week-17

JX-Ma

2024/11/23

1 INTORDUCE

This week's work is as follows:

I optimized the index precomputation on the Im2win data layout on the CPU and wrote the following versions of the algorithm.

The first algorithm is an unoptimized version that uses tuples to store the index of each multiply add element, which wastes a lot of space storing the index and is not very easy to optimize on.

The second algorithm utilizes the characteristic that the convolution window size of each output tensor is the same, reducing the size of the storage index space. This version only needs to store the first address of the input tensor convolution window corresponding to each output tensor element and the address of the corresponding convolution kernel. But the index still stores many duplicate elements.

The third algorithm optimizes the space to the greatest extent possible. Firstly, the input tensor only stores the first address of the convolution window required by the output tensor, and the convolution kernel only stores the first address of each batch. Because we perform convolution calculations in the order of access to the output tensor, we do not need additional space to store the indices of the output tensor elements.

The last algorithm modified the order of convolution, minimizing backtracking in accessing input tensor elements, that is, repeatedly accessing input tensor elements. At this point, the data layout of the output tensor is NHWC