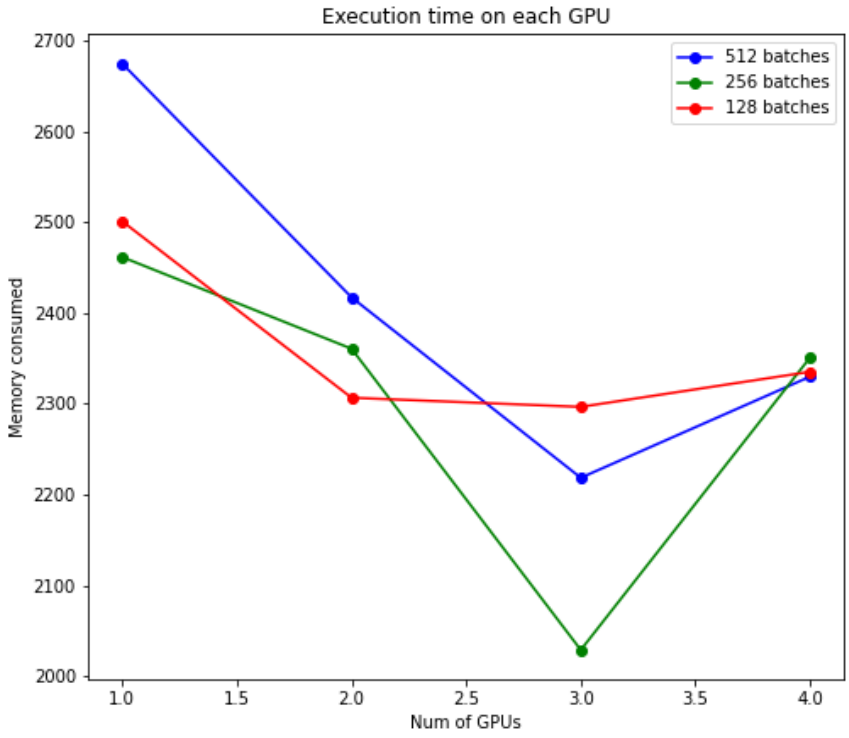



```
Sun Apr 12 14:27:02 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.48                 Driver Version: 410.48                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:18:00.0 Off |                    0 |
| N/A   44C    P0    63W / 300W |   3046MiB / 32480MiB |      3%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:3B:00.0 Off |                    0 |
25N/A   40C    P0    57W / 300W |   29125iB / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  On   | 00000000:86:00.0 Off |                    0 |
25N/A   39C    P0    58W / 300W |   291025B / 32480MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  On   | 00000000:AF:00.0 Off |                    0 |
249/A   44C116 P0    24W / 300W |   2910M B / 32480MiB |      7%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0    118130      C   python                                     3035MiB |
|    1    118130      C   python                                     2899MiB |
|    2    118130      C   python                                     2899MiB |
|    3    118130      C   python                                     2899MiB |
+-----------------------------------------------------------------------------+
```

Experiment 3	GPU 0 (GB)	GPU 1 (GB)	GPU 2 (GB)	GPU 3 (GB)	Exec. Time (secs)	Data Parallel	Train Acc	Valid Acc	Test Acc
Using 1GPU	7.403	NA	NA	NA	2674.47	No	97.86	98.93	98.62
Using 2GPU	3.219	3.113	NA	NA	2417.13	Yes	99.25	99.77	99.71
Using 3GPU	3.633	3.497	3.477	NA	2218.39	Yes	99.19	99.67	99.42
Using 4GPU	3.035	2.899	2.899	2.899	2329.82	Yes	98.91	99.53	99.42

