week-1

JX-Ma

2024/7/26

1 Introduce

This weeks'work is as follow:

- I redraw chart of tflops in different coalescing strategy.
- I had been compiled libtorch of cudnn version, and test performance of im2col with cublas library and used cudnn to accelerate. what's more, I test performance of im2win in GPU by Lu's code.
- I try to used memory-analysis tools of GPU.

2 redraw

I redraw three chart to represent tflops of different coalescing strategy.first chart show tflops of coalescing four dimensions in the first four layers of loop and it's size is 1x1, the next chart show tflops of coalescing three dimensions in the first four layers of loop and it's size is 2x2. the final chart show tflops of coalescing two dimensions in the first four layers of loop and it's size is 6x1.

3 experiment

3.1 experimental environment

- operational system: centOS7.
- CUDA-Version: 11.3
- CUDNN-Version: 8.2.1
- GCC-Version: 9.5
- C++-Version: 201703
- Openmp-Version: 4.5
- NVCC-architecture-flag:-gencode;arch=compute_86,code=sm_86
- CPU-capability-usage: AVX512

3.2 portion of experiment

The following chart represent Tflops of im2win_cudnn, im2win_cublas, libtorch_cublas and libtorch_cudnn.

- the batch size of input tensor is 128 in all experiment
- im2win_cublas: Use libtorch of cublas version, Runing Function is im2winConvCUDAimplentNCHWHPC in convImplentCUDA.cu.
- im2win_cudnn: Use libtorch of cudnn version, Runing Function is im2winConvCUDAimplentNCHWHPC in convImplentCUDA.cu.
- libtorch_cublas: Use libtorch of cublas version, Runing Function is conv2d in cublas library.
- libtorch_cudnn: Use libtorch of cudnn version, Runing Function is conv2d in cublas library.

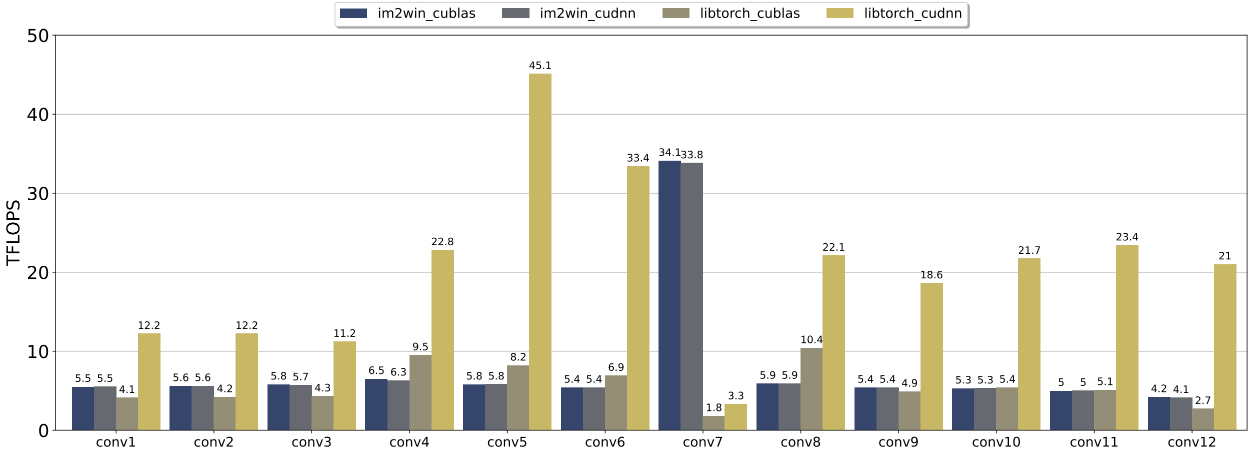

图 1: gflops

3.3 problems of experiment

1. today's problems has benn resolved, I modify parameter of blockDim, I change the size of TLIE_M_PER_BLOCK(in the im2winConvCUDAimplentNCHWHPC) from 128 to 32 and TLIE_N_PER_BLOCK from 128 to 32.the program could run. I attempt to change the size of TLIE_K_PER_BLOCK from 8 to 2 and find the tflops from 4k to 5k gflops in conv1. the size of TLIE_K_PER_BLOCK is 2 on above experiment of im2win.
2. Based on the experimental result, we found the tflops of conv7 in im2win that can reach 33Tflops, the correctness of conv1 to conv12 in im2win convolution has not been verified.
3. for the chart's colors, I use the seaborn library and set palette is "Viridis", it is a gradient colors from deep purple to yellow, it has good readability for color blind individuals.

4 memory-tool

I attempt to use the nvprof to analyse the memory of GPU code, but it will issue a warning, the warning can be concluse ”the nvprof is suitable for sm_80 or lower version.” so I attempt to use the NVIDIA Nsight System to analyse the memory.