Final Report - ECE 408
Elijiah Baird - ebaird2
Jai Agrawal - jagrawa2
Kevin Wang - klwang4

1. Baseline Results

I. M1.1: mxnet CPU layer correctness and elapsed time for the whole python program. (These were found by running ‘time’ before the call to rai)

```
real: 0m13.103s
user: 0m0.300s
sys:  0m0.056s
```

II. M1.2/M1.3: mxnet GPU layer performance results

The Cuda implementation in m1.2 spends the majority of its time not calling a kernel function. The main call to `implicit_convolve_sgemm` only takes about 50 ms. This is the most expensive call taking 37% of the kernel time. This is compared to the most expensive CPU time which was 1.8 seconds.

III. M2.1: baseline cpu implementation correctness and performance results

```
Op Time: 12.190647
Correctness: 0.8562 Model: ece408-high
```

IV. M3.1: baseline gpu implementation correctness and performance results

```
Op Time: 2.867524
Correctness: 0.8562 Model: ece408-high

Op Time: 2.921161
Correctness: 0.629 Model: ece408-low
```

2. Optimization Approach and Results

1) Local y for sum

In our very first implementation of our kernel, we were saving y to global memory each addition. We knew this was a terrible idea and remedied this immediately when beginning to optimize. Later on with unrolling, we combined multiple summations into one line of code. We believed this would increase performance due to less memory accesses. However, this didn’t improve performance much if at all because the local_y variable was a register in the thead and has very fast access time.

2) Shared memory (k and x)

Another optimization we made was to use shared memory, in the form of local_k and local_x. Utilizing shared memory over global memory greatly improved our initial time as we did not have to wait for global memory accesses. This was the a factor in dropping our runtime from ~2.5 seconds to ~300 milliseconds in our first round of optimization. Later in our process we tried bypassing the defined k4d function when loading into shared memory. We did not see a drastic speed increases from, but the fastest time we saw run was using this method so we left it in our final code.

3) blockdim (32,32) vs (1024,1,1)

Since each image in the dataset had a resolution of 28x28, 32 was the lowest power of two that could satisfy block dimensions, resolution in the fewest number of unused threads. However with this setup, every warp was diverging since we only used 28 of the 32 threads in the x dimension. We remedied this by launching 1024 threads only in the x dimension and them mapping those to the x and y dimension. This made it so that only 1 warp would diverge in the calculations. This was the other optimization that dropped our runtime from ~2.5 seconds to ~300 milliseconds.

4) griddim dimensions (b,1,1) vs (b,m,1)

Changing the dimensions from (b,1,1) to (b,m,1) wasn't worth it because of the extra overhead used. We thought that this would be beneficial because the M loop would not need to be run in each thread, but thus this simply did not work and was fruitless optimization. Our runtime when using (B,1,1) was 0.2529 while using (B,M,1) was 0.448.

5) local_y_array

We tried saving the resulting y values into an array and then writing them back after the main loop of the kernel. We were unable to test this because each thread did not have enough registers to hold this array.

6) Loop Unrolling

We unrolled the innermost for loops. We thought the overhead from checking an if statement for each loop was much greater than the actual computation. We unrolled these operations into 25 lines. This cause our runtime on ece-high to drop from 0.20315 to 0.078242. After this we were confused as to why the compiler wouldn't do this automatically. We replaced the 'K' in the conditional statement with a defined 'CONST_K' as 5. This got our runtime to a similar speed as explicitly unrolling the loops.

We also replaced the ‘M’ in the outer loop with 50. This optimization didn’t lead to as drastic of an improvement though. After further runs we decided to leave in the explicit unrolling, as it consistently gave us a better ece-low run-time even though it made our ece-high slightly slower. The ece-low became 0.0689 from 0.077 seconds.

7) Using Constants

One optimization we made was to use constants instead of using the variables the passed in values. Specializing the code for the specific instance would allow us to use less resources resulting in faster times. This also made the compiler automatically unroll our q and p for loops. This kept our code cleaner while staying efficient.

Work Load:

We all met up on multiple occasions at Grainger library. We hooked up to a monitor and pair programmed. Elijah was driving as we all were offering optimizations and error correction.

In the end we improved our run time on ece-high to 0.0768 and ece-low to 0.0686 seconds. Our time on the submission was 68.25 ms. This was a vast improvement of our milestone 3 runtime of 2.867 seconds on ece-high.

3. References

- ECE 408 Lecture Notes

4. Appendix

NVProf outputs:

