World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:17, No:06, 2023

Operator Optimization Based on Hardware Architecture Alignment Requirements

Authors : Qingqing Gai, Junxing Shen, Yu Luo

Abstract : Due to the hardware architecture characteristics, some operators tend to acquire better performance if the input/output tensor dimensions are aligned to a certain minimum granularity, such as convolution and deconvolution commonly used in deep learning. Furthermore, if the requirements are not met, the general strategy is to pad with 0 to satisfy the requirements, potentially leading to the under-utilization of the hardware resources. Therefore, for the convolution and deconvolution whose input and output channels do not meet the minimum granularity alignment, we propose to transfer the W-dimensional data to the C-dimension for computation (W2C) to enable the C-dimension to meet the hardware requirements. This scheme also reduces the number of computations in the W-dimension. Although this scheme substantially increases computation, the operator's speed can improve significantly. It achieves remarkable speedups on multiple hardware accelerators, including Nvidia Tensor cores, Qualcomm digital signal processors (DSPs), and Huawei neural processing units (NPUs). All you need to do is modify the network structure and rearrange the operator weights offline without retraining. At the same time, for some operators, such as the Reducemax, we observe that transferring the Cdimensional data to the W-dimension(C2W) and replacing the Reducemax with the Maxpool can accomplish acceleration under certain circumstances.

Keywords : convolution, deconvolution, W2C, C2W, alignment, hardware accelerator
Conference Title : ICS 2023 : International Conference on Supercomputing
Conference Location : Paris, France
Conference Dates : June 22-23, 2023