ECE408

Group: pascalpros

Members: Kelvin Yang, Samarth Jain, Varun Govind

Milestone 1

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

Time(%) Time Calls Avg Min Max Name

34.00% 118.48ms 9 13.164ms 13.149ms 13.179ms void fermiPlusCgemmLDS128_batched<bool=0, bool=1, bool=0, bool=0, int=4, int=4, int=4, int=3, int=3, bool=1, bool=1>(float2**, float2**, float2**, float2*, float2 const *, float2 const *, int, int, int, int, int, int, __int64, __int64, __int64, float2 const *, float2 const *, float2, float2, int)

26.95% 93.911ms 1 93.911ms 93.911ms 93.911ms 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)

12.65% 44.082ms 9 4.8980ms 2.6815ms 6.2784ms void fft2d_c2r_32x32<float, bool=0, unsigned int=0, bool=0, bool=0>(float*, float2 const *, int, int, int, int, int, int, int, int, int, int, float, float, cudnn::reduced_divisor, bool, float*, float*)

8.28% 28.840ms 1 28.840ms 28.840ms 28.840ms sgemm_sm35_ldg_tn_128x8x256x16x32

6.88% 23.961ms 14 1.7115ms 1.5360us 23.131ms [CUDA memcpy HtoD]

4.07% 14.173ms 2 7.0866ms 251.80us 13.921ms 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*)

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

Time(%) Time Calls Avg Min Max Name

37.37% 1.42798s 18 79.332ms 18.065us 713.65ms cudaStreamCreateWithFlags

29.12%	1.11277s	10	111.28ms	1.2480us	299.87ms	cudaFree
23.44%	895.76ms	27	33.176ms	272.95us	887.44ms	cudaMemGetInfo
8.44%	322.35ms	29	11.115ms	6.5980us	194.00ms	cudaStreamSynchronize

Report: Include an explanation of the difference between kernels and API calls

A kernel is a low-level program interfacing with the hardware (CPU, RAM, disks, network, ...) on top of which applications are running. It is the lowest level program running on computers. They are direct calls from user land to the kernel, in order to serve a particular request, which can not be directly handled by the user program.

An API is a generic term defining the interface developers have to use when writing code using libraries and a programming language. It is normally a set of functions and objects (in case of an Object oriented language) in the user land.

For example, a user program API cannot decide which file is to be stored where in the hard disk. The kernel calls take care of the low level details. If a user program has to copy a file from one location to another, it will issue a system call to the kernel and ask for that purpose to be served. The kernel will take care, which part of the hard disk is to be read, and which part will store the new copied file.

Report: 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': 0.8444}

Report: List program run time

12.60user 6.20system 0:08.29elapsed 226%CPU (0avgtext+0avgdata 2822936maxresident)k

0inputs+0outputs (0major+37352minor)pagefaults 0swaps

Report: Show output of rai running MXNet on the GPU

*Running /usr/bin/time python m1.2.py

Loading fashion-mnist data...

done

Loading model...

[02:22:03] 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

New Inference

EvalMetric: {'accuracy': 0.8444}

Report: List program run time

2.12user 1.10system 0:02.70elapsed 119%CPU (0avgtext+0avgdata 1137172maxresident)k

0inputs+512outputs (0major+156982minor)pagefaults 0swaps

Milestone 2

Report: List whole program execution time

*Running /usr/bin/time python m2.1.py

Loading fashion-mnist data...

done

Loading model...

done

New Inference

Op Time: 6.188896

Op Time: 19.258472

Correctness: 0.8451 Model: ece408

29.82user 1.56system 0:29.29elapsed 107%CPU (0avgtext+0avgdata 2824132maxresident)k

0inputs+0outputs (0major+36647minor)pagefaults 0swaps

Report: List Op Times

*Running python m2.1.py

Loading fashion-mnist data...

done

Loading model...

done

New Inference

Op Time: 6.183510

Op Time: 19.244188

Correctness: 0.8451 Model: ece408