M1.2:

```
✱ Running nvprof python /src/m1.2.py
Loading fashion-mnist data... done
==310== NVPROF is profiling process 310, command: python /src/m1.2.py
Loading model...[23:39:19] 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}
==310== Profiling application: python /src/m1.2.py
==310== Profiling result:
Time(%)      Time     Calls       Avg       Min       Max  Name
36.98%  49.974ms         1  49.974ms  49.974ms  49.974ms  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)
 28.69%  38.765ms         1  38.765ms  38.765ms  38.765ms  sgemm_sm35_ldg_tn_128x8x256x16x32
 14.34%  19.373ms         2  9.6863ms  459.03us  18.914ms  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*)
10.69%  14.444ms         1  14.444ms  14.444ms  14.444ms  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)
  4.53%  6.1197ms        13  470.74us  1.5680us  4.1874ms  [CUDA memcpy HtoD]
  2.76%  3.7341ms         1  3.7341ms  3.7341ms  3.7341ms  sgemm_sm35_ldg_tn_64x16x128x8x32
  0.82%  1.1119ms         1  1.1119ms  1.1119ms  1.1119ms  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%  748.76us        12  62.396us  2.1120us  378.01us  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%  433.46us         2  216.73us  16.607us  416.86us  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%  392.12us         1  392.12us  392.12us  392.12us  sgemm_sm35_ldg_tn_32x16x64x8x16
0.02%  23.359us         1  23.359us  23.359us  23.359us  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%  9.6960us         1  9.6960us  9.6960us  9.6960us  [CUDA memcpy DtoH]
==310== API calls:
Time(%)      Time     Calls       Avg       Min       Max  Name
 46.81%  1.87342s        18  104.08ms  19.408us  936.53ms  cudaStreamCreateWithFlags
```

```
 28.77%  1.15136s        10  115.14ms  1.0350us  326.94ms  cudaFree
 20.62%  825.09ms        24  34.379ms  235.71us  818.08ms  cudaMemGetInfo
  3.20%  128.21ms        25  5.1285ms  6.0260us  83.268ms  cudaStreamSynchronize
  0.31%  12.463ms         8  1.5579ms  7.4010us  4.2909ms  cudaMemcpy2DAsync
  0.16%  6.4391ms        42  153.31us  10.179us  1.1833ms  cudaMalloc
0.03%  1.3871ms         4  346.79us  338.84us  365.44us  cuDeviceTotalMem
  0.03%  1.0322ms       114  9.0540us     755ns  421.82us  cudaEventCreateWithFlags
  0.02%  842.23us       352  2.3920us     242ns  63.095us  cuDeviceGetAttribute
  0.01%  451.56us        23  19.632us  11.277us  83.386us  cudaLaunch
  0.01%  449.96us         6  74.993us  26.503us  127.35us  cudaMemcpy
  0.01%  291.49us         4  72.871us  48.612us  95.757us  cudaStreamCreate
  0.00%  101.99us         4  25.497us  18.311us  31.348us  cuDeviceGetName
  0.00%  89.824us       110     816ns     453ns  2.6350us  cudaDeviceGetAttribute
  0.00%  79.262us        32  2.4760us     715ns  7.6000us  cudaSetDevice
  0.00%  58.119us       147     395ns     254ns  1.3410us  cudaSetupArgument
  0.00%  47.941us         2  23.970us  22.990us  24.951us  cudaStreamCreateWithPriority
  0.00%  28.305us        10  2.8300us  1.3630us  7.0100us  cudaGetDevice
  0.00%  27.804us        23  1.2080us     456ns  2.2990us  cudaConfigureCall
  0.00%  8.1480us         1  8.1480us  8.1480us  8.1480us  cudaBindTexture
  0.00%  7.9560us        16     497ns     374ns     752ns  cudaPeekAtLastError
  0.00%  5.5790us         1  5.5790us  5.5790us  5.5790us  cudaStreamGetPriority
  0.00%  5.2010us         6     866ns     453ns  1.9160us  cuDeviceGetCount
  0.00%  4.1290us         2  2.0640us  1.7770us  2.3520us  cudaDeviceGetStreamPriorityRange
  0.00%  3.8720us         2  1.9360us  1.6050us  2.2670us  cudaEventRecord
  0.00%  3.8020us         2  1.9010us  1.4370us  2.3650us  cudaStreamWaitEvent
  0.00%  3.3870us         6     564ns     381ns     792ns  cuDeviceGet
  0.00%  3.1500us         6     525ns     338ns     737ns  cudaGetLastError
  0.00%  2.8510us         3     950ns     841ns  1.1140us  cuInit
  0.00%  2.0330us         3     677ns     611ns     752ns  cuDriverGetVersion
  0.00%  1.5010us         1  1.5010us  1.5010us  1.5010us  cudaUnbindTexture
  0.00%  1.0690us         1  1.0690us  1.0690us  1.0690us  cudaGetDeviceCount
```

