note of 《Deep Tensor Convolution on Multicores》

zxp

August 24, 2024

1 content

《Deep Tensor Convolution on Multicores》 is a paper from 2017, which is quite old and focuses on memory issues. The method proposed in this paper is based on Winograd and primarily extends Winograd to n dimensions. The paper suggests that the advantages of the Winograd algorithm are fewer computation operations and sparsity of the matrix. the paper argues that the cost of exchanging data between CPUs and GPUs is too high. To address the issue of insufficient memory, the paper proposes distributing the workload across multiple devices. The paper suggests that the transformed matrices based on the Winograd algorithm are more suitable for this distribution, and many frameworks (such as TensorFlow) support this approach. The paper considers it as a breakthrough for overcoming memory limitations. Unlike other Winograd-based algorithms, this paper does not focus on finding the minimal Winograd shape but mainly extends the Winograd algorithm to n dimensions.

2 Feelings

Based on Winograd, expanding to n-dimensional convolution, it has a sense of warping the future.