6. Observations

6.1 Observation 1

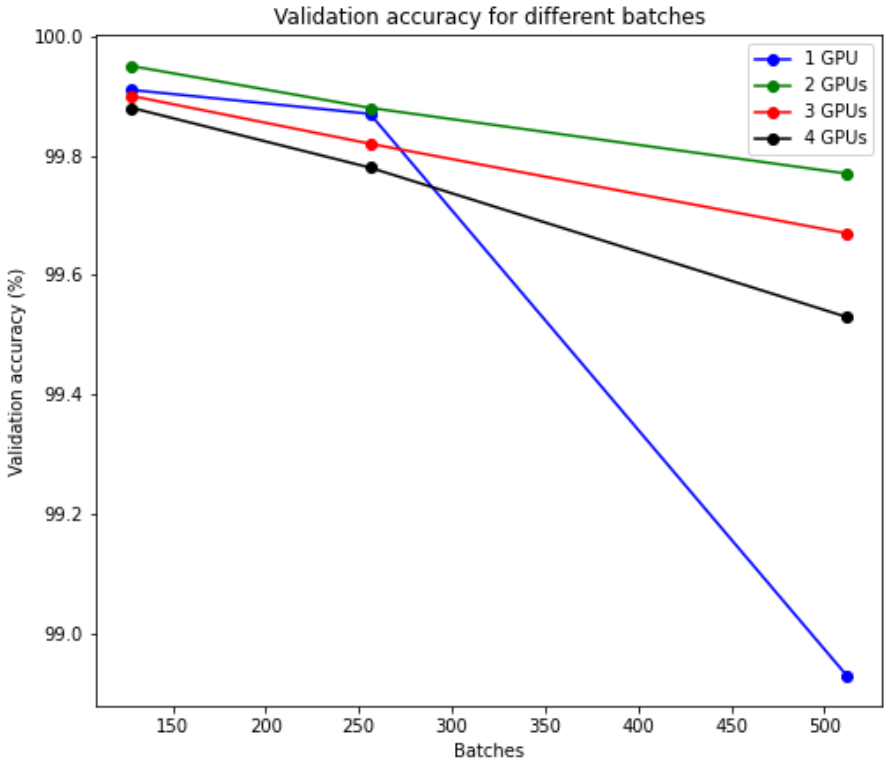


Using more GPU's does not necessarily improve performance. There is a sweet spot where for a specific batch size we get good performance. In the figure above, 256 batches with 3 GPUs in Data Parallelism performs best. Our hypothesis was:

Hypothesis 2: Using more GPU's always speeds up computation and saves time.

Increasing GPUs does not always increase performance of model and hence null hypothesis is true and the above earlier stated alternate hypothesis is false.

6.2 Observation 2



Using larger batches does not necessarily improve accuracy or model approximation capabilities. In fact, more the batch size, lesser the generalization capabilities in our case. In the figure above, 128 batches with 2 GPUs in Data Parallelism performs best. Our hypothesis was

Hypothesis 3: Larger batch size means better model approximation.

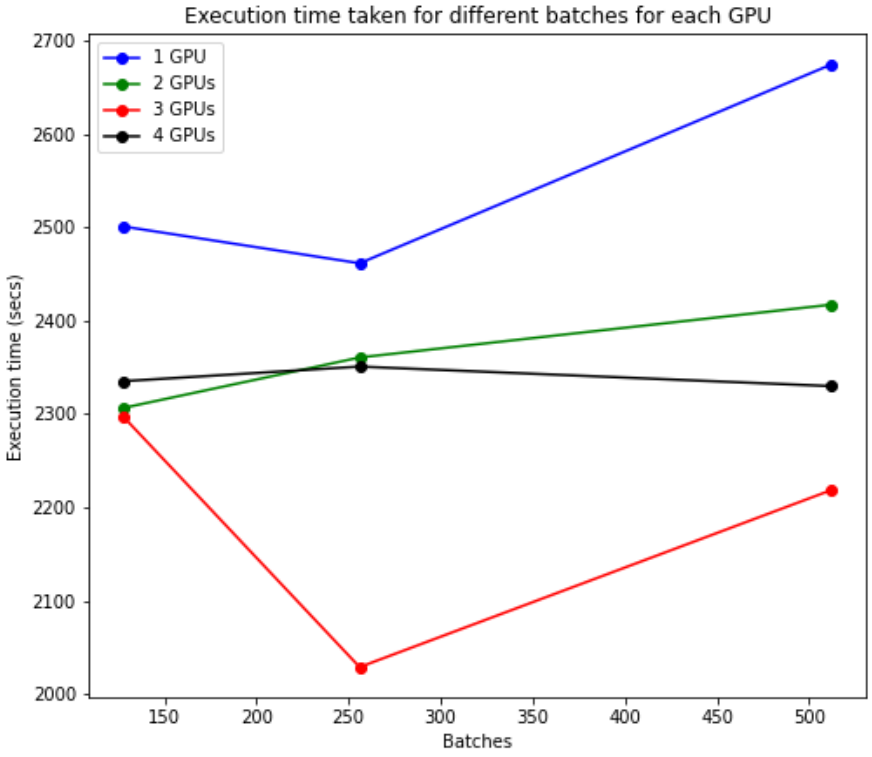
Larger Batch size does not always lead to better model approximation. Hence the null hypothesis is true and the above mentioned alternate hypothesis is false.

