M1.1 Milestone 1.1

```
* Running /usr/bin/time python m1.1.py
New Inference
Loading fashion-mnist data... done
Loading model... done
EvalMetric: {'accuracy': 0.8673}
8.38user 3.77system 0:06.77elapsed 179%CPU (0avgtext+0avgdata 2799016m
axresident)k
106936inputs+2624outputs (265major+34986minor)pagefaults 0swaps
```

M1.2 Milestone 1.2

```
* Running /usr/bin/time python m1.2.py
New Inference
Loading fashion-mnist data... done
Loading model...[03:58:20] src/operator/././cudnn_algoreg-inl.h:112: R
unning performance tests to find the best convolution algorithm, this
can take a while... (setting env variable MXNET_CUDNN_AUTOTUNE_DEFAULT
 to 0 to disable)
done
EvalMetric: {'accuracy': 0.8673}
1.80user 1.00system 0:04.05elapsed 69%CPU (0avgtext+0avgdata 910152max
resident)k
301840inputs+3136outputs (1194major+156674minor)pagefaults 0swaps
```

M1.3 Milestone 1.3

most time spent
cudnn::detail::implicit_convolve_sgemm
sgemm_sm35_ldg_tn_128x8x256x16x32
cudnn::detail::activation_fw_4d_kernel

```
* Running nvprof python m1.2.py
New Inference
Loading fashion-mnist data... done
==311== NVPROF is profiling process 311, command: python m1.2.py
Loading model...[04:18:40] 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_DEFAULT to 0 to disable)
done
EvalMetric: {'accuracy': 0.8673}
==311== Profiling application: python m1.2.py
==311== Profiling result:
Time(%)   Time   Calls    Avg     Min     Max Name
37.00% 49.994ms    1 49.994ms 49.994ms 49.994ms void cudnn::detail::implicit_convolve_sgemm<
    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, cudnn::detail::implicit_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 *, kernel_conv_params, int
    , float, float, int, float const *, float const *, int, int)
 28.65% 38.718ms    1 38.718ms 38.718ms 38.718ms sgemm_sm35_ldg_tn_128x8x256x16x32
14.35% 19.395ms    2 9.6975ms 459.22us 18.936ms void cudnn::detail::activation_fw_4d_kernel<
    float, float, int=128, int=1, int=4, cudnn::detail::tanh_func<float>>(cudnnTensorStruct, float
    const *, cudnn::detail::activation_fw_4d_kernel<float, float, int=128, int=1, int=4,
cudnn::detail::tanh_func<float>>, cudnnTensorStruct*, float, cudnnTensorStruct*, int,
    cudnnTensorStruct*)
 10.70% 14.454ms    1 14.454ms 14.454ms 14.454ms 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*,
     cudnnPoolingStruct, float, cudnnPoolingStruct, int, cudnn::reduced_divisor, float)
  4.62% 6.2386ms   13 479.89us 1.5360us 4.3049ms [CUDA memcpy HtoD]
  2.67% 3.6088ms    1 3.6088ms 3.6088ms 3.6088ms sgemm_sm35_ldg_tn_64x16x128x8x32
0.82% 1.1129ms    1 1.1129ms 1.1129ms 1.1129ms 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% 748.53us   12 62.377us 2.1440us 377.94us 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)
0.32% 433.56us    2 216.78us 16.864us 416.70us 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)
  0.29% 392.86us    1 392.86us 392.86us 392.86us sgemm_sm35_ldg_tn_32x16x64x8x16
0.02% 23.296us    1 23.296us 23.296us 23.296us 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% 9.7920us    1 9.7920us 9.7920us 9.7920us [CUDA memcpy DtoH]

==311== API calls:
Time(%)   Time   Calls    Avg     Min     Max Name
47.67% 1.95762s   18 108.76ms 18.434us 978.57ms cudaStreamCreateWithFlags
 28.13% 1.15530s   10 115.53ms 1.0540us 329.31ms cudaFree
 20.48% 840.99ms   24 35.041ms 201.88us 833.88ms cudaMemGetInfo
  3.12% 128.23ms   25 5.1290ms 5.3410us 83.284ms cudaStreamSynchronize
  0.29% 12.057ms    8 1.5071ms 12.714us 4.3691ms cudaMemcpy2DAsync
  0.17% 7.0582ms   42 168.05us 12.671us 1.2564ms cudaMalloc
  0.03% 1.3813ms    4 345.32us 339.12us 359.33us cuDeviceTotalMem
  0.03% 1.1027ms  114 9.6720us  622ns 575.44us cudaEventCreateWithFlags
  0.02% 895.80us  352 2.5440us  245ns 64.087us cuDeviceGetAttribute
  0.02% 667.36us   23 29.015us 13.729us 109.50us cudaLaunch
  0.02% 660.55us    4 165.14us 34.367us 521.49us cudaStreamCreate
  0.01% 372.36us    6 62.060us 27.144us 119.86us cudaMemcpy
  0.00% 114.46us    4 28.615us 22.175us 31.693us cuDeviceGetName
  0.00% 89.017us   32 2.7810us  936ns 6.8700us cudaSetDevice
  0.00% 71.263us  147   484ns   256ns 1.1800us cudaSetupArgument
  0.00% 70.089us  110   637ns   416ns 2.5950us cudaDeviceGetAttribute
  0.00% 51.066us    2 25.533us 19.515us 31.551us cudaStreamCreateWithPriority
  0.00% 33.654us   23 1.4630us  454ns 2.9810us cudaConfigureCall
  0.00% 17.223us   10 1.7220us 1.2930us 2.0230us cudaGetDevice
  0.00% 12.472us    1 12.472us 12.472us 12.472us cudaBindTexture
  0.00% 11.077us   16   692ns   468ns   934ns cudaPeekAtLastError
  0.00% 7.0290us    1 7.0290us 7.0290us 7.0290us cudaStreamGetPriority
  0.00% 4.4450us    6   740ns   259ns 1.5510us cuDeviceGetCount
  0.00% 4.2550us    2 2.1270us 1.4500us 2.8050us cudaStreamWaitEvent
  0.00% 3.9800us    2 1.9900us 1.8400us 2.1400us cudaDeviceGetStreamPriorityRange
  0.00% 3.7880us    2 1.8940us 1.3230us 2.4650us cudaEventRecord
  0.00% 3.6990us    6   616ns   476ns 1.0270us cudaGetLastError
  0.00% 3.2250us    6   537ns   342ns   780ns cuDeviceGet
  0.00% 3.1840us    1 3.1840us 3.1840us 3.1840us cudaUnbindTexture
0.00% 2.8110us    3   937ns   908ns   969ns cuInit
```

```
 0.00% 2.1660us    3    722ns  671ns   780ns cuDriverGetVersion
 0.00% 1.1840us    1 1.1840us 1.1840us 1.1840us cudaGetDeviceCount

* Running python /src/m2.1.py
Loading fashion-mnist data... done
Loading model... done
Op Time: 12.146829
Correctness: 0.8562 Model: ece408-high
```