Nayman Leung (NAYMANL2)
Joey Bahary (JBAHARY2)
Team-name: convoluted-convolutions

ECE 408 Final Project Milestone 1

DELIVERABLES:

1. List of all kernels that collectively consume more than 90% of program time
    a. CUDA memcpy HtoD (39.39%)
    b. void cudnn::detail::implicit_convolve_sgemm (20.65%)
    c. Volta_cgemm_64x32_tn (12.11%)
    d. Op_generic_tensor_kernel (7.15%)
    e. Fft2d_c2r_32x32 (5.74%)
    f. Volta_sgemm_128x128_tn (5.72%)
2. List of CUDA API that consume more than 90% of program time
    a. cudaStreamCreateWithFlags (42.32%)
    b. cudaMemGetInfo (33.58%)
    c. cudaFree (21.37%)
3. Explanation of difference between kernels and API calls
    a. CUDA kernels are essentially C functions defined by the user that are executed by threads on the GPU. CUDA API calls extend functionality through the runtime and Driver APIs which also hold the context. The context holds all of the management data to control and use the device (allocated memory, loaded modules that contain device code, mapping between CPU and GPU memory, etc). (https://stackoverflow.com/questions/43244645/what-is-a-cuda-context)
4. Output of RAI running on MXNet on the CPU (time m1.1.py)

```
EvalMetric: {'accuracy': 0.8236}
8.83user 3.76system 0:05.01elapsed 251%CPU (0avgtext+0avgdata
2470596maxresident)k
0inputs+2824outputs (0major+667706minor)pagefaults 0swaps
```

5. List Program Run time
    a. 5.01 seconds
6. Output of RAI running on MXNet on the GPU

```
EvalMetric: {'accuracy': 0.8236}
4.28user 3.32system 0:04.32elapsed 176%CPU (0avgtext+0avgdata
2843476maxresident)k
8inputs+4552outputs (0major+660709minor)pagefaults 0swaps
```

7. List Program Run time
    a. 4.32 seconds

Example NVPROF output

```
==383== NVPROF is profiling process 383, command: python m1.2.py
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
==383== Profiling application: python m1.2.py
==383== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   39.39%  16.127ms        20  806.35us  1.0880us  15.480ms  [CUDA memcpy HtoD]
                   20.65%  8.4531ms         1  8.4531ms  8.4531ms  8.4531ms  void cudnn::detail::implicit_convolve_sgemm<float, float, int=1024, int=5, int=5, int=3, int=3, int=3, int=1, bool=1, bool=0, bool=1>(int, int, int, float const *, int, float*, cudnn::detail::implicit_convolve_sgemm<float, float, int=1024, int=5, int=5, int=3, int=3, int=3, int=1, bool=1, bool=0, bool=1>*, kernel_conv_params, int, float, float, int, float, float, int, int)
                   12.11%  4.9587ms         1  4.9587ms  4.9587ms  4.9587ms  volta_cgemm_64x32_tn
                    7.15%  2.9281ms         2  1.4641ms  24.864us  2.9033ms  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)
                    5.74%  2.3506ms         1  2.3506ms  2.3506ms  2.3506ms  void fft2d_c2r_32x32<float, bool=0, bool=0, unsigned int=1, bool=0, bool=0>(float*, float2 const *, int, int, int, int, int, int, int, int, int, float, float, cudnn::reduced_divisor, bool, float*, float*, int2, int, int)
                    5.72%  2.3400ms         1  2.3400ms  2.3400ms  2.3400ms  volta_sgemm_128x128_tn
                    4.60%  1.8821ms         1  1.8821ms  1.8821ms  1.8821ms  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)
                    3.80%  1.5541ms         1  1.5541ms  1.5541ms  1.5541ms  void fft2d_r2c_32x32<float, bool=0, unsigned int=0, bool=0>(float2*, float const *, int, int, int, int, int, int, int, int, cudnn::reduced_divisor, bool, int2, int, int)
                    0.37%  152.42us         1  152.42us  152.42us  152.42us  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)
                    0.18%  75.072us         1  75.072us  75.072us  75.072us  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)
0.07% 30.144us 13 2.3180us 1.2160us 7.5200us 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)
0.06% 25.440us 1 25.440us 25.440us 25.440us volta_sgemm_32x128_tn
0.06% 23.776us 2 11.888us 2.5920us 21.184us 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)
0.04% 15.968us 1 15.968us 15.968us 15.968us void fft2d_r2c_32x32<float, bool=0, unsigned int=1, bool=0>(float2*, float const *, int, int, int, int, int, int, int, int, cudnn::reduced_divisor, bool, int2, int, int)
0.02% 10.016us 9 1.1120us 992ns 1.5360us [CUDA memset]
0.02% 7.3280us 1 7.3280us 7.3280us 7.3280us [CUDA memcpy DtoH]
0.01% 4.8000us 1 4.8000us 4.8000us 4.8000us 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)
0.01% 3.2640us 1 3.2640us 3.2640us 3.2640us void flip_filter<float, float>(float*, float const *, int, int, int, int)
0.01% 2.6550us 1 2.6550us 2.6550us 2.6550us compute_gemm_pointers(float2**, float2 const *, int, float2 const *, int, float2 const *, int, int)
API calls: 42.32% 3.00705s 22 136.68ms 12.851us 1.56016s cudaStreamCreateWithFlags
33.58% 2.38633s 24 99.430ms 104.11us 2.38104s cudaMemGetInfo
21.37% 1.51843s 19 79.917ms 834ns 408.31ms cudaFree
1.41% 99.952ms 912 109.60us 308ns 53.410ms cudaFuncSetAttribute
0.46% 32.478ms 9 3.6086ms 33.713us 15.513ms cudaMemcpy2DAsync
0.31% 21.688ms 29 747.87us 3.4170us 9.9361ms cudaStreamSynchronize
0.18% 12.549ms 68 184.55us 5.7180us 2.8536ms cudaMalloc
0.12% 8.4657ms 216 39.192us 889ns 5.9292ms cudaEventCreateWithFlags
0.10% 7.2073ms 6 1.2012ms 1.1090us 7.1385ms cudaEventCreate
0.07% 4.7327ms 4 1.1832ms 424.50us 1.7514ms cudaGetDeviceProperties