M3.1:

```
* Checking your athentication credentials.
* Preparing your project directory for upload.
* Uploading your project directory. This may take a few minutes.
 15.40 KiB / 15.40 KiB [████████████████████████████████████████] 100.00% 108.67
KiB/s 0s
* Folder uploaded. Server is now processing your submission.
* Your job request has been posted to the queue.
* Server has accepted your job submission and started to configure the container.
* Downloading your code.
* Using cwpearson/2017fa_ece408_mxnet_docker:amd64-gpu-latest as container image.
* Starting container.
* Running /bin/bash -c "cp -rv /src/* /build"
'/src/README.md' -> '/build/README.md'
'/src/build_example' -> '/build/build_example'
'/src/build_example/Makefile' -> '/build/build_example/Makefile'
'/src/build_example/main.cu' -> '/build/build_example/main.cu'
'/src/ece408_src' -> '/build/ece408_src'
```

```
'/src/ece408_src/new-forward.cuh' -> '/build/ece408_src/new-forward.cuh'
'/src/ece408_src/new-forward.h' -> '/build/ece408_src/new-forward.h'
'/src/ece408_src/new-inl.h' -> '/build/ece408_src/new-inl.h'
'/src/ece408_src/new.cc' -> '/build/ece408_src/new.cc'
'/src/ece408_src/new.cu' -> '/build/ece408_src/new.cu'
'/src/final.py' -> '/build/final.py'
'/src/m1.1.py' -> '/build/m1.1.py'
'/src/m1.2.py' -> '/build/m1.2.py'
'/src/m2.1.py' -> '/build/m2.1.py'
'/src/m3.1.py' -> '/build/m3.1.py'
'/src/rai_build.yml' -> '/build/rai_build.yml'
'/src/reader.py' -> '/build/reader.py'
* Running /bin/bash -c "for src in ece408_src/*; do cp -v $src /mxnet/src/operator/custom/.;
done"
'ece408_src/new-forward.cuh' -> '/mxnet/src/operator/custom/./new-forward.cuh'
'ece408_src/new-forward.h' -> '/mxnet/src/operator/custom/./new-forward.h'
'ece408_src/new-inl.h' -> '/mxnet/src/operator/custom/./new-inl.h'
'ece408_src/new.cc' -> '/mxnet/src/operator/custom/./new.cc'
'ece408_src/new.cu' -> '/mxnet/src/operator/custom/./new.cu'
* Running nice -n20 make -C /mxnet
make: Entering directory '/mxnet'
g++ -std=c++11 -c -DMSHADOW_FORCE_STREAM -Wall -Wsign-compare -O3 -DNDEBUG=1 -I/mxnet/mshadow/
-I/mxnet/dmlc-core/include -fPIC -I/mxnet/nnvm/include -I/mxnet/dlpack/include -Iinclude
-funroll-loops -Wno-unused-variable -Wno-unused-parameter -Wno-unknown-pragmas
-Wno-unused-local-typedefs -msse3 -I/usr/local/cuda/include -DMSHADOW_USE_CBLAS=1
-DMSHADOW_USE_MKL=0 -DMSHADOW_RABIT_PS=0 -DMSHADOW_DIST_PS=0 -DMSHADOW_USE_PASCAL=0
-DMXNET_USE_PROFILER=1 -DMXNET_USE_OPENCV=0 -fopenmp -DMXNET_USE_LAPACK -DMSHADOW_USE_CUDNN=1
-I/usr/include/openblas -Wno-strict-aliasing -Wno-sign-compare -ftrack-macro-expansion=0
-Wno-misleading-indentation -I/mxnet/cub -DMXNET_USE_NVRTC=0 -MMD -c src/operator/custom/new.cc
-o build/src/operator/custom/new.o
cd /mxnet/dmlc-core; make libdmlc.a USE_SSE=1 config=/mxnet/config.mk; cd /mxnet
make[1]: Entering directory '/mxnet/dmlc-core'
make[1]: 'libdmlc.a' is up to date.
make[1]: Leaving directory '/mxnet/dmlc-core'
/usr/local/cuda/bin/nvcc -std=c++11 -Xcompiler -D_FORCE_INLINES -O3 -ccbin g++  -gencode
arch=compute_30,code=sm_30 -gencode arch=compute_35,code=sm_35 -gencode
arch=compute_50,code=sm_50 -gencode arch=compute_52,code=sm_52 -gencode
arch=compute_60,code=sm_60 -gencode arch=compute_61,code=[sm_61,compute_61] --fatbin-options
-compress-all -Xcompiler "-DMSHADOW_FORCE_STREAM -Wall -Wsign-compare -O3 -DNDEBUG=1
-I/mxnet/mshadow/ -I/mxnet/dmlc-core/include -fPIC -I/mxnet/nnvm/include -I/mxnet/dlpack/include
-Iinclude -funroll-loops -Wno-unused-variable -Wno-unused-parameter -Wno-unknown-pragmas
-Wno-unused-local-typedefs -msse3 -I/usr/local/cuda/include -DMSHADOW_USE_CBLAS=1
-DMSHADOW_USE_MKL=0 -DMSHADOW_RABIT_PS=0 -DMSHADOW_DIST_PS=0 -DMSHADOW_USE_PASCAL=0
-DMXNET_USE_PROFILER=1 -DMXNET_USE_OPENCV=0 -fopenmp -DMXNET_USE_LAPACK -DMSHADOW_USE_CUDNN=1
-I/usr/include/openblas -Wno-strict-aliasing -Wno-sign-compare -ftrack-macro-expansion=0
-Wno-misleading-indentation -I/mxnet/cub -DMXNET_USE_NVRTC=0" -M -MT
build/src/operator/custom/new_gpu.o src/operator/custom/new.cu
>build/src/operator/custom/new_gpu.d
/usr/local/cuda/bin/nvcc -c -o build/src/operator/custom/new_gpu.o -std=c++11 -Xcompiler
-D_FORCE_INLINES -O3 -ccbin g++  -gencode arch=compute_30,code=sm_30 -gencode
arch=compute_35,code=sm_35 -gencode arch=compute_50,code=sm_50 -gencode
arch=compute_52,code=sm_52 -gencode arch=compute_60,code=sm_60 -gencode
arch=compute_61,code=[sm_61,compute_61] --fatbin-options -compress-all -Xcompiler
"-DMSHADOW_FORCE_STREAM -Wall -Wsign-compare -O3 -DNDEBUG=1 -I/mxnet/mshadow/
-I/mxnet/dmlc-core/include -fPIC -I/mxnet/nnvm/include -I/mxnet/dlpack/include -Iinclude
-funroll-loops -Wno-unused-variable -Wno-unused-parameter -Wno-unknown-pragmas