6.3 Observation 3

We observe that execution time is optimal for a particular value of batch size. In the figure below, 256 batches seem to be the optimal batch size for Data Parallelism followed by 128 batches. Our hypothesis was:

Hypothesis 4: Keeping number of GPUs constant, there exists a linear relationship between number of batches and time taken. In short, as number of batches increases, time decreases.

The null hypothesis holds true for this case and the above-mentioned alternate hypothesis is false. There does not exist any specific relationship and factors like GPU memory and dataset size play an important role.

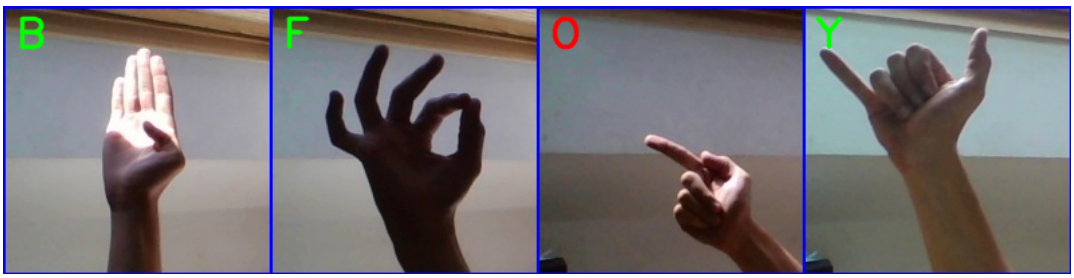


6.4 Observation 4

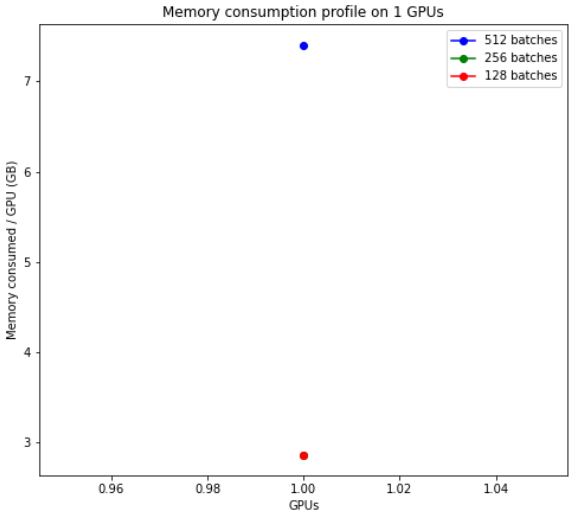
Performance of the model is exceptional with a few false classifications. A real time system is yet to be implemented wherein a stream of images comes from a camera input and we get corresponding output. Our hypothesis was:

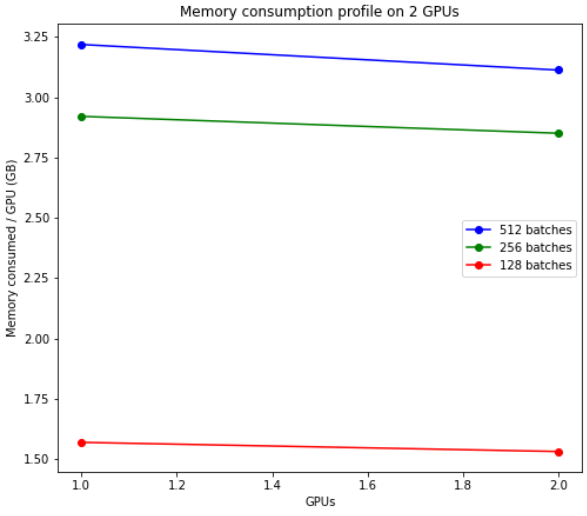
Hypothesis 1: We can recognize hand signs from images or stream of images using state of the art techniques.

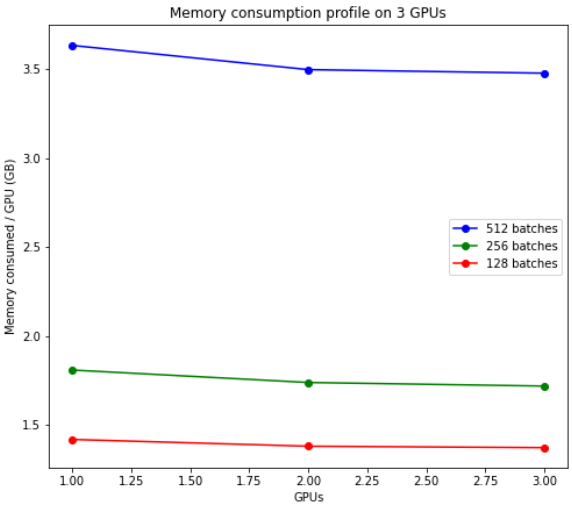
Yes, we can develop a good system using state-of-the art methods for recognizing hand signs. So, our above alternate hypothesis is true.

