CS483/ECE408 Final Project

Team: kriskr
Team member:

Ruoxi Yang	ryang28
Rongzi Wang	rwang67
Shiqi Sun	shiqis2

Milestone 1

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

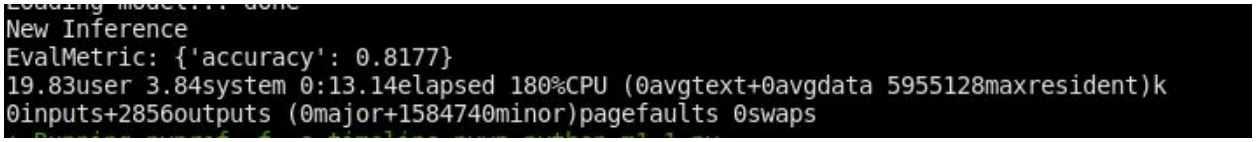
Name	Time	Percentage
CUDA memcpy HtoD	37.345ms	38.63%
volta_scudnn_128x32_relu_interior_nn_v1	20.831ms	21.55%
implicit_convolve_sgemm	19.117ms	19.78%
activation_fw_4d_kernel	7.4430ms	7.70%
volta_sgemm_128x128_tn	6.7882ms	7.02%
pooling_fw_4d_kernel	4.4341ms	4.59%

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

Name	Time	Percentage
cudaStreamCreateWithFlags	2.80865s	39.54%
cudaMemGetInfo	2.47336s	34.82%
cudaFree	1.55123s	21.84%

Explanation of the difference between kernels and API calls:
Kernel is a C function that programer define and expect the cuda threads to execute. The runtime API eases device code management by providing implicit initialization, context management, and module management, making it easier to code. During the runtime, all the kernels are automatically loaded during initialization and stay loaded for as long as the program runs.

Output of rai running MXNet on the CPU: The accuracy is 0.8177.
Program run time: The elapsed time is 0:13:14



Output of rai running MXNet on the GPU: accuracy is 0.8177
Program run time: The elapsed time is 0:05.80

```
New Inference
EvalMetric: {'accuracy': 0.8177}
4.31user 2.85system 0:05.80elapsed 123%CPU (0avgtext+0avgdata 2836900maxresident)k
0inputs+4568outputs (0major+703473minor)pagefaults 0swaps
```

Milestone 2

Name	Layer 1	Layer 2
Op Time	22.041992s	105.005034s

Whole program execution time: 127.09s