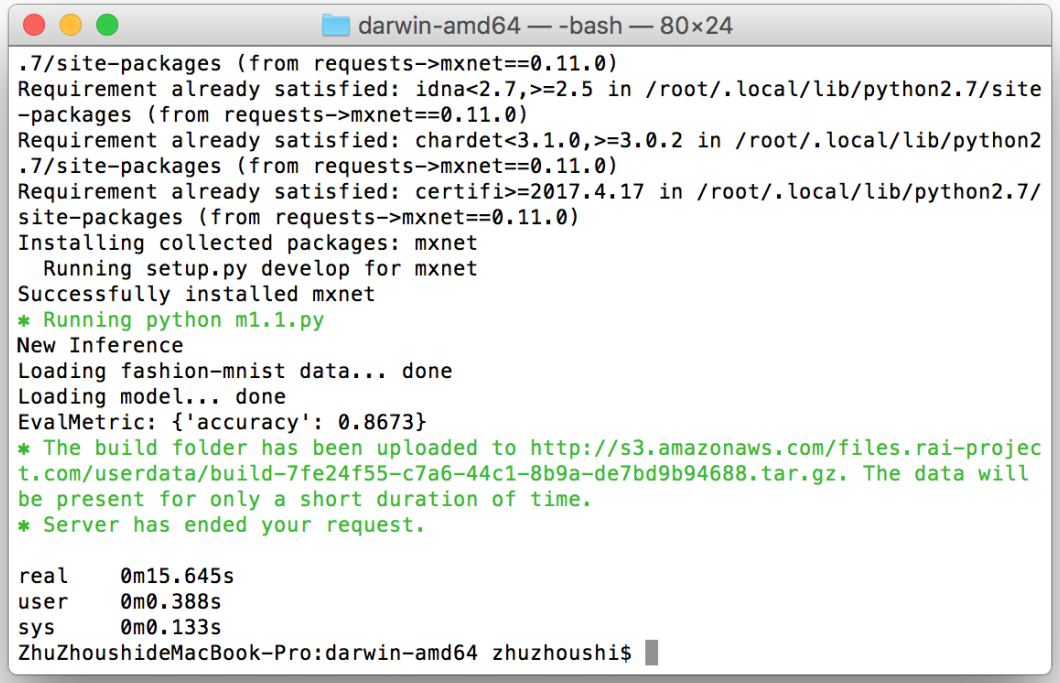
ECE408 Final Project Report

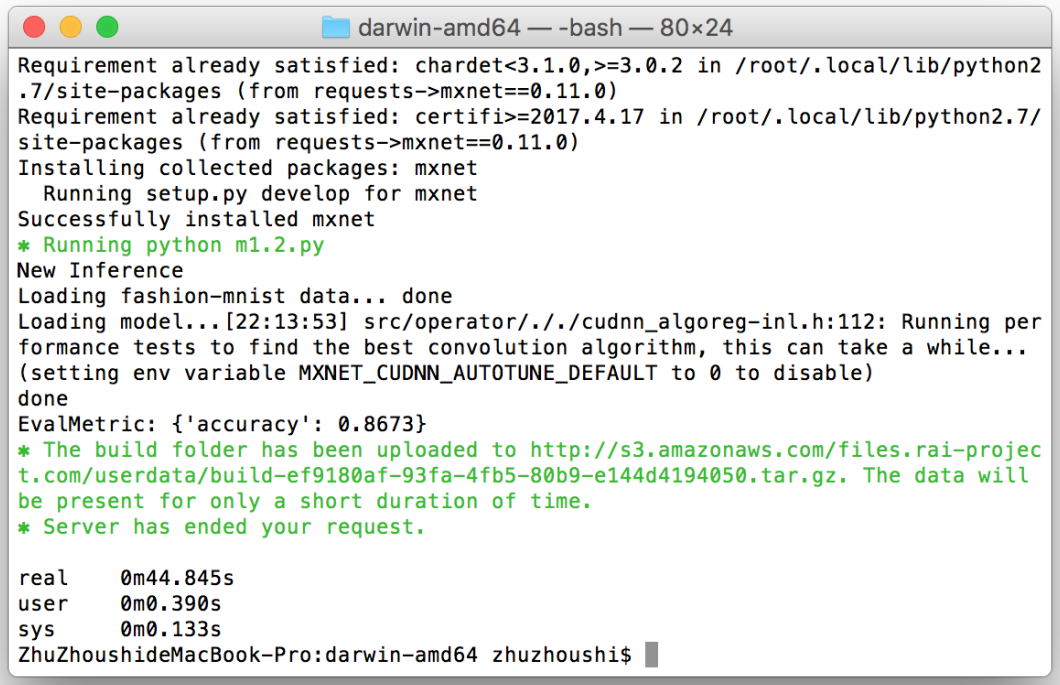
Minghao Jiang/Zhoushi Zhu
mjiang24/zzhu31