-Wno-unused-local-typedefs -msse3 -I/usr/local/cuda/include -DMSHADOW_USE_CBLAS=1
```

```
-DMSHADOW_USE_MKL=0 -DMSHADOW_RABIT_PS=0 -DMSHADOW_DIST_PS=0 -DMSHADOW_USE_PASCAL=0
-DMXNET_USE_PROFILER=1 -DMXNET_USE_OPENCV=0 -fopenmp -DMXNET_USE_LAPACK -DMSHADOW_USE_CUDNN=1
-I/usr/include/openblas -Wno-strict-aliasing -Wno-sign-compare -ftrack-macro-expansion=0
-Wno-misleading-indentation -I/mxnet/cub -DMXNET_USE_NVRTC=0" src/operator/custom/new.cu
ar crv lib/libmxnet.a build/src/operator/custom/new.o build/src/operator/custom/new_gpu.o
a - build/src/operator/custom/new.o
a - build/src/operator/custom/new_gpu.o
g++ -DMSHADOW_FORCE_STREAM -Wall -Wsign-compare -O3 -DNDEBUG=1 -I/mxnet/mshadow/
-I/mxnet/dmlc-core/include -fPIC -I/mxnet/nnvm/include -I/mxnet/dlpack/include -Iinclude
-funroll-loops -Wno-unused-variable -Wno-unused-parameter -Wno-unknown-pragmas
-Wno-unused-local-typedefs -msse3 -I/usr/local/cuda/include -DMSHADOW_USE_CBLAS=1
-DMSHADOW_USE_MKL=0 -DMSHADOW_RABIT_PS=0 -DMSHADOW_DIST_PS=0 -DMSHADOW_USE_PASCAL=0
-DMXNET_USE_PROFILER=1 -DMXNET_USE_OPENCV=0 -fopenmp -DMXNET_USE_LAPACK -DMSHADOW_USE_CUDNN=1
-I/usr/include/openblas -Wno-strict-aliasing -Wno-sign-compare -ftrack-macro-expansion=0
-Wno-misleading-indentation -I/mxnet/cub -DMXNET_USE_NVRTC=0 -shared -o lib/libmxnet.so
build/src/operator/nn/softmax.o build/src/operator/tensor/elemwise_binary_broadcast_op_extended.o
build/src/operator/tensor/elemwise_binary_op_extended.o build/src/operator/tensor/init_op.o
build/src/operator/tensor/elemwise_binary_broadcast_op_basic.o
build/src/operator/tensor/broadcast_reduce_op_index.o
build/src/operator/tensor/broadcast_reduce_op_value.o build/src/operator/tensor/control_flow_op.o
build/src/operator/tensor/elemwise_binary_op_basic.o
build/src/operator/tensor/elemwise_unary_op.o build/src/operator/tensor/elemwise_sum.o
build/src/operator/tensor/elemwise_binary_scalar_op_extended.o
build/src/operator/tensor/indexing_op.o build/src/operator/tensor/ordering_op.o
build/src/operator/tensor/elemwise_binary_broadcast_op_logic.o build/src/operator/tensor/la_op.o
build/src/operator/tensor/elemwise_binary_op_logic.o
build/src/operator/tensor/elemwise_binary_scalar_op_basic.o build/src/operator/tensor/matrix_op.o
build/src/operator/tensor/elemwise_binary_scalar_op_logic.o
build/src/operator/contrib/multibox_target.o build/src/operator/contrib/count_sketch.o
build/src/operator/contrib/dequantize.o build/src/operator/contrib/deformable_psroi_pooling.o
build/src/operator/contrib/fft.o build/src/operator/contrib/multibox_prior.o
build/src/operator/contrib/ctc_loss.o build/src/operator/contrib/multi_proposal.o
build/src/operator/contrib/psroi_pooling.o build/src/operator/contrib/quantize.o
build/src/operator/contrib/deformable_convolution.o build/src/operator/contrib/ifft.o
build/src/operator/contrib/multibox_detection.o build/src/operator/contrib/proposal.o
build/src/operator/custom/native_op.o build/src/operator/custom/ndarray_op.o
build/src/operator/custom/new.o build/src/operator/custom/custom.o
build/src/operator/random/sample_multinomial_op.o build/src/operator/random/multisample_op.o
build/src/operator/random/sample_op.o build/src/operator/nnpack/nnpack_util.o
build/src/operator/mkl/mkl_cppwrapper.o build/src/operator/mkl/mkl_memory.o build/src/io/io.o
build/src/io/image_aug_default.o build/src/io/iter_image_det_recordio.o build/src/io/image_io.o
build/src/io/image_det_aug_default.o build/src/io/iter_csv.o build/src/io/iter_image_recordio.o
build/src/io/iter_mnist.o build/src/io/iter_image_recordio_2.o build/src/common/mxrtc.o
build/src/nnvm/legacy_op_util.o build/src/nnvm/legacy_json_util.o build/src/ndarray/autograd.o
build/src/ndarray/ndarray_function.o build/src/ndarray/ndarray.o
build/src/operator/instance_norm.o build/src/operator/svm_output.o
build/src/operator/bilinear_sampler.o build/src/operator/pooling.o build/src/operator/crop.o
build/src/operator/spatial_transformer.o build/src/operator/swapaxis.o
build/src/operator/convolution_v1.o build/src/operator/softmax_output.o
build/src/operator/operator_util.o build/src/operator/sequence_reverse.o
build/src/operator/batch_norm_v1.o build/src/operator/rnn.o build/src/operator/operator.o
build/src/operator/correlation.o build/src/operator/deconvolution.o
build/src/operator/optimizer_op.o build/src/operator/lrn.o build/src/operator/convolution.o
build/src/operator/pooling_v1.o build/src/operator/pad.o build/src/operator/fully_connected.o
build/src/operator/sequence_mask.o build/src/operator/sequence_last.o
build/src/operator/grid_generator.o build/src/operator/cudnn_algoreg.o
```

