Names: Brian Tang (bt3), Joseph Adamo (jdadamo2), Kyle Jew (kjew2)

Team Name: thread_beast

Affiliation: On-Campus students

List of kernels consuming more than 90% of program time:

None, highest is [CUDA memcpy HtoD] at 30.04% of time.

List of CUDA API calls consuming more than 90% of program time:

None, highest is cudaStreamCreateWithFlags at 41.19% of time.

Kernels and API calls difference:

Kernels are programmer-defined C functions and when launched, are executed N times in parallel by N different threads. However, CUDA API calls are pre-defined extensions to the C language and meant for easing the experience for programmers to set up programs for execution by the device.

Output of rai running MXNET on CPU:

“Loading fashion-mnist data... done

Loading model... done

New Inference

EvalMetric: {'accuracy': 0.8154}

18.26user 4.46system 0:09.56elapsed 237%CPU (0avgtext+0avgdata 6047060maxresident)k

0inputs+2824outputs (0major+1601873minor)pagefaults 0swaps”

Program run time: 9.56 seconds

Output of rai running MXNET on GPU:

“Loading fashion-mnist data... done

Loading model... done

New Inference

EvalMetric: {'accuracy': 0.8154}

4.97user 3.25system 0:04.59elapsed 179%CPU (0avgtext+0avgdata 2968484maxresident)k

0inputs+4536outputs (0major+733238minor)pagefaults 0swaps”

Program run time: 4.59 seconds

Whole Program Execution Time: 1 minute 16.47 seconds

Op Times:

Python m2.1.py:

Op Time: 12.307846

Op Time: 59.309954

Correctness: 0.7653

At 100 images:

Op Time: 1.082469

Op Time: 5.923644

Correctness: 0.767

At 1000 images:

Op Time: 0.108870

Op Time: 0.590093

Correctness: 0.76

At 10000 images:

Op Time: 10.855807

Op Time: 60.478481

Correctness: 0.7653