1. Baseline Result

M1.1 Results - time



M1.2 / M1.3 Result - time



M1.2 / M1.3 Result - NVPROF

a. ✱ Running nvprof python m1.2.py
b. New Inference
c. Loading fashion-mnist data... done
d. ==309== NVPROF is profiling process 309, command: python m1.2.py
e. Loading model...[22:15:59] 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)
f. done
g. EvalMetric: {'accuracy': 0.8673}
h. ==309== Profiling application: python m1.2.py
i. ==309== Profiling result:
j. Time(%) Time Calls Avg Min Max Name
k. 36.49% 49.292ms 1 49.292ms 49.292ms 49.292ms 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)

l. 28.21% 38.116ms 1 38.116ms 38.116ms 38.116ms sgemm_sm35_ldg_tn_128x8x256x16x32
m. 14.34% 19.371ms 2 9.6853ms 454.63us 18.916ms 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*)
n. 10.65% 14.394ms 1 14.394ms 14.394ms 14.394ms 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)
o. 5.61% 7.5849ms 13 583.46us 1.6000us 5.5487ms [CUDA memcpy HtoD]
p. 2.70% 3.6527ms 1 3.6527ms 3.6527ms 3.6527ms sgemm_sm35_ldg_tn_64x16x128x8x32
q. 0.82% 1.1016ms 1 1.1016ms 1.1016ms 1.1016ms 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)
r. 0.55% 738.14us 12 61.511us 2.0800us 372.52us 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)
s. 0.32% 430.22us 2 215.11us 16.959us 413.26us 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)
t. 0.29% 385.74us 1 385.74us 385.74us 385.74us sgemm_sm35_ldg_tn_32x16x64x8x16
u. 0.02% 22.079us 1 22.079us 22.079us 22.079us 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)

v. 0.01% 9.9840us 1 9.9840us 9.9840us 9.9840us [CUDA memcpy DtoH]

w.

x. ==309== API calls:

