ECE408 Project Milestone 3 Report

Shuyue Lai (shuyuel2), Yaxin Peng (yaxinp2), Jingyuan Zhang (jz61)

Team: tensor

UIUC on Campus Students

Deliverable 1: Correctness and timing with 3 different dataset sizes

✱ **Running python m3.1.py 100**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.000295
Op Time: 0.000998
Correctness: 0.76 Model: ece408
```

✱ **Running python m3.1.py 1000**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.003069
Op Time: 0.010254
Correctness: 0.767 Model: ece408
```

✱ **Running python m3.1.py 10000**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.031567
Op Time: 0.098181
Correctness: 0.7653 Model: ece408
```

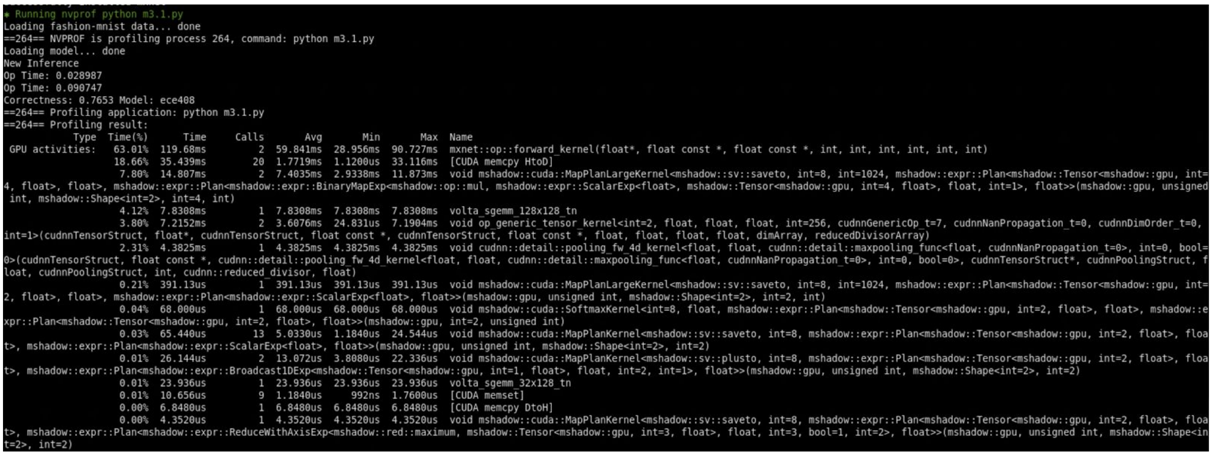
Deliverable 2: Demonstrate nvprof profiling the execution

GPU activities:

1. **Time(%):** 63.01% **Time:**119.68ms **Calls:**2
   Acg:59.841ms **Min:**28.956ms **Max:**90.727ms
   Name:mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int, int, int, int)
2. **Time(%):** 18.66% **Time:**35.439ms **Calls:**20
   Acg:1.7719ms **Min:**1.1200us **Max:**33.116ms
   Name:[CUDA memcpy HtoD]
3. **Time(%):** 7.80% **Time:**14.807ms **Calls:**2
   Acg:7.4035ms **Min:**2.9338ms **Max:**11.873ms
   Name: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)
4. **Time(%):** 4.12% **Time:**7.8308ms **Calls:**1
   Avg:7.8308ms **Min:**7.8308ms **Max:**7.8308ms
   Name:volta_sgemm_128x128_tn
5. **Time(%):** 3.80% **Time:**7.2152ms **Calls:**2
   Avg:3.6076ms **Min:**24.831us **Max:**7.1904ms
   Name: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)
6. **Time(%):** 2.31% **Time:**4.3825ms **Calls:**1
   Avg:4.3825ms **Min:**4.3825ms **Max:**4.3825ms
   Name: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)
7. **Time(%):** 0.21% **Time:**391.13us **Calls:**1
   Avg:391.13us **Min:**391.13us **Max:**391.13us
   Name: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)
8. **Time(%):** 0.04% **Time:**68.000us **Calls:**1
   Avg:68.000us **Min:**68.000us **Max:**68.000us

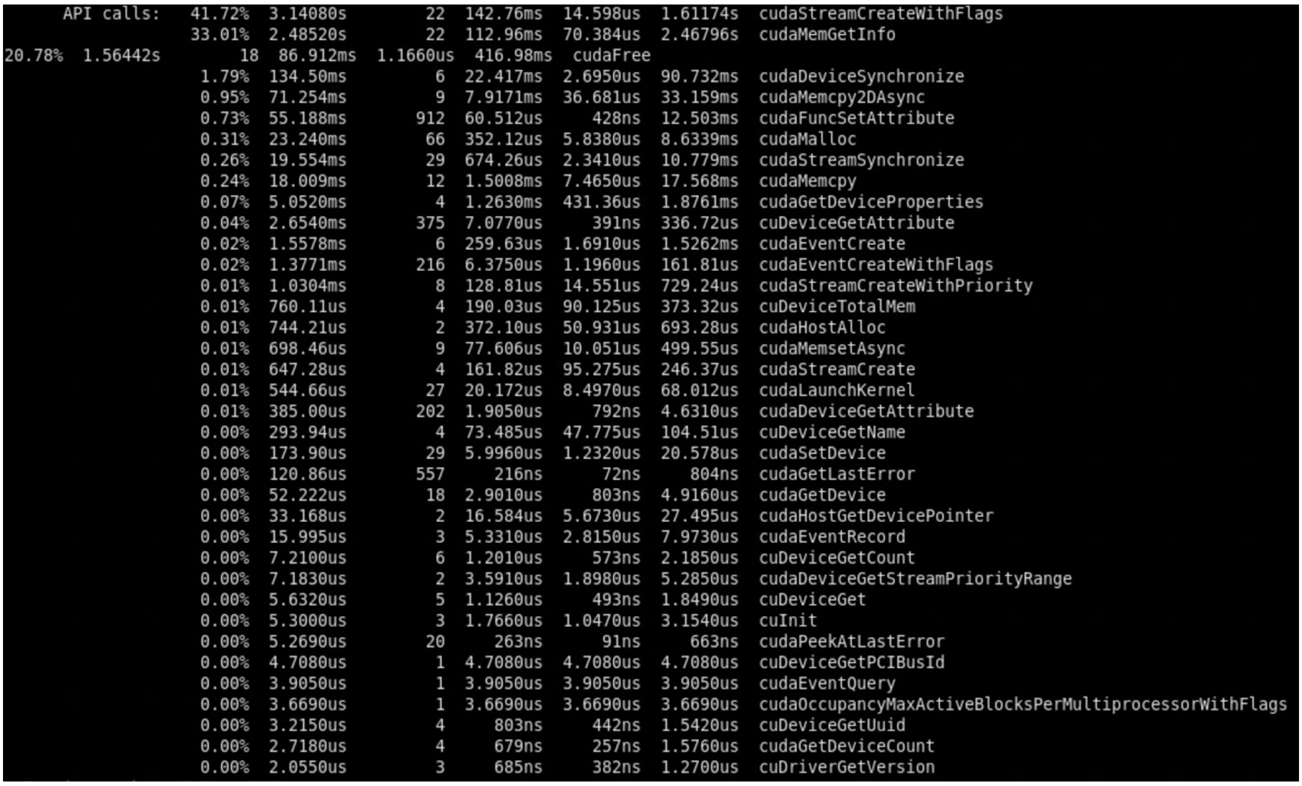
Name:void mshadow::cuda::SoftmaxKernel<int=8, float, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>>(mshadow::gpu, int=2, unsigned int)

9. **Time(%):** 0.03% **Time:**65.440us **Calls:**13
   Avg:5.0330us **Min:**1.1840us **Max:**24.544us
   Name: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)
10. **Time(%):** 0.01% **Time:**26.144us **Calls:**2
    Avg:13.072us **Min:**3.8080us **Max:**22.336us
    Name: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)
11. **Time(%):** 0.01% **Time:**23.936us **Calls:**1
    Avg:23.936us **Min:**23.936us **Max:**23.936us
    Name:volta_sgemm_32x128_tn
12. **Time(%):** 0.01% **Time:**10.656us **Calls:**9
    Avg:1.1840us **Min:**992ns **Max:**1.7600us
    Name:[CUDA memset]
13. **Time(%):** 0.00% **Time:**6.8480us **Calls:**1
    Avg:6.8480us **Min:**6.8480us **Max:**6.8480us
    Name:[CUDA memcpy DtoH]
14. **Time(%):** 0.00% **Time:**4.3520us **Calls:**1
    Acg:4.3520us **Min:**4.3520us **Max:**4.3520us
    Name: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)



