ECE 408 Final Project Report

Team Name: tbd

Yiqing Zhou (yiqing2)
Nuocheng Pan(np9)

March 29, 2019

Milestone 1: RAI Setup

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

```
40.45%:   [CUDA memcpy HtoD]
20.32%:   implicit_convolve_sgemm
11.88%:   volta_cgemm_64x32_tn
7.07%:    op_generic_tensor_kernel
5.62%:    volta_sgemm_128x128_tn
5.61%:    fft2d_c2r_32x32
4.52%:    pooling_fw_4d_kernel
3.70% :   fft2d_r2c_32x32
```

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

```
42.61%    cudaStreamCreateWithFlags
34.35%    cudaMemGetInfo
21.02%    cudaFree
```

Explanation of difference between kernels and API calls:

Kernels are functions programed by users. Kernels are launched by host and run on devices. APIs are provided by CUDA runtime system and could be directly called by users.

CPU output and runtime: (runtime is bolded)

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
8.98user 3.57system 0:05.07elapsed 247%CPU (0avgtext+0avgdata
2470144maxresid
ent)k
0inputs+2824outputs (0major+668695minor)pagefaults 0swaps
```

GPU output and runtime: (runtime is bolded)

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
4.40user 3.12system 0:04.38elapsed 171%CPU (0avgtext+0avgdata
2840696maxresident)k
0inputs+4552outputs (0major+660254minor)pagefaults 0swaps
```

Milestone 2: CPU Convolution Implementation

OP and Exec Time for different input data size:

- **Running /usr/bin/time python m2.1.py 100**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.034094
Op Time: 0.075474
Correctness: 0.84 Model: ece408
2.87user 2.76system 0:01.00elapsed 562%CPU (0avgtext+0avgdata
203620maxresident)k
0inputs+8outputs (0major+61034minor)pagefaults 0swaps
```

- **Running /usr/bin/time python m2.1.py 1000**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.245769
Op Time: 0.749210
Correctness: 0.852 Model: ece408
4.29user 3.00system 0:02.00elapsed 363%CPU (0avgtext+0avgdata
331980maxresident)k
0inputs+2824outputs (0major+110686minor)pagefaults 0swaps
```

- **Running /usr/bin/time python m2.1.py 10000**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 2.446601
Op Time: 7.594124
Correctness: 0.8397 Model: ece408
15.54user 4.46system 0:11.65elapsed 171%CPU (0avgtext+0avgdata
1617164maxresident)k
0input
s+2824outputs (0major+617305minor)pagefaults 0swaps
```

Milestone 3: GPU Forward Convolution

Nvprof result:

```
==278== NVPROF is profiling process 278, command: python m3.1.py
Loading model... done
New Inference
Op Time: 0.005826
Op Time: 0.031620
Correctness: 0.8397 Model: ece408
==278== Profiling application: python m3.1.py
==278== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   60.09%  37.378ms         2  18.689ms  5.7804ms  31.598ms  
mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int, int, int, 
int)
                   27.21%  16.927ms        20  846.34us  1.1200us  16.409ms  [CUDA memcpy 
HtoD]
                    3.96%  2.4646ms         2  1.2323ms  20.864us  2.4438ms  
volta_sgemm_32x128_tn
                    3.93%  2.4415ms         2  1.2207ms  737.50us  1.7040ms  void 
mshadow::cuda::MapPlanLargeKernel<mshadow::sv::saveto, int=8, int=1024, 
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=4, float>, float>, 
mshadow::expr::Plan<mshadow::expr::BinaryMapExp<mshadow::op::mul, 
mshadow::expr::ScalarExp<float>, mshadow::Tensor<mshadow::gpu, int=4, float>, float, 
int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=4, int)
                    2.62%  1.6307ms         2  815.34us  22.368us  1.6083ms  void 
op_generic_tensor_kernel<int=2, float, float, float, int=256, cudnnGenericOp_t=7, 
cudnnNanPropagation_t=0, cudnnDimOrder_t=0, int=1>(cudnnTensorStruct, float*, 
cudnnTensorStruct, float const *, cudnnTensorStruct, float const *, float, float, float, 
float, dimArray, reducedDivisorArray)
                    1.70%  1.0576ms         1  1.0576ms  1.0576ms  1.0576ms  void 
cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_func<float, 
cudnnNanPropagation_t=0>, int=0, bool=0>(cudnnTensorStruct, float const *, 
cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_func<float, 
cudnnNanPropagation_t=0>, int=0, bool=0>, cudnnTensorStruct*, cudnnPoolingStruct, float, 
cudnnPoolingStruct, int, cudnn::reduced_divisor, float)
                    0.25%  152.89us         1  152.89us  152.89us  152.89us  void 
mshadow::cuda::MapPlanLargeKernel<mshadow::sv::saveto, int=8, int=1024, 
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, 
mshadow::expr::Plan<mshadow::expr::ScalarExp<float>, float>>(mshadow::gpu, unsigned int, 
mshadow::Shape<int=2>, int=2, int)
                    0.12%  72.416us         1  72.416us  72.416us  72.416us  void 
mshadow::cuda::SoftmaxKernel<int=8, float, 
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, 
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>>(mshadow::gpu, 
int=2, unsigned int)
                    0.04%  27.614us        13  2.1240us  1.1520us  6.5280us  void 
mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, 
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, 
mshadow::expr::Plan<mshadow::expr::ScalarExp<float>, float>>(mshadow::gpu, unsigned int, 
mshadow::Shape<int=2>, int=2)
```