y. Time(%) Time Calls Avg Min Max Name

z. 46.41% 1.96308s 18 109.06ms 17.990us 981.19ms cudaStreamCreateWithFlags

aa. 28.81% 1.21868s 10 121.87ms 702ns 345.09ms cudaFree

bb. 21.11% 893.00ms 24 37.208ms 216.58us 885.86ms cudaMemGetInfo

cc. 3.00% 127.01ms 25 5.0805ms 5.1280us 82.522ms cudaStreamSynchronize

dd. 0.37% 15.548ms 8 1.9435ms 8.8320us 5.6891ms cudaMemcpy2DAsync

ee. 0.17% 7.0099ms 42 166.90us 9.1630us 1.2212ms cudaMalloc

ff. 0.03% 1.3561ms 4 339.03us 325.85us 350.78us cuDeviceTotalMem

gg. 0.03% 1.2785ms 4 319.63us 41.523us 1.0619ms cudaStreamCreate

hh. 0.02% 879.81us 114 7.7170us 632ns 299.26us cudaEventCreateWithFlags

ii. 0.02% 869.92us 352 2.4710us 249ns 67.838us cuDeviceGetAttribute

jj. 0.01% 574.13us 23 24.962us 10.424us 115.76us cudaLaunch

kk. 0.01% 363.69us 6 60.615us 23.365us 123.51us cudaMemcpy

ll. 0.00% 107.03us 4 26.757us 18.648us 33.786us cuDeviceGetName

mm. 0.00% 86.619us 32 2.7060us 642ns 8.5410us cudaSetDevice

nn. 0.00% 77.809us 110 707ns 421ns 2.9440us cudaDeviceGetAttribute

oo. 0.00% 62.630us 147 426ns 258ns 1.3900us cudaSetupArgument

pp. 0.00% 38.201us 2 19.100us 18.753us 19.448us cudaStreamCreateWithPriority

qq. 0.00% 28.214us 23 1.2260us 461ns 3.7020us cudaConfigureCall

rr. 0.00% 22.207us 10 2.2200us 1.4150us 7.1160us cudaGetDevice

ss. 0.00% 12.331us 1 12.331us 12.331us 12.331us cudaBindTexture

tt. 0.00% 8.6790us 16 542ns 360ns 889ns cudaPeekAtLastError

uu. 0.00% 6.5060us 1 6.5060us 6.5060us 6.5060us cudaStreamGetPriority

vv. 0.00% 5.3930us 6 898ns 334ns 2.4280us cuDeviceGetCount

ww. 0.00% 4.3000us 2 2.1500us 1.4530us 2.8470us cudaStreamWaitEvent

xx. 0.00% 4.1450us 2 2.0720us 1.3010us 2.8440us cudaEventRecord

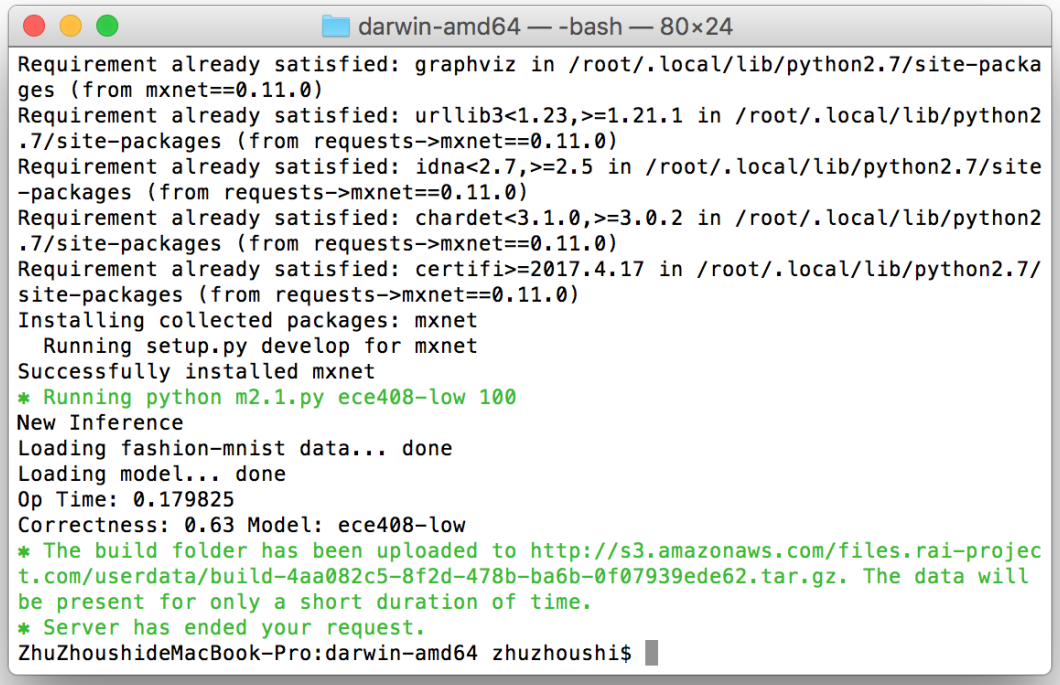
yy. 0.00% 4.1410us 6 690ns 402ns 1.3660us cuDeviceGet

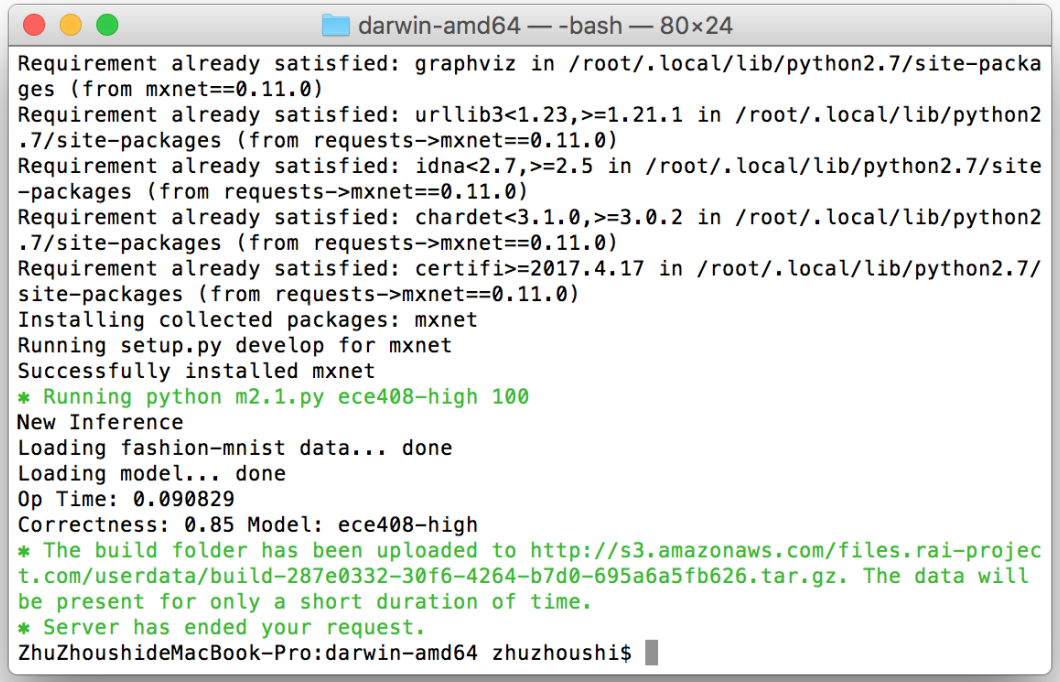
zz. 0.00% 3.5930us 6 598ns 308ns 955ns cudaGetLastError

