note of experiment in week2

zxp

September 21, 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 code

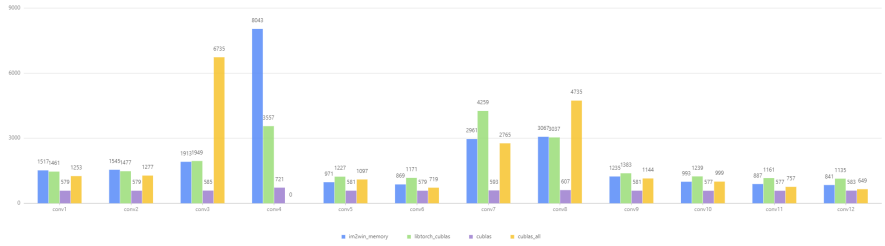

Figure 1: memory

3 Experiment

As for the method of measuring memory, no other method has been found, many tools are based on the interface of the previous tool. In order to better measure the memory, I chose to loop several times in the computation transport section (after the memory transport to the GPU and before the memory transport back to the CPU) to increase the GPU running time. This week I tested my own written call cublas implementation of convolution, writing two versions, one that evaluates one batch at a time (purple in img) and one that evaluates all batches at once (yellow in img, doesn't work in conv4). There are a few problems here, one is that the cublas requires the data to be in the GPU, the other is that the cublas is arranged in a slightly different way than the blas in CPU, and the result is transposed and restored to a tensor. said before that the server suspected that there are other people using memory, this week found the reason, the probability is that the program running before the correct exit led to the application of video memory was not released.

3.1 Analysis

Matrix multiplication in libtorch(cublas) may not compute all batches at once, especially conv3 and conv4, which cannot compute all batches at once.