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 Optimized V3 Implementation

```
MNIST Neural Network - Optimized GPU Implementation (V3)

Epoch 1 - Loss: 0.2675 - Train Accuracy: 91.86% - Time: 7.566s
Epoch 2 - Loss: 0.1062 - Train Accuracy: 96.83% - Time: 7.181s
Epoch 3 - Loss: 0.0734 - Train Accuracy: 97.82% - Time: 7.501s
Total training time: 22.248s
Test Accuracy: 97.07%
```

Gprof Profile

```
Flat profile:

Each sample counts as 0.01 seconds.
  %   cumulative   self              self     total
 time   seconds   seconds    calls  ms/call  ms/call  name
 32.56      0.14     0.14        1   140.00   159.47  trainGPU(NeuralNetwork*, NeuralNetworkGPU*, float**, float**, int)
 18.60      0.22     0.08        2    40.00    40.00  loadMNISTImagesPinned(char const*, int, float**)
  9.30      0.26     0.04                             cudaLaunchKernel
  4.65      0.28     0.02                             __cudart545
  4.65      0.30     0.02                             _init
  4.65      0.32     0.02                             cudaStreamSynchronize
  2.33      0.33     0.01   190000     0.00     0.00  __device_stub__Z19forwardOutputKernelPfS_S_S_ii(float*, float*, float*, float*, int, int)
  2.33      0.34     0.01   180000     0.00     0.00  __device_stub__Z20hiddenGradientKernelPfS_S_S_ii(float*, float*, float*, float*, int, int)
  2.33      0.35     0.01                             __cudart1057
  2.33      0.36     0.01                             __cudart1608
  2.33      0.37     0.01                             __cudart504
  2.33      0.38     0.01                             __cudart513
  2.33      0.39     0.01                             __cudart590
  2.33      0.40     0.01                             __cudart643
  2.33      0.41     0.01                             __cudart798
  2.33      0.42     0.01                             cudaFree
  2.33      0.43     0.01                             cudaMemcpyAsync
  0.00      0.43     0.00   190000     0.00     0.00  forwardGPU(NeuralNetworkGPU*, float*, float*, float*)
  0.00      0.43     0.00   190000     0.00     0.00  __device_stub__Z13softmaxKernelPfi(float*, int)
  0.00      0.43     0.00   190000     0.00     0.00  __device_stub__Z19forwardHiddenKernelPfS_S_S_ii(float*, float*, float*, float*, int, int)
  0.00      0.43     0.00   180000     0.00     0.00  backwardGPU(NeuralNetworkGPU*, float*, float*, float*, float*)
  0.00      0.43     0.00   180000     0.00     0.00  __device_stub__Z20outputGradientKernelPfS_S_i(float*, float*, float*, int)
  0.00      0.43     0.00   180000     0.00     0.00  __device_stub__Z25updateHiddenWeightsKernelPfS_S_S_fii(float*, float*, float*, float*, float, int, int)
  0.00      0.43     0.00   180000     0.00     0.00  __device_stub__Z25updateOutputWeightsKernelPfS_S_S_fii(float*, float*, float*, float*, float, int, int)
  0.00      0.43     0.00        4     0.00     0.00  freePinnedMatrix(float**, float*)
  0.00      0.43     0.00        4     0.00     0.00  allocatePinnedMatrix(int, int, float**)
  0.00      0.43     0.00        2     0.00     0.00  loadMNISTLabelsPinned(char const*, int, float**)
  0.00      0.43     0.00        1     0.00     0.53  evaluateGPU(NeuralNetworkGPU*, float**, float**, int)
  0.00      0.43     0.00        1     0.00     0.00  freeNetwork(NeuralNetwork*)
  0.00      0.43     0.00        1     0.00     0.00  createNetwork()
  0.00      0.43     0.00        1     0.00     0.00  freeNetworkGPU(NeuralNetworkGPU*)
  0.00      0.43     0.00        1     0.00     0.00  transferNetworkToCPU(NeuralNetworkGPU*, NeuralNetwork*)
  0.00      0.43     0.00        1     0.00     0.00  transferNetworkToGPU(NeuralNetwork*)
```

Flat Profile:
Total execution time: 38.444 seconds.

	Functions	**Time (%)**	**Run time (s)**	**Calls**
Most Expensive	trainGPU()	32.56	0.14	1
Secondary	loadMNISTImages()	18.60	0.08s	2
Negligible	ALMOST ALL (non-CUDA)	~0.0	0.25s	
Most time taking overall	trainGPU()	37.1	0.16 total	1

- ➔ trainGPU() is the most time-intensive function, consuming 37.1% of total execution time (0.14s out of 0.43s).
- ➔ forwardGPU() and backwardGPU() are called 190,000 and 180,000 times respectively, but each call is extremely fast and together account for only 0.02s of self time.
- ➔ CUDA runtime and kernel launch functions (e.g., cudaLaunchKernel, cudaStreamSynchronize) together account for a significant portion of the remaining time.