ECE 408/CS 483 Final Project

Team: elegant_and_easygoing_boys
Team members:
Licheng Luo (ll6)
Zhengqi Fang (zf4)
Ruian Pan (ruianp2)
Affiliation: on campus

Milestone 2

All kernels that collectively consume more than 90% of the program time

volta_scudnn_128x64_relu_interior_nn_v1
volta_gcgemm_64x32_nt
fft2d_c2r_32x32
volta_sgemm_128x128_tn
op_generic_tensor_kernel
fft2d_r2c_32x32
cudnn::detail::pooling_fw_4d_kernel

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

cudaStreamCreateWithFlags
cudaMemGetInfo
cudaFree

Explanation of the difference between kernels and API calls

Kernels are the codes that run on GPU and do the parallel computations.
API calls are the calls to the CUDA's APIs, which are defined by CUDA(NVIDIA). They are usually used to do the initializations such as memory allocations and transfer.

Output of RAI running MXNet on the CPU (m1.1)

EvalMetric: {'accuracy': 0.8154}

Run time

User	20.89
System	7.39
Elapsed	0:10.28

Output of RAI running MXNet on the GPU (m1.2)

EvalMetric: {'accuracy': 0.8154}
Run time

User	5.10
System	2.69
Elapsed	0:05.01

CPU Implementation

Correctness: 0.7653 Model: ece408
Run time (m2.1)

User	90.33
System	10.19
Elapsed	1:19.22

Op Times

Op Time: 13.540072
Op Time: 60.894102