note of experiment in week22

zxp

August 3, 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

It was written that the optimization was added to the 4x4 microkernel, but it was not tested.

3 Experiment

I have tried many different block sizes (from 4 to 128) and compared Lu's code to the convolution in libtorch, but all the code that works has errors greater than 10^4, most are much greater than 10^4.Ever have the odd result which is inf or -1. There are also quite a few different cases in which different errors are reported and the program does not run.

After Fixing setting zero operation in register, results are the same.

3.1 Analysis

Adjusting the block size doesn't make sense in terms of correctness. if the code is correct, these sizes should give the correct result. This code is wrong. It doesn't work. I do not know why it's wrong, The code needs be analyzed in great detail. there are too many threads on GPU. Very tricky.

4 Experiment2

Try to record the GPU memory usage. Use 'nvidia-smi –query-gpu=memory.used –format=csv -L-1' to output the memory usage of the GPU per second. But on the server is not good to open two terminals, and the server has eight GPUs. Has not tried on the server.Still trying how to use a script to test memory and run the program combined. because the program is very fast, 1 second interval is a little long. Do not know if there is a better way to achieve this.