ECE408: Applied Parallel Programming **Fall 2017**

Project Report

Yiming Wang(teamname: q80) *Handed In: November 17, 2017*

1. **Milestone 1**

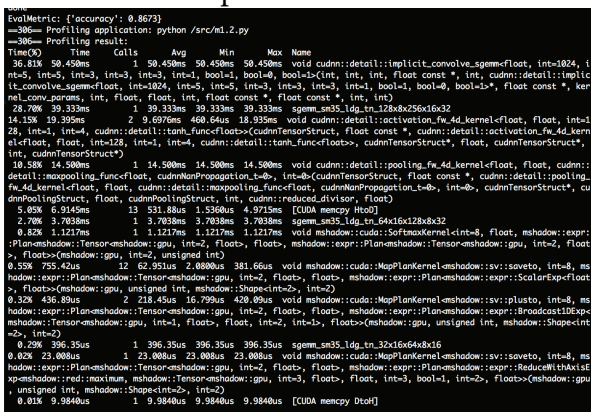
   (a) The time elapsed is 16.373s

```
Loading fashion-mnist data... done
Loading model... done
EvalMetric: {'accuracy': 0.8673}
* The build folder has been uploaded
t.com/userdata/build-02e09e7f-715c-48
be present for only a short duration
* Server has ended your request.

real    0m16.373s
user    0m0.386s
sys     0m0.123s
```

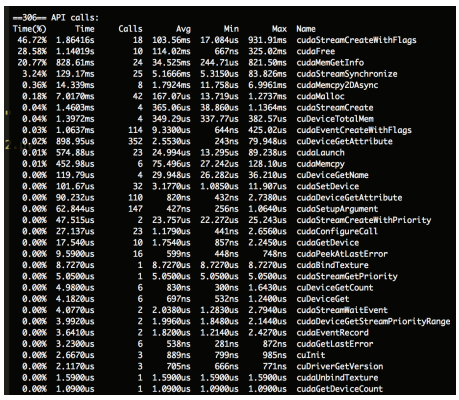
   (b) The time elapsed is 44.267s

```
done
EvalMetric: {'accuracy': 0.8673}
* The build folder has been uploaded
t.com/userdata/build-94bcc972-3cad-4
be present for only a short duration
* Server has ended your request.

real    0m44.267s
user    0m0.378s
sys     0m0.093s
```