```
                    0.04%  23.711us         2  11.855us  2.4000us  21.311us  void
mshadow::cuda::MapPlanKernel<mshadow::sv::plusto, int=8,
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>,
mshadow::expr::Plan<mshadow::expr::Broadcast1DExp<mshadow::Tensor<mshadow::gpu, int=1,
float>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>,
int=2)
                    0.02%  11.712us        10  1.1710us     992ns  1.5680us  [CUDA memset]
                    0.01%  7.6160us         1  7.6160us  7.6160us  7.6160us  [CUDA memcpy
DtoH]
                    0.01%  4.9920us         1  4.9920us  4.9920us  4.9920us  void
mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8,
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>,
mshadow::expr::Plan<mshadow::expr::ReduceWithAxisExp<mshadow::red::maximum,
mshadow::Tensor<mshadow::gpu, int=3, float>, float, int=3, bool=1, int=2>,
float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
      API calls:   41.86%  2.98971s        22  135.90ms  13.172us  1.56053s
cudaStreamCreateWithFlags
                   33.80%  2.41345s        22  109.70ms  95.187us  2.40157s  cudaMemGetInfo
                   21.03%  1.50207s        18  83.449ms
841ns  399.12ms  cudaFree
                    1.13%  80.808ms       912  88.604us     305ns  23.436ms
cudaFuncSetAttribute
                    0.62%  43.958ms       216  203.51us     855ns  25.633ms
cudaEventCreateWithFlags
                    0.56%  39.842ms         6  6.6404ms  2.2730us  31.600ms
cudaDeviceSynchronize
                    0.48%  34.204ms         9  3.8005ms  27.692us  16.468ms
cudaMemcpy2DAsync
                    0.14%  9.9121ms        66  150.18us  5.5080us  2.0783ms  cudaMalloc
                    0.11%  8.1087ms         4  2.0272ms  499.17us  2.6772ms
cudaGetDeviceProperties
                    0.09%  6.1239ms        12  510.33us  6.6460us  5.1190ms  cudaMemcpy
                    0.07%  4.7508ms        29  163.82us  2.1090us  2.2024ms
cudaStreamSynchronize
                    0.04%  2.6436ms       375  7.0490us     272ns  334.18us
cuDeviceGetAttribute
                    0.01%  909.36us         4  227.34us  45.866us  691.55us
cuDeviceGetName
                    0.01%  887.76us         8  110.97us  13.114us  688.13us
cudaStreamCreateWithPriority
                    0.01%  726.15us         2  363.08us  52.207us  673.94us  cudaHostAlloc
                    0.01%  713.23us        10  71.323us  7.5440us  480.63us
cudaMemsetAsync
                    0.01%  646.22us         4  161.55us  92.941us  275.56us
cuDeviceTotalMem
                    0.01%  617.82us         4  154.46us  74.072us  259.14us
cudaStreamCreate
                    0.01%  540.37us        27  20.013us  8.2670us  52.173us
cudaLaunchKernel
                    0.00%  292.81us       202  1.4490us     545ns  4.5470us
cudaDeviceGetAttribute
                    0.00%  150.39us        29  5.1850us  1.0190us  16.214us  cudaSetDevice
                    0.00%  133.47us         6  22.244us  1.1630us  86.376us
cudaEventCreate
```

```
                    0.00%  113.29us       557     203ns      75ns     812ns
cudaGetLastError
                    0.00%  44.747us        18  2.4850us     581ns  4.4910us  cudaGetDevice
                    0.00%  26.923us         2  13.461us  5.0500us  21.873us
cudaHostGetDevicePointer
                    0.00%  16.492us         4  4.1230us  1.7760us  7.1760us
cudaEventRecord
                    0.00%  7.4790us         2  3.7390us  2.7320us  4.7470us  cudaEventQuery
                    0.00%  6.1870us        20     309ns     140ns     599ns
cudaPeekAtLastError
                    0.00%  5.9340us         2  2.9670us  1.7720us  4.1620us
cudaDeviceGetStreamPriorityRange
                    0.00%  5.4820us         6     913ns     441ns  1.8620us
cuDeviceGetCount
                    0.00%  5.2120us         5  1.0420us     505ns  1.5460us  cuDeviceGet
                    0.00%  4.4920us         3  1.4970us     793ns  2.6370us  cuInit
                    0.00%  3.6930us         1  3.6930us  3.6930us  3.6930us
cuDeviceGetPCIBusId
                    0.00%  2.7100us         4     677ns     328ns  1.3210us
cuDeviceGetUuid
                    0.00%  2.0830us         3     694ns     338ns  1.3080us
cuDriverGetVersion
                    0.00%  1.9460us         4     486ns     239ns     806ns
cudaGetDeviceCount
```