build/src/operator/identity_attach_KL_sparse_reg.o build/src/operator/activation.o
build/src/operator/upsamplin
g.o build/src/operator/cudnn_batch_norm.o build/src/operator/loss_binary_op.o
build/src/operator/regression_output.o build/src/operator/l2_normalization.o
build/src/operator/slice_channel.o build/src/operator/concat.o build/src/operator/leaky_relu.o
build/src/operator/roi_pooling.o build/src/operator/batch_norm.o build/src/operator/dropout.o
build/src/operator/cross_device_copy.o build/src/operator/softmax_activation.o
build/src/operator/make_loss.o build/src/engine/profiler.o build/src/engine/naive_engine.o
build/src/engine/threaded_engine_pooled.o build/src/engine/threaded_engine.o
build/src/engine/engine.o build/src/engine/threaded_engine_perdevice.o
build/src/storage/storage.o build/src/c_api/c_api_symbolic.o build/src/c_api/c_api_ndarray.o
build/src/c_api/c_api_executor.o build/src/c_api/c_predict_api.o build/src/c_api/c_api_function.o
build/src/c_api/c_api.o build/src/c_api/c_api_error.o
build/src/executor/inplace_addto_detect_pass.o build/src/executor/graph_executor.o
build/src/executor/attach_op_execs_pass.o build/src/executor/attach_op_resource_pass.o
build/src/kvstore/kvstore.o build/src/resource.o build/src/initialize.o
/mxnet/dmlc-core/libdmlc.a build/src/operator/nn/softmax_gpu.o
build/src/operator/tensor/indexing_op_gpu.o
build/src/operator/tensor/elemwise_binary_op_extended_gpu.o
build/src/operator/tensor/elemwise_binary_scalar_op_extended_gpu.o
build/src/operator/tensor/elemwise_binary_scalar_op_basic_gpu.o
build/src/operator/tensor/ordering_op_gpu.o
build/src/operator/tensor/elemwise_binary_broadcast_op_basic_gpu.o
build/src/operator/tensor/elemwise_binary_op_basic_gpu.o
build/src/operator/tensor/elemwise_sum_gpu.o build/src/operator/tensor/init_op_gpu.o
build/src/operator/tensor/elemwise_binary_scalar_op_logic_gpu.o
build/src/operator/tensor/broadcast_reduce_op_index_gpu.o
build/src/operator/tensor/matrix_op_gpu.o
build/src/operator/tensor/broadcast_reduce_op_value_gpu.o
build/src/operator/tensor/control_flow_op_gpu.o build/src/operator/tensor/elemwise_unary_op_gpu.o
build/src/operator/tensor/elemwise_binary_broadcast_op_extended_gpu.o
build/src/operator/tensor/elemwise_binary_broadcast_op_logic_gpu.o
build/src/operator/tensor/la_op_gpu.o build/src/operator/tensor/elemwise_binary_op_logic_gpu.o
build/src/operator/contrib/ctc_loss_gpu.o build/src/operator/contrib/psroi_pooling_gpu.o
build/src/operator/contrib/quantize_gpu.o build/src/operator/contrib/deformable_convolution_gpu.o
build/src/operator/contrib/ifft_gpu.o build/src/operator/contrib/multibox_detection_gpu.o
build/src/operator/contrib/multibox_target_gpu.o build/src/operator/contrib/proposal_gpu.o
build/src/operator/contrib/count_sketch_gpu.o build/src/operator/contrib/dequantize_gpu.o
build/src/operator/contrib/deformable_psroi_pooling_gpu.o build/src/operator/contrib/fft_gpu.o
build/src/operator/contrib/multibox_prior_gpu.o build/src/operator/contrib/multi_proposal_gpu.o
build/src/operator/custom/native_op_gpu.o build/src/operator/custom/new_gpu.o
build/src/operator/random/sample_multinomial_op_gpu.o build/src/operator/random/sample_op_gpu.o
build/src/ndarray/ndarray_function_gpu.o build/src/operator/sequence_reverse_gpu.o
build/src/operator/optimizer_op_gpu.o build/src/operator/lrn_gpu.o
build/src/operator/pooling_v1_gpu.o build/src/operator/fully_connected_gpu.o
build/src/operator/sequence_mask_gpu.o build/src/operator/swapaxis_gpu.o
build/src/operator/regression_output_gpu.o build/src/operator/leaky_relu_gpu.o
build/src/operator/identity_attach_KL_sparse_reg_gpu.o build/src/operator/activation_gpu.o
build/src/operator/roi_pooling_gpu.o build/src/operator/cudnn_batch_norm_gpu.o
build/src/operator/loss_binary_op_gpu.o build/src/operator/convolution_gpu.o
build/src/operator/make_loss_gpu.o build/src/operator/batch_norm_gpu.o
build/src/operator/upsampling_gpu.o build/src/operator/slice_channel_gpu.o
build/src/operator/dropout_gpu.o build/src/operator/softmax_activation_gpu.o
build/src/operator/svm_output_gpu.o build/src/operator/pad_gpu.o
build/src/operator/deconvolution_gpu.o build/src/operator/correlation_gpu.o
build/src/operator/instance_nor

```
m_gpu.o build/src/operator/concat_gpu.o build/src/operator/l2_normalization_gpu.o
build/src/operator/grid_generator_gpu.o build/src/operator/sequence_last_gpu.o
build/src/operator/pooling_gpu.o build/src/operator/convolution_v1_gpu.o
build/src/operator/crop_gpu.o build/src/operator/spatial_transformer_gpu.o
build/src/operator/softmax_output_gpu.o build/src/operator/bilinear_sampler_gpu.o
