CS483/ECE408 report

Guohao Dou(gdou2), Liyu Liu(lliu79), Qichao Gao(qgao10)

March 6, 2018

1 Milestone 1

1.1 a list of all kernels that collectively consume more than 90% of the program time

- fermiPlusCgemmLDS128_batched: 34.09%
- cudnn::detail::implicit_convolve_sgemm: 27.01%
- fft2d_c2r_32x32: 12.65%
- sgemm_sm35_ldg_tn_128x8x256x16x32: 8.20%
- CUDA memcpy HtoD: 6.43%
- cudnn::detail::activation_fw_4d_kernel: 4.07%

Total: 92.45%

1.2 a list of all CUDA API calls that collectively consume more than 90% of the program time.

- cudaStreamCreateWithFlags: 37.37%
- cudaFree: 29.25%
- cudaMemGetInfo: 23.42%

Total: 90.01%

1.3 an explanation of the difference between kernels and API calls

Kernels are the code that will be run by GPU threads. It's usually launched on the host. It does the actual computation.
API calls are utilities provided by CUDA to do chores like memory allocation on GPU and data transfer e.t.c. Their names are usually started by 'cuda'.

1.4 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': 0.8444}
12.48user 6.36system 0:08.40elapsed 224%CPU (0avgtext+0avgdata 2829772maxresident)k
0inputs+2624outputs (0major+38196minor)pagefaults 0swaps
```

1.5 List program run time

8.40 sec.

1.6 output of rai running MXNet on the GPU

```
 Running /usr/bin/time python m1.2.py
Loading fashion-mnist data...
done
Loading model...
21:36:15 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_D
to 0 to disable)
done
New Inference
EvalMetric: 'accuracy': 0.8444
2.14user 1.08system 0:02.70elapsed 119%CPU (0avgtext+0avgdata 1137876maxresident)k
0inputs+512outputs (0major+156089minor)pagefaults 0swaps
```

1.7 List program run time

2.7 secs. (faster than CPU)