Project Milestone 1-2
Report

Caiting Wu (cwu72)
Hua Chen (huachen4)
Zekun Hu (zekunhu2)

Milestone 1

1. A List of all Kernels that Collectively Consume More than 90% of the Program Time:

Time(%)	Time(ms)	Name
34.00	**118.46**	void fermiPlusCgemmLDS128_batched<...>
26.94	**93.879**	void cudnn::detail::implicit_convolve_sgemm<...>
12.65	**44.060**	void fft2d_c2r_32x32<...>
8.19	**28.540**	sgemm_sm35_ldg_tn_128x8x256x16x32
6.65	**23.153**	[CUDA memcpy HtoD]
4.06	**14.157**	void cudnn::detail::activation_fw_4d_kernel<...>
3.81	**13.289**	void cudnn::detail::pooling_fw_4d_kernel<...>
1.71	**5.9454**	void fft2d_r2c_32x32<...>
1.16	**4.0542**	sgemm_sm35_ldg_tn_64x16x128x8x32

2. A list of all CUDA API calls that collectively consume more than 90% of the program time:

Time(%)	Time(ms)	Name
43.48	**1.92**	cudaStreamCreateWithFlags
27.11	**1.20**	cudaFree
20.71	**916.54**	cudaMemGetInfo
7.31	**323.45**	cudaStreamSynchronize
1.08	**47.75**	cudaMemcpy2DAsync
0.16	**7.15**	cudaMalloc
0.03	**1.37**	cuDeviceTotalMem

3. Explain the difference between kernels and API calls

Kernels are function launched to be executed on the device while APIs are used to communicate between different software components.

4. Show Output of rai Running MXNet on the CPU:

```
Successfully installed mxnet
*  Running /usr/bin/time python m1.1.py
Loading fashion-mnist data...
done
Loading model...
done
New Inference
EvalMetric: {'accuracy': 0.8444}
13.16user 12.04system 0:11.76elapsed 214%CPU (0avgtext+0avgdata 2830016maxresident)k
0inputs+2624outputs (0major+36851minor)pagefaults 0swaps
```

Figure: the snapshot of the result we have by running MXNet on the CPU

5. Program Run-Time (CPU):

From the result shown in the PowerShell, we can tell the run-time on the CPU is 11.76 seconds.

6. Show Output of rai Running MXNet on the GPU:

```
Successfully installed mxnet
*  Running /usr/bin/time python m1.2.py
Loading fashion-mnist data...
done
Loading model...
[04:54:48] src/operator/./../cudnn_algoreg-inl.h:112: Running performance tests to find
o disable)
done
New Inference
EvalMetric: {'accuracy': 0.8444}
2.22user 1.09system 0:02.81elapsed 117%CPU (0avgtext+0avgdata 1139152maxresident)k
0inputs+3136outputs (0major+159479minor)pagefa
ults 0swaps
```

Figure: the snapshot of the result we have by running MXNet on the GPU

7. Program Run-Time (GPU):

From the result shown in the PowerShell, we can tell the run-time on the GPU is 2.81 seconds.

Milestone 2

1. List whole program execution time:

a). 10000 images (default): 30.10 *s*

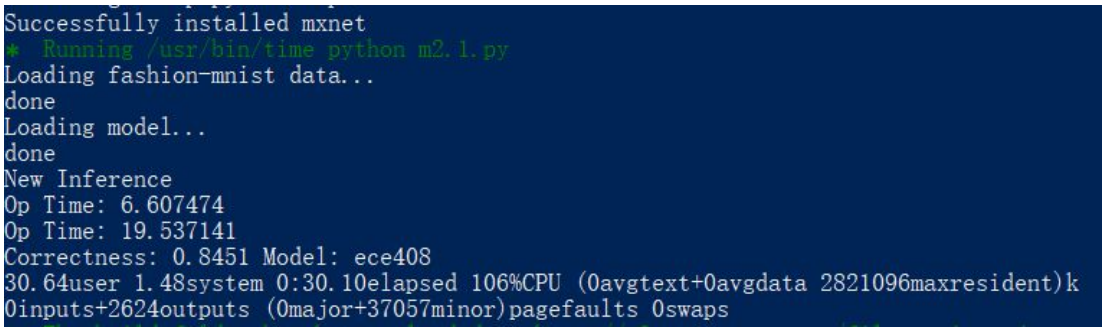

Figure: the snapshot of the result we have by running ConvNet on the CPU with 10000 images

b). 100 images: 1.15 *s*

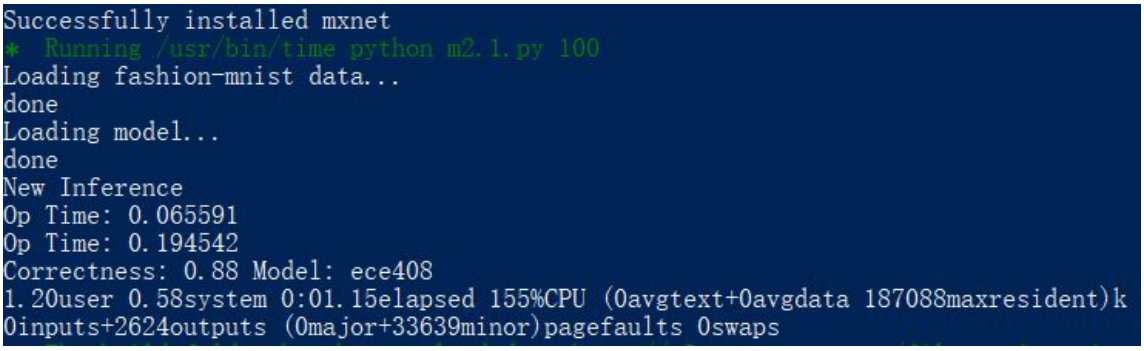

Figure: the snapshot of the result we have by running ConvNet on the CPU with 100 images

c). 10 images (default): 0.87 *s*

```
Successfully installed mxnet
*  Running /usr/bin/time python m2.1.py 10
Loading fashion-mnist data...
done
Loading model...
done
New Inference
Op Time: 0.006572
Op Time: 0.019520
Correctness: 1.0 Model: ece408
0.86user 0.51system 0:00.87elapsed 156%CPU (0avgtext+0avgdata 170392maxresident)k
0inputs+2624outputs (0major+31144minor)pag
efaults 0swaps
```

Figure: the snapshot of the result we have by running ConvNet on the CPU with 10 images

2. List Op Times:

a). 10000 images (default):

First Layer Op Time:	6.60747 *s*
Second Layer Op Time:	19.537141 *s*

b). 100 images:

First Layer Op Time:	0.065591 *s*
Second Layer Op Time:	0.194542 *s*

c). 10 images:

First Layer Op Time:	0.006572 *s*
Second Layer Op Time:	0.019520 *s*