build/src/operator/batch_norm_v1_gpu.o build/src/operator/rnn_gpu.o -pthread -lm -lcudart
-lcublas -lcurand -lcusolver -L/usr/local/cuda/lib64 -L/usr/local/cuda/lib -lopenblas -fopenmp
-lrt -llapack -lcudnn  -lcuda -lcufft \
 -Wl,--whole-archive /mxnet/nnvm/lib/libnnvm.a -Wl,--no-whole-archive
make: Leaving directory '/mxnet'
* Running pip install --user -e /mxnet/python
Obtaining file:///mxnet/python
Requirement already satisfied: numpy in /root/.local/lib/python2.7/site-packages (from
mxnet==0.11.0)
Requirement already satisfied: requests in /root/.local/lib/python2.7/site-packages (from
mxnet==0.11.0)
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 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: 2.867524
Correctness: 0.8562 Model: ece408-high
==310== Profiling application: python m3.1.py
==310== Profiling result:
Time(%)      Time     Calls       Avg       Min       Max  Name
96.46%  2.84794s         1  2.84794s  2.84794s  2.84794s  mxnet::op::forward_kernel(float*, float
const *, float const *, int, int, int, int, int, int)
  1.31%  38.617ms         1  38.617ms  38.617ms  38.617ms  sgemm_sm35_ldg_tn_128x8x256x16x32
0.66%  19.549ms         1  19.549ms  19.549ms  19.549ms  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)
  0.66%  19.368ms         2  9.6838ms  455.74us  18.912ms  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*)
0.49%  14.402ms         1  14.402ms  14.402ms  14.402ms  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)
  0.21%  6.2264ms        13  478.95us  1.5360us  4.3090ms  [CUDA memcpy HtoD]
  0.12%  3.6684ms         1  3.6684ms  3.6684ms  3.6684ms  sgemm_sm35_ldg_tn_64x16x128x8x32
0.04%  1.1073ms         1  1.1073ms  1.1073ms  1.1073ms  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)
  0.03%  742.01us        12  61.833us  2.0480us  374.36us  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)
0.01%  432.51us         2  216.25us  17.472us  415.04us  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)
  0.01%  383.32us         1  383.32us  383.32us  383.32us  sgemm_sm35_ldg_tn_32x16x64x8x16
0.00%  23.071us         1  23.071us  23.071us  23.071us  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)
  0.00%  9.4710us         1  9.4710us  9.4710us  9.4710us  [CUDA memcpy DtoH]
==310== API calls:
Time(%)      Time     Calls       Avg       Min       Max  Name
 42.28%  2.86748s         1  2.86748s  2.86748s  2.86748s  cudaDeviceSynchronize
 27.37%  1.85599s        18  103.11ms  16.100us  927.67ms  cudaStreamCreateWithFlags
 16.77%  1.13733s        10  113.73ms  1.0400us  323.36ms  cudaFree
 12.08%  819.31ms        23  35.622ms  237.32us  812.53ms  cudaMemGetInfo
  1.15%  77.739ms        25  3.1095ms  5.6080us  41.640ms  cudaStreamSynchronize
  0.19%  12.746ms         8  1.5932ms  18.624us  4.4363ms  cudaMemcpy2DAsync
  0.09%  6.3168ms        41  154.07us  9.8160us  1.1125ms  cudaMalloc
  0.02%  1.3725ms         4  343.14us  338.25us  357.09us  cuDeviceTotalMem
  0.01%  918.33us       114  8.0550us     619ns  306.65us  cudaEventCreateWithFlags
  0.01%  863.74us       352  2.4530us     244ns  86.349us  cuDeviceGetAttribute
  0.01%  505.84us        24  21.076us  9.8180us  51.508us  cudaLaunch
  0.01%  427.22us         4  106.81us  38.946us  290.45us  cudaStreamCreate
  0.01%  349.62us         6  58.269us  23.525us  121.04us  cudaMemcpy
  0.00%  96.433us         4  24.108us  17.544us  29.717us  cuDeviceGetName
  0.00%  70.648us        30  2.3540us     600ns  7.1850us  cudaSetDevice
  0.00%  67.790us       104     651ns     416ns  1.5050us  cudaDeviceGetAttribute
  0.00%  62.766us       145     432ns     253ns  1.5700us  cudaSetupArgument
  0.00%  38.107us         2  19.053us  18.592us  19.515us  cudaStreamCreateWithPriority
  0.00%  28.006us        24  1.1660us     377ns  2.5840us  cudaConfigureCall
  0.00%  24.203us        10  2.4200us  1.3150us  6.0590us  cudaGetDevice
  0.00%  9.1710us        17     539ns     394ns     832ns  cudaPeekAtLastError
  0.00%  4.0210us         6     670ns     281ns  1.5810us  cuDeviceGetCount
  0.00%  3.7880us         1  3.7880us  3.7880us  3.7880us  cudaStreamGetPriority
  0.00%  3.6650us         6     610ns     370ns  1.2370us  cuDeviceGet
  0.00%  3.6330us         2  1.8160us  1.3980us  2.2350us  cudaStreamWaitEvent
  0.00%  3.3750us         2  1.6870us  1.3070us  2.0680us  cudaEventRecord
  0.00%  2.8650us         2  1.4320us  1.3390us  1.5260us  cudaDeviceGetStreamPriorityRange
  0.00%  2.6890us         3     896ns     769ns  1.0080us  cuInit
```

```
  0.00%  2.5220us         5     504ns     377ns     584ns  cudaGetLastError
  0.00%  2.4490us         3     816ns     777ns     857ns  cuDriverGetVersion
