note of ’Overview Of Tensor Layout In Modern Neural Network Accelerator’ and ’FFT Blitz: The Tensor Cores Strike Back’

zxp

November 23, 2024

1 content

These two papers are too short, so put them together, ¡ Overview Of Tensor Layout In Modern Neural Network Accelerator¿ only 4 page, ¡FFT Blitz: The Tensor Cores Strike Back¿ Only two page

1.1 ’Overview Of Tensor Layout In Modern Neural Network Accelerator’

This is a 2021 paper, the Core is to analyze the utilization of cache by different Tensor layouts. The background of the paper is GPU and im2col and Tensor Core. The paper mentions two layouts, NHWC and N(C/X)HWX. or N(C/32)HW32. The figures in the paper describing the different data layouts are close to those in the Nvidia documentation and can be found here.

1.1.1 NHWC and N(C/X)HWX

C (channel) stores first, then comes the W (width) dimension, H (height) dimension is after W , N(batch size) is stored at last; Fig.2(c) C (channel) is divided into (C/x) = (64/32) = 2 groups, each group is HWx (i.e. HW32), for HWx, x channels are stored first, then comes the W and H dimension as HWC.

1.1.2 Implementation of Profiling Framework

includes Block Splitting Engine (BSE), Tensor Data Loading Engine (TDLE) and a Hierarchy Memory Structure (HMS).

BSE:split the result Matrix C into multiple fixed size Blocks, from top to bottom and left to right.The paper gave no details

TDLE loads the raw tensor data from HMS and outputs the cache usage of the current layer, such as hit/miss, load/store, and cull counts

For HMS, the paper describes L1 and L2 caches and lru, and describes the flow of reads with caches. Here's the paper's description: "TDLE first sends request to L1 for tensor data, if L1 hit, L1 will return tensor data to TDLE directly. Otherwise, L1 will go to L2 to load tensor data, L2 will do the hit/miss test, if hit, L2 will return tensor data to L1, otherwise, L2 will load tensor data from main memory."

1.1.3 Experimental

This paper compares the cache hit and other parameters of two different data layouts. No data layout conclusions are given, just that what is presented in the paper can analyze the cache hit ratio of different data

1.2 'FFT Blitz: The Tensor Cores Strike Back'

The paper is also 21 years old. The paper mentions the Cooley-Tukey FFT, which fits nicely into the Tensor Core, on which the work is based. Cooley-Tukey FFT decomposes large DFT computations into sequences of parallel, small baseline DFT computations that can fit well into Tensor Cores.

The flow of the paper: 1.data movement and pre-computations; 2.CUDA kernel is launched which computes the column wise DFT operation using the Tensor Cores; 3.a per thread element wise complex number multiplication is done with the resultant output; 4. move towards a row based FFT computation and appropriate element wise twiddle factor computation; 5.matrix transpose to dump the final output back to global memory; 6.next batch.

The conclusion of the paper, the method of the paper is fast, did not say why

2 Feelings

These two articles are too short, some descriptions can refer to