Title: Fast Algorithms for Convolutional Neural Networks

JX-Ma

2024/9/15

ABSTRACT

Conventional FFT based convolution is fast for large filters, but state of the art convolutional neural networks use small, 3×3 filters. This paper introduce a new class of fast algorithms for convolutional neural networks using Winograd's minimal filtering algorithms.

INTRODUCE

Deep learning networks have achieved good results in image processing, but require a lot of time and resources. The algorithm created in this article is based on the minimum filtering algorithm pioneered by Winograd. The fast algorithm of convolutional neural network can reduce the arithmetic complexity of convolutional network layers by up to 4 times, and almost all arithmetic is performed through dense matrix multiplication.

Convolutional Neural Networks

This section mainly introduces the process formula of convolution

Fast Algorithms

Convolution with a 2x3 matrix and a 3x1 matrix would normally require 6 floating-point multiplications and 6 floating-point additions. Use Winograd to reduce multiplication to 4 times, including 4 data additions, 3 filter additions, 2 constant multiplications, and 4 floating-point multiplications.

Later on, it was mentioned that when the input matrix is 4x4 and convolved with a 3x3 matrix, the stride is 1, and the standard requires 36 multiplications. However, using the Winograd algorithm only requires 16 multiplications, which I haven't fully understood yet. One thing is that using the Winograd algorithm mainly reduces the number of floating-point multiplications.