  0.00%  1.1450us         1  1.1450us  1.1450us  1.1450us  cudaGetDeviceCount
* The build folder has been uploaded to
http://s3.amazonaws.com/files.rai-project.com/userdata/build-e1e8137b-311b-462d-b71b-c52ce03392a6
.tar.gz. The data will be present for only a short duration of time.
* Server has ended your request.

====================
* Running nvprof python m3.1.py ece408-low
New Inference
Loading fashion-mnist data... done
==310== NVPROF is profiling process 310, command: python m3.1.py ece408-low
Loading model... done
Op Time: 2.921161
Correctness: 0.629 Model: ece408-low
==310== Profiling application: python m3.1.py ece408-low
==310== Profiling result:
Time(%)      Time     Calls       Avg       Min       Max  Name
 96.50%  2.90158s         1  2.90158s  2.90158s  2.90158s  mxnet::op::forward_kernel(float*,
float const *, float const *, int, int, int, int, int, int)
  1.31%  39.248ms         1  39.248ms  39.248ms  39.248ms  sgemm_sm35_ldg_tn_128x8x256x16x32
0.65%  19.546ms         1  19.546ms  19.546ms  19.546ms  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)
  0.65%  19.395ms         2  9.6973ms  460.92us  18.934ms  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*)
  0.48%  14.503ms         1  14.503ms  14.503ms  14.503ms  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)
  0.21%  6.1797ms        13  475.36us  1.5040us  4.2534ms  [CUDA memcpy HtoD]
  0.12%  3.6519ms         1  3.6519ms  3.6519ms  3.6519ms  sgemm_sm35_ldg_tn_64x16x128x8x32
0.04%  1.1224ms         1  1.1224ms  1.1224ms  1.1224ms  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)
  0.03%  754.97us        12  62.913us  2.1120us  380.99us  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)
0.01%  437.82us         2  218.91us  17.344us  420.48us  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)
  0.01%  398.04us         1  398.04us  398.04us  398.04us  sgemm_sm35_ldg_tn_32x16x64x8x16
```

```
  0.00%  23.711us         1  23.711us  23.711us  23.711us  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)
  0.00%  9.5680us         1  9.5680us  9.5680us  9.5680us  [CUDA memcpy DtoH]

==310== API calls:
Time(%)      Time     Calls       Avg       Min       Max  Name
 42.87%  2.92112s         1  2.92112s  2.92112s  2.92112s  cudaDeviceSynchronize
 26.99%  1.83951s        18  102.19ms  17.255us  919.41ms  cudaStreamCreateWithFlags
 16.53%  1.12655s        10  112.65ms     793ns  321.60ms  cudaFree
 12.08%  823.04ms        23  35.784ms  207.79us  816.34ms  cudaMemGetInfo
  1.16%  78.800ms        25  3.1520ms  6.0090us  42.562ms  cudaStreamSynchronize
  0.18%  12.107ms         8  1.5134ms  7.9860us  4.3067ms  cudaMemcpy2DAsync
  0.10%  6.4866ms        41  158.21us  12.232us  1.1072ms  cudaMalloc
  0.04%  2.4761ms         4  619.03us  33.347us  2.3431ms  cudaStreamCreate
  0.02%  1.3934ms         4  348.36us  339.14us  363.08us  cuDeviceTotalMem
  0.01%  851.48us       352  2.4180us     244ns  63.493us  cuDeviceGetAttribute
  0.01%  678.71us       114  5.9530us     668ns  303.36us  cudaEventCreateWithFlags
  0.01%  516.79us        24  21.532us  10.424us  51.474us  cudaLaunch
  0.01%  414.20us         6  69.033us  57.630us  85.653us  cudaMemcpy
  0.00%  117.14us         4  29.284us  17.267us  45.302us  cuDeviceGetName
  0.00%  70.550us        30  2.3510us     651ns  6.7700us  cudaSetDevice
  0.00%  62.945us       104     605ns     411ns  2.0310us  cudaDeviceGetAttribute
  0.00%  58.437us       145     403ns     254ns  1.4480us  cudaSetupArgument
  0.00%  36.884us         2  18.442us  17.928us  18.956us  cudaStreamCreateWithPriority
  0.00%  28.615us        24  1.1920us     391ns  2.8760us  cudaConfigureCall
  0.00%  16.469us        10  1.6460us  1.1680us  2.8640us  cudaGetDevice
  0.00%  9.2660us        17     545ns     334ns     807ns  cudaPeekAtLastError
  0.00%  4.5100us         6     751ns     285ns  1.5780us  cuDeviceGetCount
  0.00%  4.1540us         6     692ns     345ns  1.2700us  cuDeviceGet
  0.00%  4.1360us         1  4.1360us  4.1360us  4.1360us  cudaStreamGetPriority
  0.00%  3.9770us         2  1.9880us  1.4020us  2.5750us  cudaStreamWaitEvent
  0.00%  3.3280us         2  1.6640us  1.2420us  2.0860us  cudaEventRecord
  0.00%  2.9340us         2  1.4670us  1.3540us  1.5800us  cudaDeviceGetStreamPriorityRange
  0.00%  2.6410us         3     880ns     825ns     954ns  cuInit
  0.00%  2.5710us         3     857ns     707ns  1.0480us  cuDriverGetVersion
0.00%  2.3600us         5     472ns     314ns     621ns  cudaGetLastError
  0.00%     949ns         1     949ns     949ns     949ns  cudaGetDeviceCount
* The build folder has been uploaded to