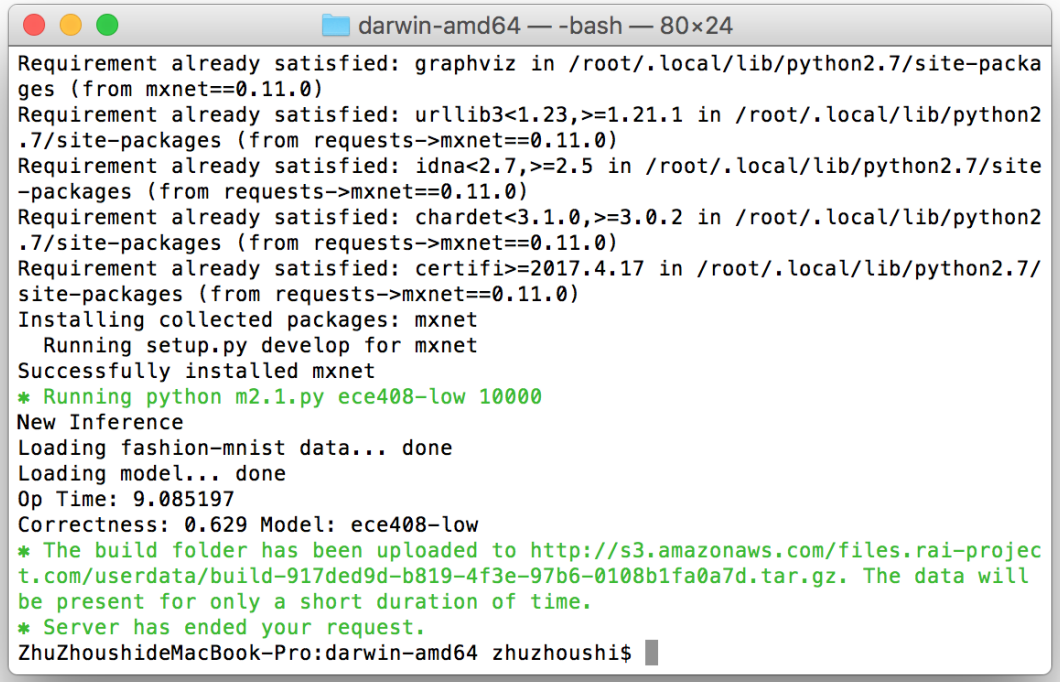
aaa. 0.00% 3.2050us 2 1.6020us 1.2850us 1.9200us cudaDeviceGetStreamPriorityRange

