note of 《Optimizing Batched Winograd Convolution on GPUs》

zxp

October 19, 2024

1 content

Optimizing Batched Winograd Convolution on GPUs is a 20 year old paper that optimizes Winograd algorithms on Gpus. The paper points out that using Winograd algorithm should be 2.25 times faster than using GEMM in theory, but Winograd on cudnn does not reach the theoretical performance, Winograd on cudnn is only 1.4 times faster than GEMM.

The work of this paper is no reference, the paper thinks cuda and nvcc are not good enough, so the main work of this paper is to achieve a better assembly languages. For optimization on the code, this paper is also to optimize the throughput and partitioning of data. For other methods, the paper mentions FFT, and also believes that Winograd algorithm has advantages in modern comparison, because FFT is more suitable for the filter is relatively large, and Winograd algorithm is more suitable for 3x3 filter.

This paper also uses a standard similar to roofline, but it is theoretical maximum, not single floating-point performance, in this part of the paper will transform and calculate the two parts apart, respectively, to show the theoretical performance that can be achieved.

The paper is also worth referring to the performance representation, the paper uses a similar table, but the depth of color to express the proportion of acceleration, which will be very intuitive.

2 Feelings

Optimizing convolution is too difficult