http://s3.amazonaws.com/files.rai-project.com/userdata/build-424a8142-6cdf-415d-8fe7-d30dc972829f
.tar.gz. The data will be present for only a short duration of time.
* Server has ended your request.
```

```
// FINAL SUBMISSION

* Running nvprof python m3.1.py ece408-high
New Inference
Loading fashion-mnist data... done
==312== NVPROF is profiling process 312, command: python m3.1.py ece408-high
```

```
Loading model... done
Op Time: 0.076815
Correctness: 0.8562 Model: ece408-high
==312== Profiling application: python m3.1.py ece408-high
==312== Profiling result:
Time(%)      Time     Calls       Avg       Min       Max  Name
35.14%  57.193ms         1  57.193ms  57.193ms  57.193ms  mxnet::op::forward_kernel(float*, float
const *, float const *, int, int, int, int, int, int)
 24.26%  39.486ms         1  39.486ms  39.486ms  39.486ms  sgemm_sm35_ldg_tn_128x8x256x16x32
12.04%  19.601ms         1  19.601ms  19.601ms  19.601ms  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)
 11.91%  19.381ms         2  9.6904ms  460.19us  18.921ms  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*)
  8.91%  14.503ms         1  14.503ms  14.503ms  14.503ms  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)
  3.83%  6.2406ms        13  480.04us  1.5360us  4.3215ms  [CUDA memcpy HtoD]
  2.23%  3.6237ms         1  3.6237ms  3.6237ms  3.6237ms  sgemm_sm35_ldg_tn_64x16x128x8x32
0.68%  1.1138ms         1  1.1138ms  1.1138ms  1.1138ms  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)
  0.46%  753.98us        12  62.831us  2.1120us  380.64us  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)
0.27%  436.25us         2  218.13us  16.800us  419.45us  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)
  0.24%  389.53us         1  389.53us  389.53us  389.53us  sgemm_sm35_ldg_tn_32x16x64x8x16
  0.01%  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%  9.6320us         1  9.6320us  9.6320us  9.6320us  [CUDA memcpy DtoH]
==312== API calls:
Time(%)      Time     Calls       Avg       Min       Max  Name
 46.22%  1.91098s        18  106.17ms  16.608us  955.15ms  cudaStreamCreateWithFlags
 27.54%  1.13843s        10  113.84ms     807ns  327.39ms  cudaFree
 21.73%  898.53ms        23  39.066ms  236.94us  891.69ms  cudaMemGetInfo
  1.90%  78.698ms        25  3.1479ms  5.5600us  42.499ms  cudaStreamSynchronize
```

```
  1.86%  76.789ms         1  76.789ms  76.789ms  76.789ms  cudaDeviceSynchronize
  0.31%  12.699ms         8  1.5874ms  8.6390us  4.4141ms  cudaMemcpy2DAsync
  0.18%  7.4028ms         4  1.8507ms  44.553us  7.1680ms  cudaStreamCreate
  0.15%  6.3441ms        41  154.73us  12.603us  1.1006ms  cudaMalloc
  0.03%  1.3656ms         4  341.39us  340.08us  344.49us  cuDeviceTotalMem
  0.02%  905.57us       114  7.9430us     629ns  305.94us  cudaEventCreateWithFlags
  0.02%  847.47us       352  2.4070us     261ns  65.021us  cuDeviceGetAttribute
  0.01%  529.41us        24  22.058us  11.471us  48.487us  cudaLaunch
  0.01%  353.84us         6  58.973us  22.834us  122.92us  cudaMemcpy
  0.00%  101.53us         4  25.382us  18.220us  30.356us  cuDeviceGetName
  0.00%  78.539us        30  2.6170us     645ns  8.9260us  cudaSetDevice
  0.00%  66.464us       145     458ns     267ns  1.5970us  cudaSetupArgument
  0.00%  66.256us       104     637ns     424ns  1.8200us  cudaDeviceGetAttribute
  0.00%  37.974us         2  18.987us  17.951us  20.023us  cudaStreamCreateWithPriority
  0.00%  30.145us        24  1.2560us     422ns  3.5950us  cudaConfigureCall
  0.00%  20.405us        10  2.0400us  1.1590us  6.5760us  cudaGetDevice
  0.00%  10.298us        17     605ns     398ns     864ns  cudaPeekAtLastError
  0.00%  4.5020us         1  4.5020us  4.5020us  4.5020us  cudaStreamGetPriority
  0.00%  4.3640us         6     727ns     280ns  1.5850us  cuDeviceGetCount
  0.00%  3.8430us         2  1.9210us  1.3150us  2.5280us  cudaStreamWaitEvent
  0.00%  3.5450us         2  1.7720us  1.1230us  2.4220us  cudaEventRecord
  0.00%  3.3750us         6     562ns     382ns     731ns  cuDeviceGet
  0.00%  3.1280us         3  1.0420us     838ns  1.4200us  cuInit
  0.00%  2.8820us         2  1.4410us  1.2500us  1.6320us  cudaDeviceGetStreamPriorityRange
  0.00%  2.4940us         5     498ns     409ns     635ns  cudaGetLastError
  0.00%  2.1200us         3     706ns     651ns     801ns  cuDriverGetVersion
  0.00%  1.1700us         1  1.1700us  1.1700us  1.1700us  cudaGetDeviceCount
* The build folder has been uploaded to
http://s3.amazonaws.com/files.rai-project.com/userdata/build-450bab9a-499b-47dc-8ae6-78cfa1d35189
.tar.gz. The data will be present for only a short duration of time.
* Server has ended your request.
```