ECE 408 Project Report

Team Name: fast_af
Names: Omar Taha, Jason Gong, Shreyas Byndoor
NetIDs: otaha2, xg8, byndoor2
Affiliation: Chicago City Scholars

Milestone 1

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

- [CUDA memcpy HtoD]
- cudnn::detail::implicit_convolve_sgemm
- volta_cgemm_64x32_tn
- op_generic_tensor_kernel
- fft2d_c2r_32x32
- volta_sgemm_128x128_tn
- cudnn::detail::pooling_fw_4d_kernel
- fft2d_r2c_32x32

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

- cudaStreamCreateWithFlags
- cudaMemGetInfo
- cudaFree

Kernels vs. API Calls:

CUDA Kernels are simply defined as regular C functions. However, unlike typical C functions, CUDA Kernels are executed N times in parallel by N different CUDA threads. Meanwhile, CUDA APIs provide C functions that execute on the host to allocate and deallocate device memory, transfer data between host memory and device memory, manage systems with multiple devices, etc. They are not executed by each CUDA thread as a Kernel is.

Output of rai running MXNet on the CPU:

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
9.30user 3.46system 0:05.34elapsed 239%CPU (0avgtext+0avgdata 2471560maxresident)k
0inputs+2824outputs (0major+666761minor)pagefaults 0swaps
```

Program run time:

```
0:05.34 elapsed
```

Output of rai running MXNet on the GPU:

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
4.41user 3.28system 0:04.34elapsed 177%CPU (0avgtext+0avgdata 2837968maxresident)k
0inputs+4552outputs (0major+661333minor)pagefaults 0swaps
```

Program run time:

0:04.34 elapsed

Milestone 2

program execution time:

0:15.52 elapsed

Op Times:

1) Op Time: 2.826305
2) Op Time: 11.143122