M1.1 Milestone 1.1

```
* Running python /src/m1.1.py
...
EvalMetric: {'accuracy': 0.8673}
0:06.77 elapsed
```

M1.2/1.3 Milestone 1.2/1.3

The GPU spends most of its compute time in
* **implicit_convolve_sgemm**
* **sgemm_sm35_ldg_tn_128x8x256x16x32**, and
* **activation_fw_4d_kernel**.

Most of the compute time is spend performing convolution. There are also memory API calls that take more time than computation.

```
* Running nvprof python /src/m1.2.py
...
==305== Profiling result:
Time(%)      Time  Calls       Avg       Min       Max  Name
36.73%   50.515ms      1  50.515ms  50.515ms  50.515ms  implicit_convolve_sgemm
28.66%   39.409ms      1  39.409ms  39.409ms  39.409ms  sgemm_sm35_ldg_tn_128x8x256
14.10%   19.396ms      2  9.6979ms  461.70us  18.934ms  activation_fw_4d_kernel
...
==305== API calls:
Time(%)      Time  Calls       Avg       Min       Max  Name
 46.82%  1.85071s     18  102.82ms  17.325us  925.01ms  cudaStreamCreateWithFlags
 28.48%  1.12580s     10  112.58ms     949ns  320.49ms  cudaFree
 20.78%  821.31ms     24  34.221ms  236.10us  814.20ms  cudaMemGetInfo
...
```

M2.1 Milestone 2.1

```
* Running python /src/m2.1.py
Loading fashion-mnist data... done
Loading model... done
Op Time: 12.146829
Correctness: 0.8562 Model: ece408-high

* Running python m2.1.py ece408-low 10000
Loading fashion-mnist data... done
Loading model... done
Op Time: 12.801514
Correctness: 0.629 Model: ece408-low
```

Team Member Contributions

Sujay: Milestone 2 code
Tanishq: Milestone 2 pdf
Gordon: pdf edits