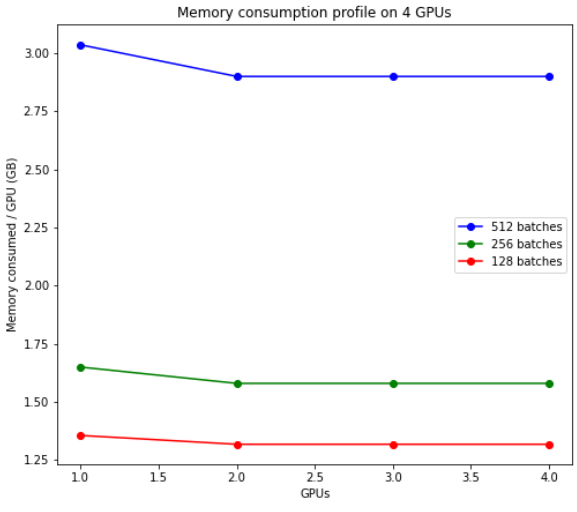


6.5 Memory Consumption on GPUs









6.6 CPU Performance

Since we had high power GPUs with us, we decided to just run 1 epoch in the CPU for verifying how much time does it take. As can been seen below, it took 1383.2038 secs for a single run. That's ~25 mins for a single epoch.

```
Validation Batch number: 064/067, Validation Loss: 1.1160, Validation Acc: 71.8750
Validation Batch number: 065/067, Validation Loss: 1.1788, Validation Acc: 63.2812
Validation Batch number: 066/067, Validation Loss: 1.3614, Validation Acc: 58.7500
-------------------------------------------------------------------------------
Epoch: 001, Train Loss: 2.0917, Train Acc: 39.3291%, Valid Loss : 1.1734, Valid Acc: 67.7767%, Time: 1383.2038s
-------------------------------------------------------------------------------

Previous Best loss: inf | New Best Loss: 1.1734 | Saving Best model...

Epoch: 2/25
Batch: 000/600, Training Loss: 1.1542, Training Acc: 67.1875
Batch: 001/600, Training Loss: 1.6125, Training Acc: 52.3438
Batch: 002/600, Training Loss: 2.4055, Training Acc: 37.5000
Batch: 003/600, Training Loss: 3.2568, Training Acc: 17.9688
Batch: 004/600, Training Loss: 1.9958, Training Acc: 40.6250
```

```
  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
333933 lad.n     20   0 43.6G 24.8G 52184 S 98.6 13.2  4:35.52 python main.py
3.6G 24.8.9R 47.6 126.8 13.2.52:42.9352184 S 36.0 13.2  0:14.99 python main.py
```

7. Conclusions

1. There exists a good tradeoff between number of batches and num of GPUs to be used in parallel. As can be seen from the above analysis, memory usage was not optimal considering amount of memory being wasted on each GPU and hence we would want to have efficient memory distribution as well as time utilization
2. There is a clear under usage of GPU resources and maybe spinning a smaller GPU would have helped me. Nvidia K-80 GPU would have helped a lot and
3. We need to use up to 90-95% of GPU memory and only if a batch does not fit a memory, we should use Data Parallelism.

4. In our analysis, 256 Batches, with 3 GPUs in data parallelism performs best with a total of 5.267GB of memory being consumed for computation.

8. References

1. S. P. More and A. Sattar, "Hand gesture recognition system using image processing," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 671-675.
2. Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
3. Imagenet classification with deep convolutional neural networks, A Krizhevsky, I Sutskever, GE Hinton - Advances in neural information processing systems, 2012
4. Measuring the effects of data parallelism on neural network training, CJ Shallue, J Lee, J Antognini, J Sohl-Dickstein… - arXiv preprint arXiv:1811.03600, 2018
5. On model parallelization and scheduling strategies for distributed machine learning, S Lee, JK Kim, X Zheng, Q Ho, GA Gibson, EP Xing - Advances in neural information processing systems, 2014