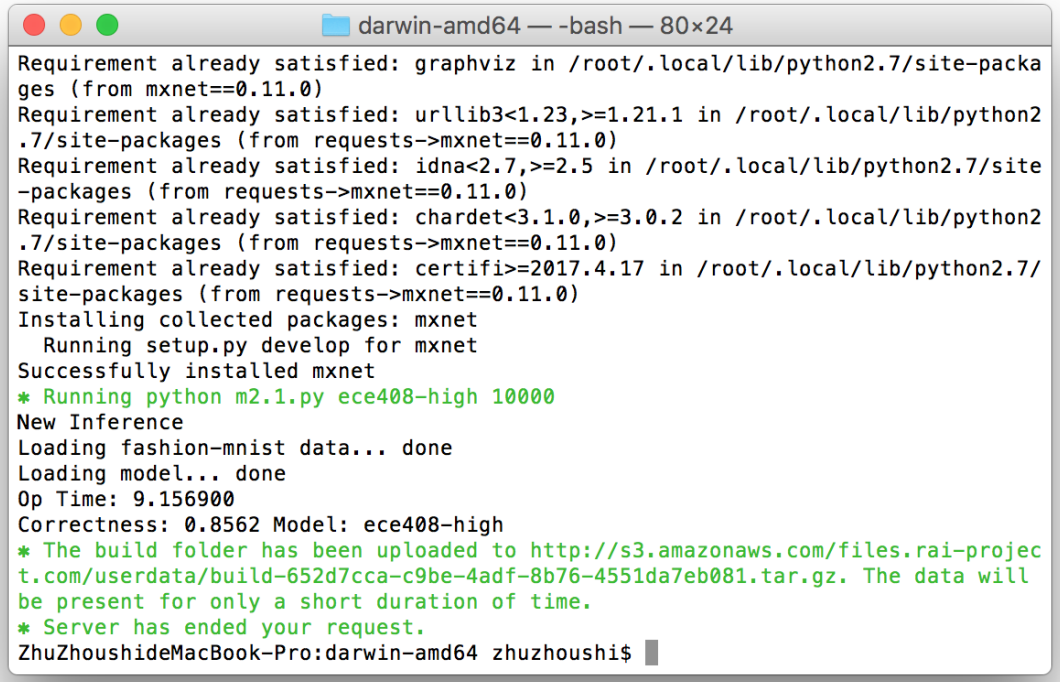
bbb. 0.00% 3.1530us 3 1.0510us 664ns 1.3460us cuInit
ccc. 0.00% 2.3390us 3 779ns 725ns 854ns cuDriverGetVersion
ddd. 0.00% 1.7490us 1 1.7490us 1.7490us 1.7490us cudaUnbindTexture
eee. 0.00% 1.3320us 1 1.3320us 1.3320us 1.3320us cudaGetDeviceCount
fff. * The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-d93fdd29-3569-4c88-992b-2dcf5cee0b32.tar.gz. The data will be present for only a short duration of time.
ggg. * Server has ended your request.

M2.1 Result - time



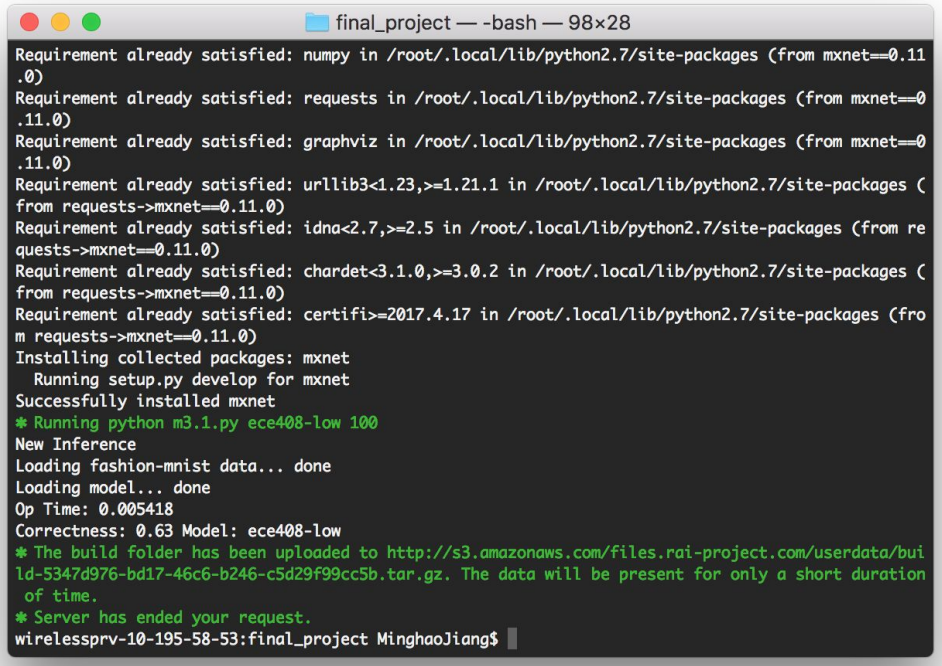

darwin-amd64 — -bash — 80×24
Requirement already satisfied: graphviz in /root/.local/lib/python2.7/site-packages (from mxnet==0.11.0)
Requirement already satisfied: urllib3<1.23,>=1.21.1 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: idna<2.7,>=2.5 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: certifi>=2017.4.17 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Installing collected packages: mxnet
Running setup.py develop for mxnet
Successfully installed mxnet
* Running python m2.1.py ece408-high 100
New Inference
Loading fashion-mnist data... done
Loading model... done
Op Time: 0.090829
Correctness: 0.85 Model: ece408-high
* The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-287e0332-30f6-4264-b7d0-695a6a5fb626.tar.gz. The data will be present for only a short duration of time.
* Server has ended your request.
ZhuZhoushideMacBook-Pro:darwin-amd64 zhuzhoushi$


