ECE 408 FINAL PROJECT MILESTONE 2 REPORT

Ahmad Syafiq bin Shamsul Bahari (shamsul2)
Anirban Banerjee (banerj10)
Srujun Thanmay Gupta (sgupta80)

1.1: Run the Baseline Forward Pass

We got result:

```
EvalMetric: {'accuracy': 0.8673}
9.98user 2.83system 0:04.12elapsed 310%CPU
```

1.2: Run the baseline GPU implementation

We got result:

```
EvalMetric: {'accuracy': 0.8673}
1.73user 0.92system 0:02.19elapsed 121%CPU
```

The CPU load is considerably less in this run, indicating that most work happened on the GPU instead.

1.3 Generate a NVPROF Profile

Most time consuming kernel calls:

Time (%)	Time (ms)	Kernel Name
37.00%	49.981ms	`void cudnn::detail::implicit_convolve_sgemm(...)`
28.65%	38.704ms	`sgemm_sm35_ldg_tn_128x8x256x16x32`
14.35%	19.385ms	`void cudnn::detail::activation_fw_4d_kernel(...)`
10.70%	14.451ms	`void cudnn::detail::pooling_fw_4d_kernel(...)`
		`...`

From this list, it seems that the GPU spends the bulk of the time in 2 kernels, the actual convolution kernel and a large single precision floating point general matrix multiplication (SGEMM).

2.1 Add a simple CPU forward implementation

Using the baseline CPU implementation, we got the following output.

```
* Running python /src/m2.1.py
New Inference
Loading fashion-mnist data... done
Loading model... done
Op Time: 11.722999
Correctness: 0.8562 Model: ece408-high
```

This matches the given Op Time and Correctness from the README.