note of 《Tensor Layout Optimization of Convolution for Inference on Digital Signal Processor》

zxp

November 16, 2024

1 content

"Tensor Layout Optimization of Convolution for Inference on Digital Signal Processor" is a paper of 2019.This paper focuses on the advantages and disadvantages of direct convolution of two different data layout (NCHW and HWCN) on Digital Signal Processor (DSP), and designs an automatic tuning method to accurately select the appropriate layout for different convolution layers in neural networks. The hybrid layout proposed in the paper is faster.

1.1 BACKGROUND

It is pointed out that the application of DSP in deep learning is mainly concentrated in the inference stage, and the real-time requirement of the inference stage is higher than that of the training stage. For embedded devices, few studies have considered how the network tensor layout affects the inference performance on digital DSP. The application 13 of the paper is to study the data layout on GPU (see later). However, the paper argues that because DSP design is more flexible than GPU, it is difficult to directly copy the performance analysis results of one DSP to another DSP. Therefore, automatic tuning is required. According to the paper, different order of tensor dimensions will be presented in different organizational forms in memory, corresponding to different layouts.

The difference in the layout will directly affect the computational efficiency of convolution.The output channel is depend on the number of the convolution kernels. The number of channels of the weight tensor is generally equal to the number of channels of the input tensor. These latitude differences result in different data layouts with different computation speeds. The tensor layout that dominates the mainstream framework is NCHW, which the W is the lowest di-

mension. This is because the NCHW tensor layout can better utilize the parallel performance of the GPUs and cudnn supports it very well.

The paper describes the two data layouts used, NCHW and HWCN. The paper has a section of different data layouts in the memory placement situation, the description is very good: (Numbers are elements in the figure in the paper) "For NCHW layout, the W dimension is the lowest dimension, arranged in the order of (1, 2, 3), (4, 5, 6) and (7, 8, 9) in the direction of W. Then the elements of the tensor are arranged in H dimension that the tuple (1, 2, 3) is placed in memory before the tuple (4, 5, 6), and the (4, 5, 6) is placed before the (7, 8, 9). Thirdly, the third arrangement dimension is C, which means that after the elements of one channel (such as the green blocks) are arranged, the elements of next channel (such as the pink blocks) will be arranged. Finally, for the N is the highest arrangement dimension of NCHW, the second kernel needs to be arranged after the arrangement of the first kernel. Similarly, HWCN layout is also arranged with the same rules which the N is the first dimension that needs to be arranged and the W is the last dimension that needs to be arranged. "the paper concludes that" It is clear that the layout of tensor and the shape of the input feature map have a very direct impact on the memory access behavior when calculating convolution."

1.2 METHODOLOGY

The experiments in this paper show that the performance of the same convolution varies greatly for different layouts, and different tensor layouts are suitable for different convolution in DSP. The strategy of the paper needs to know the layout information of the next layer in advance, and the paper points out that the layout information determined according to the next layer can directly correspond to the storage of convolution results, which can reduce the cost of conversion between different tensor layouts, and the time of transform can be almost ignored without recording.

The paper's base method for selecting data is that "for NCHW tensor layout, W is the lowest dimension, and when the filter w (or output w) is larger than the output c, it can take full advantage of the data locality of the weight and output width dimensions. For the HWCN tensor layout, N is the lowest dimension, and when the output c is greater than the output h, it can take full advantage of the data locality of the channel of the output tensor."

The paper regards the selection of data layout as a classification problem, which is not reliable to judge by experience, and it is too difficult to model the memory access mode in convolution, and the accuracy is difficult to guarantee. Therefore, the paper tests the performance of different classification algorithms, and analyzes the tim e spent by the selector to select the appropriate data layout

theory to explain why the method in the paper uses SVM.

1.3 EXPERIMENT

This paper is an experiment to compare the performance of two different data layout and mixed data layout, to prove that the effect of different data layout is different, and the effect of mixed data layout is better. We also compare the accuracy of different classifiers in selecting the optimal data layout to illustrate why SVM is selected

2 Feelings

The paper mention an uncommon data layout, HWCN. the paper chooses the appropriate data layout through selectors, but the background of the paper is DSP, and there may be different in the CPU and GPU.