note of 《The Indirect Convolution Algorithm》

zxp

September 14, 2024

1 content

"The Indirect Convolution Algorithm" is a paper of 2019, which is also the problem of the high memory cost of im2col and im2row transformation. The algorithm proposed in this paper is simple, using many Pointers to rows of the input tensor, and reading data directly from the input tensor, so as to reduce memory cost. The paper points out that the method in the paper has a low memory cost, but is slower than im2col when stride is 1 or when the filter size is 1x1. The paper points out that the disadvantage of direct convolution is that different size need special optimization.

2 Feelings

I feel that the method of the paper does not solve the shortcomings of direct convolution. The premise of the paper is that the size of input tensor and filter tensor are fixed in the network