ECE 408 Final Project Milestone 1 Report

Yiqing Zhou (yiqing2)
Nuocheng Pan(np9)

March 10, 2019

Kernels that collectively consume more than 90% of the program time:

```
40.45%:   [CUDA memcpy HtoD]
20.32%:   implicit_convolve_sgemm
11.88%:   volta_cgemm_64x32_tn
7.07%:    op_generic_tensor_kernel
5.62%:    volta_sgemm_128x128_tn
5.61%:    fft2d_c2r_32x32
4.52%:    pooling_fw_4d_kernel
3.70% :   fft2d_r2c_32x32
```

CUDA API calls that collectively consume more than 90% of the program time:

```
42.61%    cudaStreamCreateWithFlags
34.35%    cudaMemGetInfo
21.02%    cudaFree
```

Explanation of difference between kernels and API calls:

Kernels are functions programed by users. Kernels are launched by host and run on devices. APIs are provided by CUDA runtime system and could be directly called by users.

CPU output and runtime: (runtime is bolded)

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
8.98user 3.57system 0:05.07elapsed 247%CPU (0avgtext+0avgdata
2470144maxresid
ent)k
0inputs+2824outputs (0major+668695minor)pagefaults 0swaps
```

GPU output and runtime: (runtime is bolded)

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
4.40user 3.12system 0:04.38elapsed 171%CPU (0avgtext+0avgdata
2840696maxresident)k
0inputs+4552outputs (0major+660254minor)pagefaults 0swaps
```