note of 《The Indirect Convolution Algorithm》

zxp

September 21, 2024

1 content

《Fast Algorithms for Convolutional Neural Networks》 is a 15-year paper on convolutional operations with smaller filters. The authors of the paper believe that small filters are becoming more popular and fft is more suitable for this case. The paper is also divided into small pieces, the paper proposes two smaller fft cores, (2x2, 3,x3) with 2.23 fewer computations and (3x3, 4x4) with 4 fewer computations (1/4 of the computations of direct convolution). There are two noteworthy points in this paper. The benchmark used to calculate the accuracy in this paper is double precision naive direct convolution, and the glops calculated by the calculation number of direct convolution exceeds that of roofline. It is concluded that the fft accuracy is not affected much when the core is small, and the accuracy requirement of convolution calculation is not particularly high.

2 Feelings

I feel that the conclusion of the paper is very consistent with the cudnn situation