ECE 408 Applied Parallel Programming, Milestone 2

deaplearners

Yiliang Wang (wang513), Shuchen Zhang (szhan114), Tao Sun (taosun2)

Q1 Include a list of all kernels that collectively consume more than 90% of the program time.

Kernel	Time	Cummulative Time
CUDA memcpy HtoD	37.73%	37.73%
volta_scudnn_128x32_relu_interior_nn_v1	21.47%	59.20%
cudnn::detail::implicit_convolve_sgemm	21.32%	80.52%
cudnn::detail::activation_fw_4d_kernel	7.45%	87.97%
volta_sgemm_128x128_tn	6.87%	94.84%

Q2 Include a list of all CUDA API calls that collectively consume more than 90% of the program time.

Kernel	Time	Cummulative Time
cudaStreamCreateWithFlags	38.74%	38.74%
cudaMemGetInfo	36.72%	75.46%
cudaFree	21.46%	96.92%

Q3 Include an explanation of the difference between kernels and API calls.
Kernels are user-defined C functions. When called, kernels are executed multiple times in parallel by a number of different CUDA threads, as opposed to regular C functions executing only once.

The API calls are an interface provided by CUDA to facilitate users familiar with the C programming language to easily write GPU programs. The API calls provide implicit initialization, context management, and module management to ease the device code management.

Q4 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}
19.83user 3.97system 0:13.42elapsed 177%CPU (0avgtext+0avgdata 5955540maxresident)k
0inputs+2856out
puts (0major+1586517minor)pagefaults 0swaps
```

Q5 List program run time on the CPU. The program run time on the CPU is 13.42 seconds.

```
19.83user 3.97system 0:13.42elapsed 177%CPU (0avgtext+0avgdata 5955540maxresident)k
0inputs+2856out
puts (0major+1586517minor)pagefaults 0swaps
```

Q6 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}
```

Q7 List program run time on the GPU.
The program run time on the CPU is 4.59 seconds.

```
4.37user 2.67system 0:04.59elapsed 153%CPU (0avgtext+0avgdata 2833892maxresident)k
0inputs+4568outputs (0major+703015minor)pagefaults 0swaps
```

Q8 List whole program execution time for the cpu convolution implementation.
The whole program execution time is 3 minutes and 2.90 seconds.

```
190.55user 7.44system 3:02.90elapsed 108%CPU (0avgtext+0avgdata 5871140maxresident)k
0inputs+2856outputs (0major+2252827minor)pagefaul
ts 0swaps
```

Q9 List Op times for the cpu convolution implementation.
The Op times for the two layers are 25.485743 seconds and 153.849626 seconds.

```
Op Time: 25.485743
Op Time: 153.849626
Correctness: 0.8171 Model: ece408
```