```
* Running nvprof python m3.1.py
Loading fashion-mnist data... done
==264== NVPROF is profiling process 264, command: python m3.1.py
Loading model... done
New Inference
Op Time: 0.028987
Op Time: 0.090747
Correctness: 0.7653 Model: ece408
==264== Profiling application: python m3.1.py
==264== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   63.01%  119.68ms         2  59.841ms  28.956ms  90.727ms  mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int, int, int, int)
                   18.66%  35.439ms        20  1.7719ms  1.1200us  33.116ms  [CUDA memcpy HtoD]
                    7.80%  14.807ms         2  7.4035ms  2.9338ms  11.873ms  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)
                    4.12%  7.8308ms         1  7.8308ms  7.8308ms  7.8308ms  volta_sgemm_128x128_tn
                    3.80%  7.2152ms         2  3.6076ms  24.831us  7.1904ms  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)
                    2.31%  4.3825ms         1  4.3825ms  4.3825ms  4.3825ms  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, f
loat, cudnnPoolingStruct, int, cudnn::reduced_divisor, float)
                    0.21%  391.13us         1  391.13us  391.13us  391.13us  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)
                    0.04%  68.000us         1  68.000us  68.000us  68.000us  void mshadow::cuda::SoftmaxKernel<int=8, float, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::e
xpr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>>(mshadow::gpu, int=2, unsigned int)
                    0.03%  65.440us        13  5.0330us  1.1840us  24.544us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, floa
t>, mshadow::expr::Plan<mshadow::expr::ScalarExp<float>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
                    0.01%  26.144us         2  13.072us  3.8080us  22.336us  void mshadow::cuda::MapPlanKernel<mshadow::sv::plusto, int=8, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, floa
t>, 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)
                    0.01%  23.936us         1  23.936us  23.936us  23.936us  volta_sgemm_32x128_tn
                    0.01%  10.656us         9  1.1840us     992ns  1.7600us  [CUDA memset]
                    0.00%  6.8480us         1  6.8480us  6.8480us  6.8480us  [CUDA memcpy DtoH]
                    0.00%  4.3520us         1  4.3520us  4.3520us  4.3520us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, floa
t>, 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<in
t=2>, int=2)
```

API calls:



