ECE408/CS483 Final Project Report

SHRIYAA MITTAL (smittal6), TIANCHENG WU (twu54) , CHUANKAI ZHAO (czhao37), UIUC

TEAM NAME: smartconvolutionteam

MILESTONE 1, DUE : OCTOBER 24, 2018

Report deliverables

Include a list of all kernels that collectively consume more than 90% of the program time

Top 10 kernels are as below:

(1) volta_scudnn_128x32_relu_interior_nn_v1
(2) Implicit_convolve_sgemm
(3) volta_sgemm_128x128_tn
(4) activation_fw_4d_kernel
(5) pooling_fw_4d_kernel
(6) MapPlanLargeKernel
(7) SoftmaxKernel
(8) MapPlanKernel
(9) volta_sgemm_32x32_sliced1x4_tn
(10) computeOffsetsKernel

Include a list of all CUDA API calls that collectively consume more than 90% of the program time

Top 10 CUDA API calls are as below:

(1) cudaStreamCreateWithFlags
(2) cudaMemGetInfo
(3) cudaFree
(4) cudaEventCreateWithFlags
(5) cudaMemcpy2DAsync
(6) cudaFuncSetAttribute
(7) cudaStreamSynchronize
(8) cudaMalloc
(9) cudaGetDeviceProperties
(10) cudaMemcpy

Include an explanation of the difference between kernels and API calls

Kernels are programmer defined functions, while API calls are built-in.

Show output of rai running MXNet on the CPU

```
* Running /usr/bin/time python m1.1.py
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8177}
```

List program run time

```
19.48user 4.09system 0:13.30elapsed 177%CPU
```

Show 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}
```

List program run time

```
4.05user 2.67system 0:04.63elapsed 145%CPU
```