   (c) The result for nvprof is shown below



```
EvalMetric: {'accuracy': 0.8673}
==306== Profiling application: python /src/m1.2.py
==306== Profiling result:
Time(%)      Time     Calls       Avg       Min       Max  Name
 36.81%  50.450ms         1  50.450ms  50.450ms  50.450ms  void cudnn::detail::implicit_convolve_sgemm<float, int=1024, i
nt=5, int=5, int=3, int=3, int=3, int=1, bool=1, bool=0, bool=1>(int, int, int, float const *, int, cudnn::detail::implic
it_convolve_sgemm<float, int=1024, int=5, int=5, int=3, int=3, int=3, int=1, bool=1, bool=0, bool=1>*, float const *, ker
nel_conv_params, int, float, float, int, float const *, int, int)
 28.70%  39.333ms         1  39.333ms  39.333ms  39.333ms  sgemm_sm35_ldg_tn_128x8x256x16x32
14.15%  19.395ms         2  9.6976ms  460.64us  18.935ms  void cudnn::detail::activation_fw_4d_kernel<float, float, int=1
28, int=1, int=4, cudnn::detail::tanh_func<float>>(cudnnTensorStruct, float const *, cudnn::detail::activation_fw_4d_kern
el<float, float, int=128, int=1, int=4, cudnn::detail::tanh_func<float>>, cudnnTensorStruct*, float, cudnnTensorStruct*,
int, cudnnTensorStruct*)
 10.58%  14.500ms         1  14.500ms  14.500ms  14.500ms  void cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::
detail::maxpooling_func<float, cudnnNanPropagation_t=0>, int=0>(cudnnTensorStruct, float const *, cudnn::detail::pooling_
fw_4d_kernel<float, float, cudnn::detail::maxpooling_func<float, cudnnNanPropagation_t=0>, int=0>, cudnnTensorStruct*, cu
dnnPoolingStruct, float, cudnnPoolingStruct, int, cudnn::reduced_divisor, float)
  5.05%  6.9145ms        13  531.88us  1.5360us  4.9715ms  [CUDA memcpy HtoD]
  2.70%  3.7038ms         1  3.7038ms  3.7038ms  3.7038ms  sgemm_sm35_ldg_tn_64x16x128x8x32
  0.82%  1.1217ms         1  1.1217ms  1.1217ms  1.1217ms  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)
0.55%  755.42us        12  62.951us  2.0800us  381.66us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, ms
hadow::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)
0.32%  436.89us         2  218.45us  16.799us  420.09us  void mshadow::cuda::MapPlanKernel<mshadow::sv::plusto, int=8, ms
hadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::BroadcastiDExp<
mshadow::Tensor<mshadow::gpu, int=1, float>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int
=2>, int=2)
  0.29%  396.35us         1  396.35us  396.35us  396.35us  sgemm_sm35_ldg_tn_32x16x64x8x16
0.02%  23.008us         1  23.008us  23.008us  23.008us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, ms
hadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::ReduceWithAxisE
xp<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%  9.9840us         1  9.9840us  9.9840us  9.9840us  [CUDA memcpy DtoH]
```



```
==306== API calls:
Time(%)      Time     Calls       Avg       Min       Max  Name
 46.72%  1.86416s        18  103.56ms  17.084us  931.91ms  cudaStreamCreateWithFlags
 28.58%  1.14019s        10  114.02ms     667ns  325.02ms  cudaFree
 20.77%  828.61ms        24  34.525ms  244.71us  821.50ms  cudaMemGetInfo
  3.24%  129.17ms        25  5.1666ms  5.3150us  83.826ms  cudaStreamSynchronize
  0.36%  14.339ms         8  1.7924ms  11.758us  6.9961ms  cudaMemcpy2DAsync
  0.18%  7.0170ms        42  167.07us  13.719us  1.2737ms  cudaMalloc
  0.04%  1.4603ms         4  365.06us  38.860us  1.1364ms  cudaStreamCreate
  0.04%  1.3972ms         4  349.29us  337.77us  382.57us  cuDeviceTotalMem
  0.03%  1.0637ms       114  9.3300us     644ns  425.02us  cudaEventCreateWithFlags
  0.02%  898.95us       352  2.5530us     243ns  79.948us  cuDeviceGetAttribute
  0.01%  574.88us        23  24.994us  13.295us  89.238us  cudaLaunch
  0.01%  452.98us         6  75.496us  27.242us  128.10us  cudaMemcpy
  0.00%  119.79us         4  29.948us  26.282us  36.210us  cuDeviceGetName
  0.00%  101.67us        32  3.1770us  1.0850us  11.907us  cudaSetDevice
  0.00%  90.232us       110     820ns     432ns  2.7380us  cudaDeviceGetAttribute
  0.00%  62.844us       147     427ns     256ns  1.0640us  cudaSetupArgument
  0.00%  47.515us         2  23.757us  22.272us  25.243us  cudaStreamCreateWithPriority
  0.00%  27.137us        23  1.1790us     441ns  2.6560us  cudaConfigureCall
  0.00%  17.540us        10  1.7540us     857ns  2.2450us  cudaGetDevice
  0.00%  9.5900us        16     599ns     448ns     748ns  cudaPeekAtLastError
  0.00%  8.7270us         1  8.7270us  8.7270us  8.7270us  cudaBindTexture
  0.00%  5.0500us         1  5.0500us  5.0500us  5.0500us  cudaStreamGetPriority
  0.00%  4.9800us         6     830ns     300ns  1.6430us  cuDeviceGetCount
  0.00%  4.1820us         6     697ns     532ns  1.2400us  cuDeviceGet
  0.00%  4.0770us         2  2.0380us  1.2830us  2.7940us  cudaStreamWaitEvent
  0.00%  3.9920us         2  1.9960us  1.8480us  2.1440us  cudaDeviceGetStreamPriorityRange
  0.00%  3.6410us         2  1.8200us  1.2140us  2.4270us  cudaEventRecord
  0.00%  3.2300us         6     538ns     281ns     872ns  cudaGetLastError
  0.00%  2.6670us         3     889ns     799ns     985ns  cuInit
  0.00%  2.1170us         3     705ns     666ns     771ns  cuDriverGetVersion
  0.00%  1.5900us         1  1.5900us  1.5900us  1.5900us  cudaUnbindTexture
  0.00%  1.0900us         1  1.0900us  1.0900us  1.0900us  cudaGetDeviceCount
```

   We can see from the result that the most time consuming kernel in profile part in *implicit_convolve_sgemm*, *sgemm_sm35_ldg_tn_128x8x256x16x32* and *activation_fw_4d_kernel*. For the API call, the most consuming kernel is *cudaStreamCreateWithFlags*, *cudaFree* and *cudaMemGetInfo*.

2. **Milestone 2**

The result is shown below.