```
      API calls:   41.72%  3.14080s        22  142.76ms  14.598us  1.61174s  cudaStreamCreateWithFlags
                   33.01%  2.48520s        22  112.96ms  70.384us  2.46796s  cudaMemGetInfo
20.78%  1.56442s        18  86.912ms  1.1660us  416.98ms  cudaFree
                    1.79%  134.50ms         6  22.417ms  2.6950us  90.732ms  cudaDeviceSynchronize
                    0.95%  71.254ms         9  7.9171ms  36.681us  33.159ms  cudaMemcpy2DAsync
                    0.73%  55.188ms       912  60.512us     428ns  12.503ms  cudaFuncSetAttribute
                    0.31%  23.240ms        66  352.12us  5.8380us  8.6339ms  cudaMalloc
                    0.26%  19.554ms        29  674.26us  2.3410us  10.779ms  cudaStreamSynchronize
                    0.24%  18.009ms        12  1.5008ms  7.4650us  17.568ms  cudaMemcpy
                    0.07%  5.0520ms         4  1.2630ms  431.36us  1.8761ms  cudaGetDeviceProperties
                    0.04%  2.6540ms       375  7.0770us     391ns  336.72us  cuDeviceGetAttribute
                    0.02%  1.5578ms         6  259.63us  1.6910us  1.5262ms  cudaEventCreate
                    0.02%  1.3771ms       216  6.3750us  1.1960us  161.81us  cudaEventCreateWithFlags
                    0.01%  1.0304ms         8  128.81us  14.551us  729.24us  cudaStreamCreateWithPriority
                    0.01%  760.11us         4  190.03us  90.125us  373.32us  cuDeviceTotalMem
                    0.01%  744.21us         2  372.10us  50.931us  693.28us  cudaHostAlloc
                    0.01%  698.46us         9  77.606us  10.051us  499.55us  cudaMemsetAsync
                    0.01%  647.28us         4  161.82us  95.275us  246.37us  cudaStreamCreate
                    0.01%  544.66us        27  20.172us  8.4970us  68.012us  cudaLaunchKernel
                    0.01%  385.00us       202  1.9050us     792ns  4.6310us  cudaDeviceGetAttribute
                    0.00%  293.94us         4  73.485us  47.775us  104.51us  cuDeviceGetName
                    0.00%  173.90us        29  5.9960us  1.2320us  20.578us  cudaSetDevice
                    0.00%  120.86us       557     216ns      72ns     804ns  cudaGetLastError
                    0.00%  52.222us        18  2.9010us     803ns  4.9160us  cudaGetDevice
                    0.00%  33.168us         2  16.584us  5.6730us  27.495us  cudaHostGetDevicePointer
                    0.00%  15.995us         3  5.3310us  2.8150us  7.9730us  cudaEventRecord
                    0.00%  7.2100us         6  1.2010us     573ns  2.1850us  cuDeviceGetCount
                    0.00%  7.1830us         2  3.5910us  1.8980us  5.2850us  cudaDeviceGetStreamPriorityRange
                    0.00%  5.6320us         5  1.1260us     493ns  1.8490us  cuDeviceGet
                    0.00%  5.3000us         3  1.7660us  1.0470us  3.1540us  cuInit
                    0.00%  5.2690us        20     263ns      91ns     663ns  cudaPeekAtLastError
                    0.00%  4.7080us         1  4.7080us  4.7080us  4.7080us  cuDeviceGetPCIBusId
                    0.00%  3.9050us         1  3.9050us  3.9050us  3.9050us  cudaEventQuery
                    0.00%  3.6690us         1  3.6690us  3.6690us  3.6690us  cudaOccupancyMaxActiveBlocksPerMultiprocessorWithFlags
                    0.00%  3.2150us         4     803ns     442ns  1.5420us  cuDeviceGetUuid
                    0.00%  2.7180us         4     679ns     257ns  1.5760us  cudaGetDeviceCount
                    0.00%  2.0550us         3     685ns     382ns  1.2700us  cuDriverGetVersion
```