darwin-amd64 — -bash — 80×24
Requirement already satisfied: graphviz in /root/.local/lib/python2.7/site-packages (from mxnet==0.11.0)
Requirement already satisfied: urllib3<1.23,>=1.21.1 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: idna<2.7,>=2.5 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: certifi>=2017.4.17 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Installing collected packages: mxnet
Running setup.py develop for mxnet
Successfully installed mxnet
* Running python m2.1.py ece408-low 10000
New Inference
Loading fashion-mnist data... done
Loading model... done
Op Time: 9.085197
Correctness: 0.629 Model: ece408-low
* The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-917ded9d-b819-4f3e-97b6-0108b1fa0a7d.tar.gz. The data will be present for only a short duration of time.
* Server has ended your request.
ZhuZhoushideMacBook-Pro:darwin-amd64 zhuzhoushi$

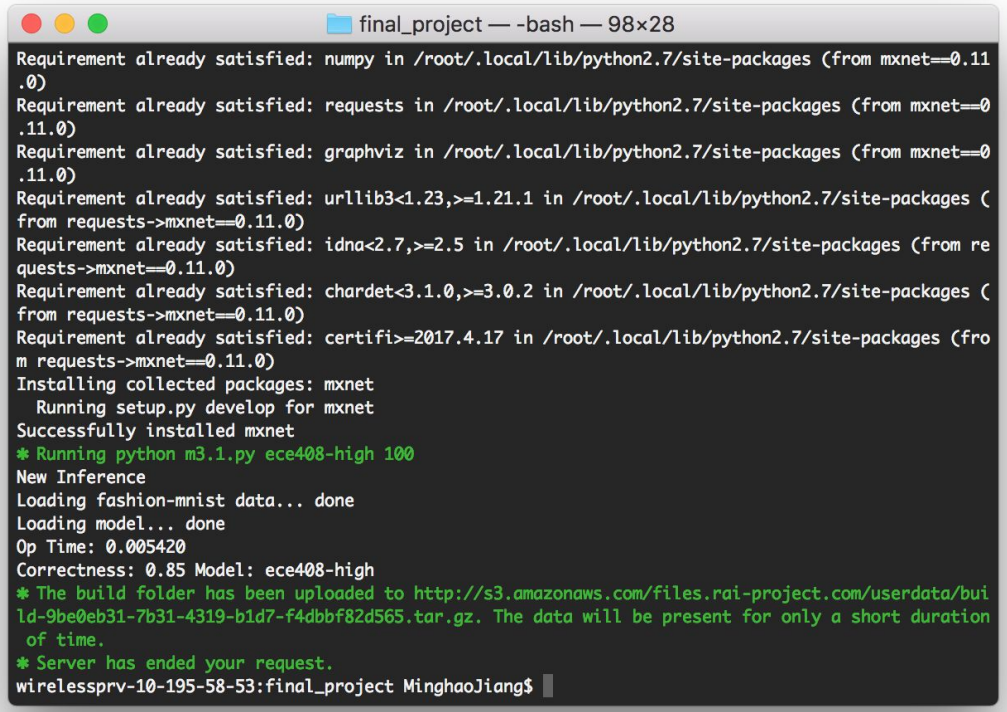

darwin-amd64 — -bash — 80×24
Requirement already satisfied: graphviz in /root/.local/lib/python2.7/site-packages (from mxnet==0.11.0)
Requirement already satisfied: urllib3<1.23,>=1.21.1 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: idna<2.7,>=2.5 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Requirement already satisfied: certifi>=2017.4.17 in /root/.local/lib/python2.7/site-packages (from requests->mxnet==0.11.0)
Installing collected packages: mxnet
Running setup.py develop for mxnet
Successfully installed mxnet
* Running python m2.1.py ece408-high 10000
New Inference
Loading fashion-mnist data... done
Loading model... done
Op Time: 9.156900
Correctness: 0.8562 Model: ece408-high
* The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-652d7cca-c9be-4adf-8b76-4551da7eb081.tar.gz. The data will be present for only a short duration of time.
* Server has ended your request.
ZhuZhoushideMacBook-Pro:darwin-amd64 zhuzhoushi$

