week-20

JX-Ma

2024/12/28

1 INTORDUCE

This week's work is as follows:
This week, we first optimized the CPU code and made a compromise between NHWC and NHCW regarding the calculation order of the output tensor. I found that it was much faster in conv1 and conv3, and there was a significant improvement in conv10. I have understood the im2win code on the GPU, but I still encountered some issues while doing index precomputation. Index precomputation mainly involves pre computing the indexes carried by threads in each thread block by moving elements from global variables to shared memory. In the convolution kernel, we store the first addresses of different convolution kernel batches in the M direction, which is the channel direction of the output tensor. The elements transferred between threads follow a regular pattern, so only the first address needs to be transferred. However, the elements in the input tensor are partially regular, so it would be better to consider whether the process of loading input tensor elements into shared memory should be row first.