note of 《The Indirect Convolution Algorithm》

zxp

September 28, 2024

1 content

Did not understand last week, this paper uses winograd, not FFT, the paper believes that FFT is suitable for large filters, and the method mentioned in the paper is based on the minimal filtering algorithms of winograd. Here, by the way, record the conclusion of winograd introduced by Ma Jixiang, the precondition is that the step size is one, and the number of calculations is approximately 1/9.

《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 winograd is more suitable for this case. The paper is also divided into small pieces, the paper proposes two smaller winograd cores, (2x2, 3,x3) with 2.23 fewer computations and (3x3, 4x4) with 4 fewer computations (1/4 of the computations of direct convolution). The memory usage of FFT in cuDNN experiments in this paper is also large

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 winograd 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