《Characterizing and Demystifying the Implicit Convolution Algorithm on Commercial Matrix-Multiplication Accelerators》论文笔记

zxp

February 3, 2024

1 论文内容

《商用矩阵乘法加速器上隐式卷积算法的表征与解密》这篇论文是争对im2col的。论文指出现在的商业神经网络加速器并不只是朴素的显式的im2col转换，而是使用了并未公开的优化过的im2col方式。这种隐式的im2col算法会比显式的im2col花费更少的空间和更好的性能，这篇论文证明显示的im2col即内存效率底和速度慢，尝试对im2col进行优化达到隐式im2col的效果。论文的实验证明英伟达的gpu和TPU都没有使用显式im2col(尽管英伟达的gpu上有这个选项)，显式im2col比隐式方法平均慢28%，显式方法中的GEMM时间几乎与隐式方法相同。与GPU上的GEMM相比，显式方法引入了大约26%的性能开销。显式方法比隐式方法慢23%。

1.1 通道优先Im2col方法

论文使用了HWC布局，这种布局受到步幅的影响比较小。当stride=1时，CHW和HWC格式之间的差距很小，因为w维通常足够大，可以使用DRAM带宽，但stride比较大的时候HWC布局连续性比较好。

1.2 优化

当输入通道尺寸很小时，比如3，通道优先设计就会变得低效，论文建议用其他块(尚未计算)的数据填充向量内存，本质上是同时计算多个块。

2 心得

这篇论文是基于GPU的，优化的是im2col，GEMM使用的和其他的商用im2col实

现完全一样。论文指出商用的im2col不是朴素的显示的im2col，而是使用了未公开的优化。论文提出一种优化im2col的方式-通道优先Im2col方法。并说明了这种方式的优劣，stride比较大的时候HWC布局连续性比较好，输入通道尺寸很小时会变得低效。