ECE 408 Final Project Report
Andrew Betbadal (betbadl2), Michael Beaudin (mbeaudi2), Atin Ganti (ganti3)
big_baller_brand_123g

Milestone 1:

Kernels that take 90% of program time

Cuda memcpy HtoD
volta_scudnn_128x32_relu_interior_nn_v1
implicit_convolve_sgemm
volta_sgemm_128x128_tn
activation_fw_4d_kernel

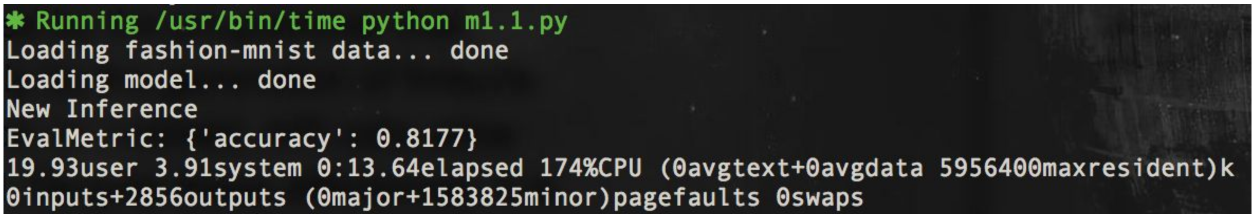
API calls that take 90 % of program time

cudaStreamCreateWithFlags
cudaMemGetInfo
cudaFree

Include an explanation of the difference between kernels and API calls

In Cuda, kernels are C functions defined by the programmer that allow a programmer to execute code parallely by the generation of threads instead of behaving like a standard C function. They must be signified by the global keyword. In contrast, API calls are a set of subroutine definitions, protocols, and tools that allow building software. API calls like cudaFree are required in order to build the software effectively.

Show output of rai running MXNet on the CPU

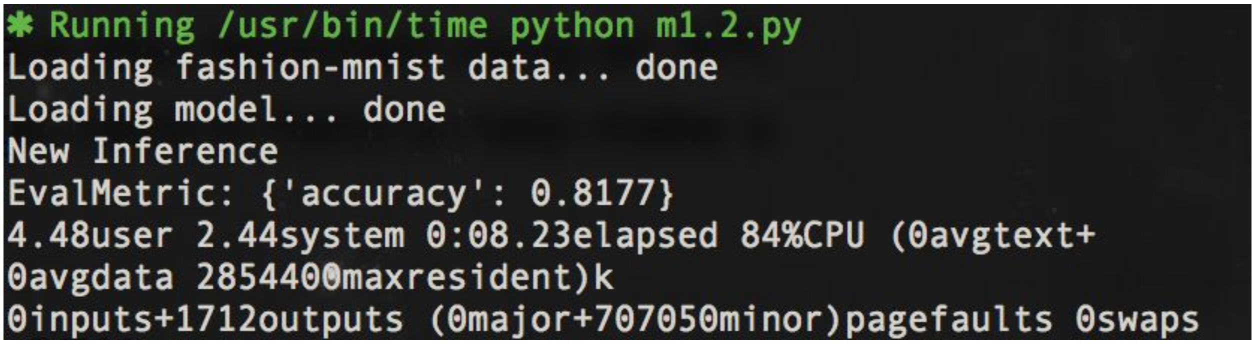


```
✱ Running /usr/bin/time python m1.1.py
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8177}
19.93user 3.91system 0:13.64elapsed 174%CPU (0avgtext+0avgdata 5956400maxresident)k
0inputs+2856outputs (0major+1583825minor)pagefaults 0swaps
```

List program run time

Run Time : 13.64

Show output of rai running MXNet on the GPU



```
✱ Running /usr/bin/time python m1.2.py
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8177}
4.48user 2.44system 0:08.23elapsed 84%CPU (0avgtext+
0avgdata 2854400maxresident)k
0inputs+1712outputs (0major+707050minor)pagefaults 0swaps
```

List program run time

Run Time : 08.23