Milestone2:
Team: coolda
Zheng Feng, zfeng8, RailD: 5d97b1bd88a5ec28f9cb9448
Long Hong: liuhong2, RailD: 5d97b1c988a5ec28f9cb9460
Dingyu Peng: dpeng4, RailD: 5d97b1f988a5ec28f9cb94b9
School Affiliation: UIUC on campus

- All kernels that collectively consume more than 90% of the program time.

```
32.40%  35.357ms        20  1.7679ms  1.0880us  33.031ms  [CUDA memcpy HtoD]
18.39%  20.071ms  1  20.071ms  20.071ms  20.071ms
volta_scudnn_128x64_relu_interior_nn_v1
16.21%  17.693ms         4  4.4232ms  4.3519ms  4.6367ms volta_gcgemm_64x32_nt
8.72%  9.5153ms         4  2.3788ms  1.9783ms  3.1088ms  void fft2d_c2r_32x32<float,
bool=0, bool=0, unsigned int=0, bool=0, bool=0>(float*, float2 const *, int, int, int, int, int,
int, int, int, int, float, float, cudnn::reduced_divisor, bool, float*, float*, int2, int, int)
7.21%  7.8714ms         1  7.8714ms  7.8714ms  7.8714ms  volta_sgemm_128x128_tn
```

- All CUDA API calls that collectively consume more than 90% of the program time.

```
41.80%  3.17683s       22  144.40ms  14.638us  1.62690s cudaStreamCreateWithFlags
33.89%  2.57551s       24  107.31ms  56.562us  2.57005s cudaMemGetInfo
21.02%  1.59732s       19  84.069ms  1.2180us  429.36ms  cudaFree
```

- Explanation of the difference between kernels and API calls
  Kernels are device-level codes, running parallel in SM, while API calls are host-to-device communication codes.

- output of rai running MXNet on the CPU
  'accuracy': 0.8154

- List program run time
  9.4s

- output of rai running MXNet on the GPU
  'accuracy': 0.8154

- List program run time
  4.52s

- CPU program execution time
  76.5s

- Op Times

	1.1	2.1 100	2.1 1000	2.1 10000
Op1 time	11.640115	0.158422	1.098643	10.917119
Op2 time	62.344837	0.636590	5.913035	59.204878

MileStone3:
Memory copy host to device: 31.82ms
Forward kernel: 123.43ms
Map Plan LargeKernel: 16.74ms
Memory copy device to host: 6.88µs
Volta segmentation: 128 * 128   time: 7.83ms

Epoch 1 Op time: 0.030191 s.
Epoch 2 Op time: 0.093297 s.