note of experiment in week24

zxp

August 17, 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

3 Experiment

This week was mainly about understanding the data transfer and computation parts of Lu's code, and how they are specifically implemented. I got it, and it's more flexible than what I wrote. After completing it, I felt there were mostly no issues except for remainder.(but due to tensor dimensions, it is still necessary to set appropriate parameters; there's no problem with setting the microkernel parameters to 8x8 and 4x4, but it's not something that can be used in every layer). Then, I remembered when Ma compared cublas and cudnn last week, the error was quite large. Previously, we used cudnn for comparison, and when we compared using culas, the error with cublas was not big, so the problem lies with cudnn, not that it is both fast and accurate; the largest error in conv5, which resulted in glops exceeding the roofline before, is due to cudnn.