ECE 408 - Final Project Report

School: UIUC

Team: wen_mei_we_clutch_that_a

Jacob Brown, jlbrown5 ◇ Brian Yan, yyan18 ◇ Brandon Yu, byu23

Milestone 1

- Kernels that collectively consume more than 90% of the program time.

Time (%)	Time	Name
34.07	118.49ms	fermiPlusCgemmLDS128_batched
27.00	93.897ms	cudnn::detail::implicit_convolve_sgemm
12.70	44.164ms	fft2d_c2r_32x32
8.20	28.518ms	sgemm_sm35_ldg_tn_128x8x256x16x32
6.43	22.368ms	[CUDA memcpy HtoD]
4.08	14.180ms	cudnn::detail::activation_fw_4d_kernel

- API calls that collectively consume more than 90% of the program time.

Time (%)	Time	Name
43.66	1.93s	cudaStreamCreateWithFlags
26.90	1.19s	cudaFree
20.615	911.5ms	cudaMemGetInfo

- Difference between kernels and API calls

  API calls are calls to the cuda api defined in cuda.h. These functions are usually called by the host to set up the kernel and get information about the GPU. All the API calls listed in the table above are high level CUDA runtime API calls, which are built on lower level CUDA driver APIs. Those CUDA runtime API calls make it easier for us to compile our kernels into executables.

  Kernels are user programs that run on the GPU. These are more traditional functions that do the work needed by the user, such as vector addition. However, unlike traditional C functions, CUDA kernels run N (number of threads) times per invocations instead of only once. Which is essentially what makes parallel computing possible.

- Output of rai running MXNet on the CPU

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8444}
```

- Program run time (On CPU)

```
12.77user 6.12system 0:08.42elapsed 224%CPU (0avgtext+0avgdata 2826068maxresident)k
0inputs+2624outputs (0major+39593minor)pagefaults 0swaps
```

- Output of rai running MXNet on the GPU

```
Loading fashion-mnist data... done
Loading model...
[05:18:37] src/operator/./../cudnn_algoreg-inl.h:112: Running performance tests to find
the best convolution algorithm, this can take a while... (setting env variable
MXNET_CUDNN_AUTOTUNE_DEFAULT to 0 to disable)
done
New Inference
EvalMetric: {'accuracy': 0.8444}
```

- Program run time (On GPU)

```
2.12user 1.11system 0:02.71elapsed 119%CPU (0avgtext+0avgdata 1134872maxresident)k
0inputs+512outputs (0major+154708minor)pagefaults 0swaps
```

Milestone 2

- Full Program Run time

Number of images	User Time (s)	System Time (s)	Elapsed Time (s)
10000	30.39	1.41	29.80
100	1.04	0.49	1.01
10	0.75	0.52	0.74

- Op Times

Number of images	Convolutional Layer 1 Op Time (s)	Convolutional Layer 2 Op Time (s)
10000	6.499919	19.373818
100	0.064483	0.193488
10	0.006535	0.019293