week-19

JX-Ma

2024/12/14

1 INTORDUCE

This week's work is as follows:
This week, I separated the index pre computation and convolution computation and calculated the tflops without the index pre computation separately. found that it is still not as fast as the previous optimization. Previously, I suspected that it was because we used dynamic arrays to store our indexes, and the space of dynamic arrays is on the heap, so reading data from the heap is slower. Afterwards, I tried using static arrays to store indexes, but the speed was still slow and segment errors occurred at certain layers. My idea is that performing index precomputation on the CPU may not achieve our ideal optimization results, as the optimization of the index computation part is already sufficient after index hashing. Recalling the indexes we stored in the index precomputation section, most of the data are arithmetic sequences, and we only need to calculate their offsets.