note of 《Reformulating the direct convolution for high-performance deep learning inference on ARM processors》

zxp

August 10, 2024

1 background

The background is clearly described in the title, which is about implementing high-performance direct convolution on ARM processors. I have read it once before; it moves the work of the paper 《zero-memory overhead direct convolution》 to ARM and the optimization is relatively common. However, the paper does not spend a lot of space on repetitive optimization. Most of it is about describing the two algorithms proposed in the paper. Because this paper ran specific networks, the experimental results section is also quite comprehensive (also a classic combination of speed and memory usage), comparing many different methods (including the two algorithms proposed by the paper authors' performance under different conditions) and comparing the network's aggregated time.

Aside from the background introduction and experiments, the structure of the paper is as follows: introduce direct convolution, introduce the method to optimize direct convolution – blocking, then introduce the memory-friendly data layout NHWC for direct convolution (the data layout mentioned above is also this, which is introduced in a separate section), and then introduce what microkernel do, explaining the similarities in doing and matrix multiplication. Then, from the perspective of microkernel, describe how the optimizations are actually implemented. Then, two new algorithms are proposed, which are actually from two perspectives, one is a dedicated microkernel, and the other is packing. The paper also compares the advantages and disadvantages of these two approaches. The dedicated microkernel approach is not flexible, troublesome, not good in migration, but high in performance with no memory overhead. The packing approach is flexible and compatible, as well as potentially using existing linear algebra libraries (the authors have done experiments for it, and the effect is not good, not as good as their own, and the authors believe that this use of dimensions is too strange), but packing will have additional memory usage.

2 Feelings

This article has been seen before. It is very similar to 《zero-memory overhead direct convolution》. However, the background has been changed to ARM architecture, with optimizations that are quite close and common.