High Performance Computing with GPUs
Neural Network Acceleration

[Github Repo]

Team Members	Roll Number
Hania Aamer	22i-1154
Ayesha Ejaz	22i-0899
Aisha Siddiqa	22i-1281

Execution Time of Naive V2 Implementation

```
MNIST Neural Network - GPU Implementation (V2)

Epoch 1 - Loss: 0.2700 - Train Accuracy: 91.80% - Time: 12.960s
Epoch 2 - Loss: 0.1072 - Train Accuracy: 96.81% - Time: 12.902s
Epoch 3 - Loss: 0.0732 - Train Accuracy: 97.85% - Time: 12.582s
Total training time: 38.444s
Test Accuracy: 97.05%
```

Gprof Profile

Flat Profile:
Total execution time: 38.444 seconds.

	Functions	Time (%)	Run time (12.45s)	Calls
Most Expensive	trainGPU()	28.57%	(0.22s)	1
Secondary	loadMNISTImages()	23.38%	(0.40s)	2
Negligible	forwardGPU, backwardGPU, transferNetworkToGPU	2.01%	(0.25s)	Not more than twice
Most time taking overall	trainGPU()	37.4%	(0.29s total)	once

- ➔ trainGPU() dominates execution time.
- ➔ forwardGPU() and backwardGPU() are called 190,000 and 180,000 times respectively, but each call is very fast (0.02s and 0.01s self time, 5.2% and 3.9% of total time)
- ➔ CUDA runtime functions (e.g., __cudart773, __cudart798) together account for a noticeable portion (~20%) of the time.