note of experiment in week16

zxp

December 28, 2024

1 environment

cpu:Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz (56 cores were applied)

gpu:rtx3090(a piece was applied)

System:CentOS7

Compiler:9.5

2 Experiment

This week i changed im2col on GPU as stated in last week's group meeting, and put the batch transformation inside the kernel function as a loop, so that the number of threads required to transform one batch at a time is the same as the number of threads required to transform all batches at once. But this way of computing all the batches at once on my own machine produced a worse gflops on all the batches than last week. As for why the performance of computing all the batches at once is not as good as computing them one at a time, we first test the time of matrix multiplication. The GPU is faster for matrix multiplication of larger matrices than smaller ones, which is as expected. However, after removing this overhead, the gflops of the all-batch version is still slower on covn1, so the im2col transform of the all-batch version is not fast enough.

Then there is the server issue, this week on the server to be able to use cublas correctly. Then I tried to run two different versions of im2col (batch=128, conv1, 2,11,12) on the server. Probably because the server's graphics card was better, the results were better than my own computer. conv12 is 4 times faster than glops, which computes all batches at once. More data is needed