```
                0.03%  2.4824ms       375  6.6190us     284ns  331.03us  cuDeviceGetAttribute
                0.01%  789.41us         2  394.70us  51.193us  738.21us  cudaHostAlloc
                0.01%  621.86us        30  20.728us  7.9970us  81.572us  cudaLaunchKernel
                0.01%  610.11us         4  152.53us  94.017us  275.55us  cuDeviceTotalMem
                0.01%  599.74us         4  149.94us  77.489us  246.88us  cudaStreamCreate
                0.01%  469.55us        12  39.128us  5.9160us  88.270us  cudaMemcpy
                0.01%  389.25us         9  43.250us  9.3750us  212.87us  cudaMemsetAsync
                0.00%  323.78us       210  1.5410us     566ns  16.920us  cudaDeviceGetAttribute
                0.00%  289.34us         4  72.334us  43.955us  103.40us  cuDeviceGetName
                0.00%  172.31us         8  21.538us  13.755us  44.913us
cudaStreamCreateWithPriority
                0.00%  155.99us        32  4.8740us  1.4400us  15.018us  cudaSetDevice
                0.00%  106.65us       564     189ns      75ns     611ns  cudaGetLastError
                0.00%  43.911us        18  2.4390us     599ns  4.7600us  cudaGetDevice
                0.00%  23.685us         6  3.9470us  1.6840us  7.0150us  cudaEventRecord
                0.00%  13.089us         1  13.089us  13.089us  13.089us  cudaBindTexture
                0.00%  9.2010us         3  3.0670us  1.8970us  4.3720us  cudaStreamWaitEvent
                0.00%  7.9230us         1  7.9230us  7.9230us  7.9230us  cuDeviceGetPCIBusId
                0.00%  7.0690us         2  3.5340us  2.3280us  4.7410us
cudaHostGetDevicePointer
                0.00%  6.1030us         6  1.0170us     401ns  2.3180us  cuDeviceGetCount
                0.00%  6.0000us         2  3.0000us  1.5100us  4.4900us
cudaDeviceGetStreamPriorityRange
                0.00%  5.2940us        18     294ns     121ns     673ns  cudaPeekAtLastError
                0.00%  4.7520us         5     950ns     474ns  1.7100us  cuDeviceGet
                0.00%  4.1730us         3  1.3910us     809ns  2.2560us  cuInit
                0.00%  3.8930us         1  3.8930us  3.8930us  3
.8930us  cudaEventQuery
                0.00%  3.3850us         1  3.3850us  3.3850us  3.3850us  cudaUnbindTexture
                0.00%  2.4530us         4     613ns     354ns  1.2000us  cuDeviceGetUuid
                0.00%  1.9340us         3     644ns     330ns  1.1950us  cuDriverGetVersion
                0.00%  1.7790us         4     444ns     262ns     777ns  cudaGetDeviceCount
```

Milestone 2

DELIVERABLES:

- List whole program run times
- List Op. times

Run #1: 100

```
*Running /usr/bin/time python m2.1.py 100
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.034078
Op Time: 0.074938
Correctness: 0.84 Model: ece408
```

Run #2: 1,000

```
*Running /usr/bin/time python m2.1.py 1000
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.243053
Op Time: 0.741502
Correctness: 0.852 Model: ece408
4.40user 2.85system 0:01.99elapsed 363%CPU (0avgtext+0avgdata 332360maxresident)
k
0inputs+2824outputs (0major+110723minor)pagefaults 0swaps
```

Default: 10,000

```
Op Time: 2.437733
Op Time: 7.488936
Correctness: 0.8397 Model: ece408
15.27user 4.59system 0:11.51elapsed 172%CPU (0avgtext+0avgdata 1617608maxresident)k
```

Milestone 3

Run #1: 100

```
*Running /usr/bin/time python m3.1.py 100
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.000075
Op Time: 0.000213
Correctness: 0.84 Model: ece408
4.16user 3.50system 0:04.18elapsed 183%CPU (0avgtext+0avgdata 2784952maxresident
)k
8inputs+2800outputs (
0major+624565minor)pagefaults 0swaps
```

Run #2: 1,000

```
*Running /usr/bin/time python m3.1.py 1000
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.000611
Op Time: 0.002004
Correctness: 0.852 Model: ece408
4.31user 3.21system 0:04.15elapsed 181%CPU (0avgtext+0avgdata 2776192maxresident
)k
0inputs+4576outputs (0major+623696minor)pag
efaults 0swaps
```

Default: 10,000

```
*Running /usr/bin/time python m3.1.py 10000
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.006043
Op Time: 0.021991
Correctness: 0.8397 Model: ece408
4.41user 3.48system 0:04.34elapsed 181%CPU (0avgtext+0avgdata 2844976maxr
esident)k
0inputs+4576outputs (0major+663183minor)pagefaults 0swaps
```