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

```
Each sample counts as 0.01 seconds.
  %   cumulative   self              self     total
 time   seconds   seconds    calls  ms/call  ms/call  name
 30.43      0.21     0.21        2   105.00   105.00  loadMNISTImages(char const*, int)
 24.64      0.38     0.17        1   170.00   189.47  trainGPU(NeuralNetwork*, NeuralNetworkGPU*, double**, double**, int)
  8.70      0.44     0.06                             cudaMemcpy
  5.80      0.48     0.04                             __cudart513
  4.35      0.51     0.03                             __cudart545
  4.35      0.54     0.03                             __cudart798
  4.35      0.57     0.03                             cudaGetLastError
  2.90      0.59     0.02                             __cudart610
  2.90      0.61     0.02                             _init
  1.45      0.62     0.01   190000     0.00     0.00  forwardGPU(NeuralNetworkGPU*, double*, double*, double*)
  1.45      0.63     0.01   180000     0.00     0.00  backwardGPU(NeuralNetworkGPU*, double*, double*, double*, double*)
  1.45      0.64     0.01                             __cudaPopCallConfiguration
  1.45      0.65     0.01                             __cudart475
  1.45      0.66     0.01                             __cudart556
  1.45      0.67     0.01                             __cudart773
  1.45      0.68     0.01                             cudaFree
  1.45      0.69     0.01                             cudaLaunchKernel
  0.00      0.69     0.00   190000     0.00     0.00  __device_stub__Z13softmaxKernelPdi(double*, int)
  0.00      0.69     0.00   190000     0.00     0.00  __device_stub__Z19forwardHiddenKernelPdS_S_S_ii(double*, double*, double*, double*, int, int)
  0.00      0.69     0.00   190000     0.00     0.00  __device_stub__Z19forwardOutputKernelPdS_S_S_ii(double*, double*, double*, double*, int, int)
  0.00      0.69     0.00   180000     0.00     0.00  __device_stub__Z20outputGradientKernelPdS_S_i(double*, double*, double*, int)
  0.00      0.69     0.00   180000     0.00     0.00  __device_stub__Z20hiddenGradientKernelPdS_S_S_ii(double*, double*, double*, double*, int, int)
  0.00      0.69     0.00   180000     0.00     0.00  __device_stub__Z25updateHiddenWeightsKernelPdS_S_S_dii(double*, double*, double*, double*, double, int, int)
  0.00      0.69     0.00   180000     0.00     0.00  __device_stub__Z25updateOutputWeightsKernelPdS_S_S_dii(double*, double*, double*, double*, double, int, int)
  0.00      0.69     0.00        4     0.00     0.00  freeMatrix(double**, int)
  0.00      0.69     0.00        2     0.00     0.00  loadMNISTLabels(char const*, int)
  0.00      0.69     0.00        1     0.00     0.53  evaluateGPU(NeuralNetworkGPU*, double**, double**, int)
  0.00      0.69     0.00        1     0.00     0.00  freeNetwork(NeuralNetwork*)
  0.00      0.69     0.00        1     0.00     0.00  createNetwork()
  0.00      0.69     0.00        1     0.00     0.00  freeNetworkGPU(NeuralNetworkGPU*)
  0.00      0.69     0.00        1     0.00     0.00  transferNetworkToCPU(NeuralNetworkGPU*, NeuralNetwork*)
  0.00      0.69     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()	28.57	0.22	1
Secondary	loadMNISTImages()	23.38	0.40	2
Negligible	forwardGPU, backwardGPU, transferNetworkToGPU	2.01	0.25	
Most time taking overall	trainGPU()	37.4	0.29 total	1

- ➔ 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.