

# ECE408 Project Milestone 2 Report

Team: BananaUber

Names and Netids: Mingren Feng(mingren3), Zhengliang Zhu(zz40), Ao Li(aol3)

School affiliation: UIUC

## 1. All kernels that collectively consume more than 90% of program time:

Time(%)	Time	Calls	Avg	Min	Max	Name
22.79%	23.207ms	1	23.207ms	23.207ms	23.207ms	volta_scudnn_128x32_relu_interior_nn_v1
20.87%	21.246ms	1	21.246ms	21.246ms	21.246ms	void cudnn::detail::implicit_convolve_sgemm<float, float, int=1024, int=5, int=5, 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=1, bool=1, bool=0, bool=1>*, kernel_conv_params, int, float, float, int, float, int, int)
7.41%	7.5455ms	1	7.5455ms	7.5455ms	7.5455ms	volta_sgemm_128x128_tn
7.28%	7.4108ms	2	3.7054ms	24.895us	7.3859ms	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*)
4.33%	4.4101ms	1	4.4101ms	4.4101ms	4.4101ms	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)
0.50%	506.36us	1	506.36us	506.36us	506.36us	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)

## 2. All CUDA API calls that collectively consume more than 90% of program time:

Time(%)	Time	Calls	Avg	Min	Max	Name
37.47%	2.55985s	22	116.36ms	13.057us	1.41544s	cudaStreamCreateWithFlags
34.29%	2.34289s	24	97.620ms	67.410us	2.33776s	cudaMemGetInfo
21.90%	1.49590s	19	78.732ms	310ns	392.33ms	cudaFree

3. The difference between the API call and the kernel call is that a CUDA API call is a function call to the functions that have been defined in the CUDA. While the kernel call are all self implemented kernel functions. Kernel calls runs on GPU.

## 4. Output of rai running MXNet on the CPU:

\* Running `nvprof python m1.1.py`

Loading fashion-mnist data... done

Loading model... done

New Inference

EvalMetric: {'accuracy': 0.8177}

\* The build folder has been uploaded to

<http://s3.amazonaws.com/files.rai-project.com/userdata/build-5bcf923d2ec41922e94ef5a9.tar.gz>.

The data will be present for only a short duration of time.

\* Server has ended your request.

Program Running Time: 13.41s

## 5. Output of rai running MXNet on the GPU:

\* Running `/usr/bin/time python m1.2.py`

Loading fashion-mnist data... done

Loading model... done

New Inference

EvalMetric: {'accuracy': 0.8177}

4.18user 2.52system 0:04.61elapsed 145%CPU (0avgtext+0avgdata 2838492maxresident)k  
0inputs+4568outputs (0major+703787minor)pagefaults 0swaps

Program Running Time: 4.61s

6. Milestone2 result:

\* Running /usr/bin/time python m2.1.py 10000

Loading fashion-mnist data... done

Loading model...[20:53:31] src/nnvm/legacy\_json\_util.cc:204: Warning: loading symbol saved by  
MXNet version 10300 with lower version of MXNet v10200. May cause undefined behavior. Please  
update MXNet if you encounter any issue

done

New Inference

Op Time: 26.281501

Op Time: 154.224093

Correctness: 0.8171 Model: ece408

192.24user 7.42system 3:04.23elapsed 108%CPU (0avgtext+0avgdata 5867948maxresident)k  
0inputs+2544outputs (0major+2252713minor)pagefau

lts 0swaps

Whole program execution time: 192.24user 7.42system 3:04.23elapsed 108%CPU

Op time for first layer: 26.281501s

Op time for second layer: 154.224093s