M3.1 Result - time

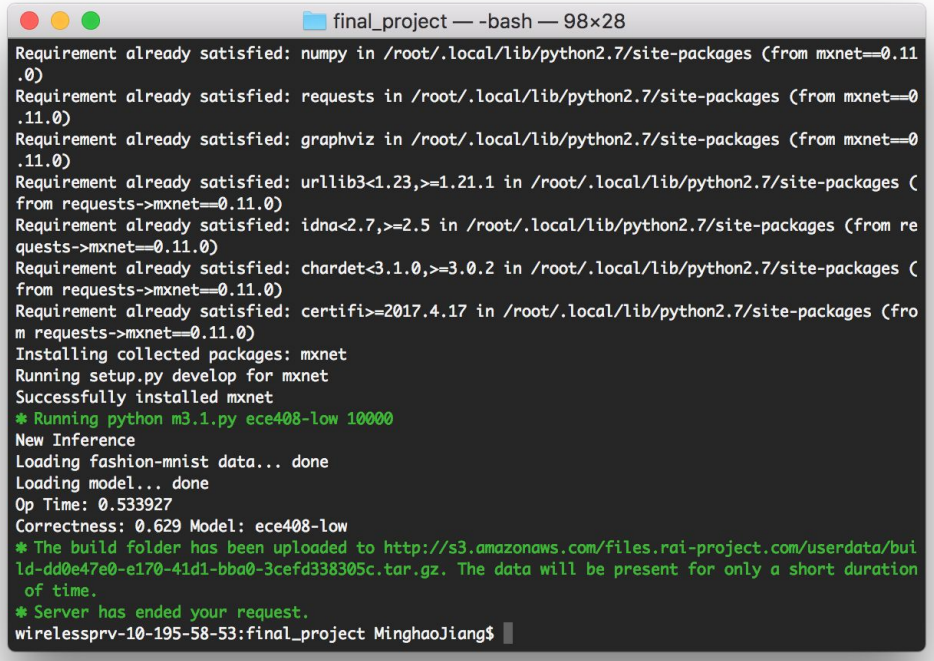
Running ece408-low with size 100



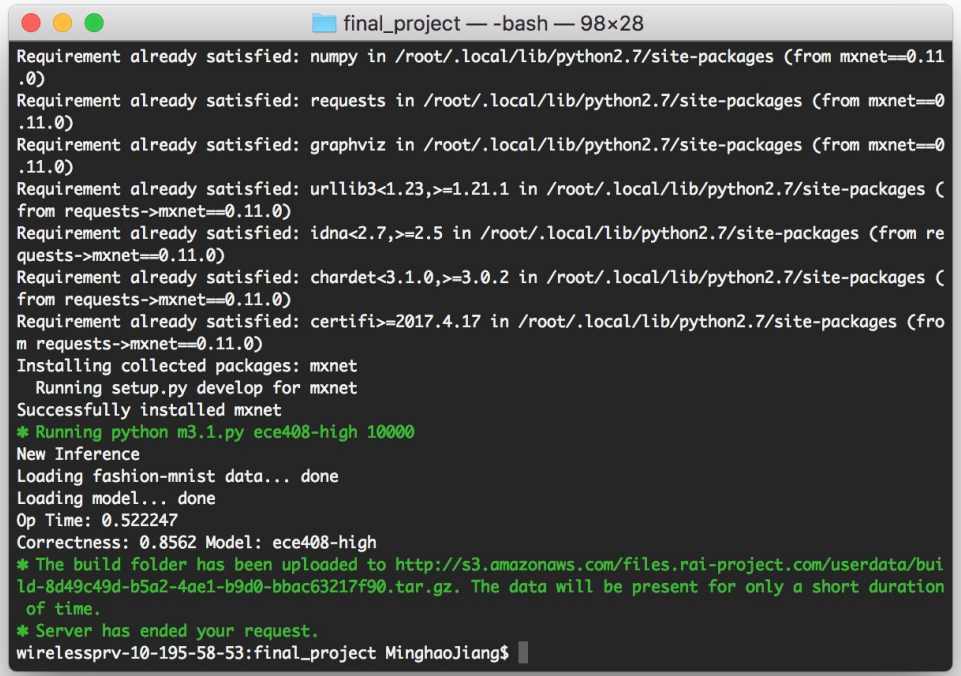
Running ece408-high with size 100



Running ece408-low with size 10000



Running ece408-high with size 10000



M3.1 Result - nvprof

- ✱ Running nvprof python m3.1.py
- New Inference
- Loading fashion-mnist data... done
- ==310== NVPROF is profiling process 310, command: python m3.1.py
- Loading model... done
- Op Time: 0.577711
- Correctness: 0.8562 Model: ece408-high
- ==310== Profiling application: python m3.1.py
- ==310== Profiling result:
- Time(%) Time Calls Avg Min Max Name
- 84.25% 558.14ms 1 558.14ms 558.14ms 558.14ms mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int, int, int, int)
- 5.82% 38.530ms 1 38.530ms 38.530ms 38.530ms sgemm_sm35_ldg_tn_128x8x256x16x32
- 2.95% 19.555ms 1 19.555ms 19.555ms 19.555ms 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)
- 2.92% 19.369ms 2 9.6844ms 455.42us 18.913ms 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*)
- 2.17% 14.392ms 1 14.392ms 14.392ms 14.392ms 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)
- 0.92% 6.1251ms 13 471.16us 1.5040us 4.2093ms [CUDA memcpy HtoD]
- 0.55% 3.6369ms 1 3.6369ms 3.6369ms 3.6369ms sgemm_sm35_ldg_tn_64x16x128x8x32
- 0.17% 1.1041ms 1 1.1041ms 1.1041ms 1.1041ms 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)

