Week 1 Reading Reflection

Hao Deng

October 2023

1 Reflection

In this paper, the author proposed an im2win data transformation algorithm and an im2win-based convolution algorithm with various optimizations. After reading this paper, I got some basic ideas about the im2win transformation algorithm and the im2win-based convolution algorithm. Compared with im2col, Im2win has better data reusability and cache hit. Most of the elements loaded in the previous operation are reused in the next operation. Additionally, im2win generates a smaller tensor compared with im2col. The im2win-based convolution algorithm is implemented similarly to direct convolution but based on the im2win transformation algorithm. There are four optimization techniques to improve the performance of the im2win-based convolution algorithm.

1. Vectorization and FMA instructions
2. Loop reordering, hoist, and loop unrolling
3. Register and cache blocking
4. Parallelization

The optimized im2win-based convolution has the shortest runtime and largest GFLOPS compared with im2col-based and direction convolution. It has less data size and memory as well compared with im2col-based convulsion.

There are some terms I don't understand or forget, including channel, batch, stride, memory footprint, brandwidth overhead, SIME, MEC, FMA, NCHW, etc. As a result, I can't fully understand the formula of im2win and its logic. Additionally, I can't fully understand the optimization part, especially loop reordering, hoist, and loop unrolling. Besides the paper itself, I am also curious about the concept of convolution itself. When and why do we use convolution? What can we do with convolution?