# ECE 408 Final Project Report

Yiqing Zhou (yiqing2)
Nuocheng Pan(np9)

March 13, 2019

# Milestone 1: RAI Setup

## 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
```

# Milestone 2: CPU Convolution Implementation

## OP and Exec Time for different input data size:

**✱ Running /usr/bin/time python m2.1.py 100**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.034094
Op Time: 0.075474
Correctness: 0.84 Model: ece408
2.87user 2.76system 0:01.00elapsed 562%CPU (0avgtext+0avgdata
203620maxresident)k
0inputs+8outputs (0major+61034minor)pagefaults 0swaps
```

**✱ Running /usr/bin/time python m2.1.py 1000**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.245769
Op Time: 0.749210
Correctness: 0.852 Model: ece408
4.29user 3.00system 0:02.00elapsed 363%CPU (0avgtext+0avgdata
331980maxresident)k
0inputs+2824outputs (0major+110686minor)pagefaults 0swaps
```

**✱ Running /usr/bin/time python m2.1.py 10000**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 2.446601
Op Time: 7.594124
Correctness: 0.8397 Model: ece408
15.54user 4.46system 0:11.65elapsed 171%CPU (0avgtext+0avgdata
1617164maxresident)k
0input
s+2824outputs (0major+617305minor)pagefaults 0swaps
```