- 0.11% 741.94us 12 61.828us 2.0480us 374.30us 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)
- 0.07% 431.10us 2 215.55us 16.640us 414.46us 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<msh adow::gpu, int=1, float>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
- 0.06% 393.37us 1 393.37us 393.37us 393.37us sgemm_sm35_ldg_tn_32x16x64x8x16
- 0.00% 22.592us 1 22.592us 22.592us 22.592us 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::maxi mum, 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)
- 0.00% 9.4720us 1 9.4720us 9.4720us 9.4720us [CUDA memcpy DtoH]
- ==310== API calls:
- Time(%) Time Calls Avg Min Max Name
- 41.33% 1.86962s 18 103.87ms 17.410us 934.41ms cudaStreamCreateWithFlags
- 25.39% 1.14879s 10 114.88ms 791ns 325.67ms cudaFree
- 18.27% 826.70ms 23 35.943ms 235.23us 819.94ms cudaMemGetInfo
- 12.77% 577.66ms 1 577.66ms 577.66ms 577.66ms cudaDeviceSynchronize
- 1.71% 77.476ms 25 3.0990ms 6.0530us 41.421ms cudaStreamSynchronize
- 0.28% 12.473ms 8 1.5591ms 13.237us 4.2810ms cudaMemcpy2DAsync
- 0.14% 6.4915ms 41 158.33us 8.7790us 1.1259ms cudaMalloc
- 0.03% 1.3604ms 4 340.09us 338.69us 343.69us cuDeviceTotalMem
- 0.02% 840.10us 114 7.3690us 480ns 295.60us cudaEventCreateWithFlags
- 0.02% 833.55us 352 2.3680us 243ns 62.967us cuDeviceGetAttribute
- 0.01% 577.92us 24 24.080us 10.141us 71.989us cudaLaunch
- 0.01% 535.25us 4 133.81us 42.062us 395.01us cudaStreamCreate
- 0.01% 351.53us 6 58.587us 24.381us 118.89us cudaMemcpy

- 0.00% 97.276us 4 24.319us 17.929us 29.121us cuDeviceGetName
- 0.00% 77.852us 145 536ns 223ns 16.722us cudaSetupArgument
- 0.00% 72.606us 30 2.4200us 630ns 7.3620us cudaSetDevice
- 0.00% 67.437us 104 648ns 355ns 1.5340us cudaDeviceGetAttribute
- 0.00% 48.768us 2 24.384us 18.672us 30.096us cudaStreamCreateWithPriority
- 0.00% 46.039us 24 1.9180us 434ns 21.469us cudaConfigureCall
- 0.00% 19.854us 10 1.9850us 1.3980us 5.7600us cudaGetDevice
- 0.00% 8.6470us 17 508ns 388ns 794ns cudaPeekAtLastError
- 0.00% 4.8760us 6 812ns 317ns 2.0070us cuDeviceGetCount
- 0.00% 4.7770us 1 4.7770us 4.7770us 4.7770us cudaStreamGetPriority
- 0.00% 4.0320us 2 2.0160us 1.3620us 2.6700us cudaStreamWaitEvent
- 0.00% 3.6750us 6 612ns 328ns 1.0610us cuDeviceGet
- 0.00% 3.6540us 2 1.8270us 1.2030us 2.4510us cudaEventRecord
- 0.00% 2.9560us 2 1.4780us 1.3190us 1.6370us cudaDeviceGetStreamPriorityRange
- 0.00% 2.6450us 3 881ns 824ns 912ns cuInit
- 0.00% 2.6400us 3 880ns 809ns 1.0010us cuDriverGetVersion
- 0.00% 2.5710us 5 514ns 349ns 662ns cudaGetLastError
- 0.00% 1.1440us 1 1.1440us 1.1440us 1.1440us cudaGetDeviceCount
- ✱ The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-27f9bc3b-2dc3-448c-8e49-ce76015ae001.tar.gz. The data will be present for only a short duration of time.
- ✱ Server has ended your request.

2. Optimization Approach and Results

TBD

3. References

TBD

4. (Optional) Suggestions for Improving Next Year

TBD