《FAST CONVOLUTION KERNELS ON PASCAL GPU WITH HIGH MEMORY EFFICIENCY》阅读笔记

zxp

April 20, 2024

1 论文内容

《具有较高的内存效率的基于PASCAL gpu的快速卷积核》这篇论文提出了内存访问被优化以实现更高的性能的在GPU上的卷积方法，有单通道和多通道两个版本。论文将数据仔细划分并分配给每个SM，以隐藏全局内存的访问延迟。

1.1 论文中提到的背景

论文中提到隐式矩阵乘法，将特征映射和滤波数据分成多个子块，并将其转换成多个子矩阵，非常节约内存，非常高效被纳入Cudnn中。并且还提到两种方法（论文没具体说但引用了）比隐式矩阵乘法还高效，但是很吃input和filter的纬度，因为固定了给每个流处理器的数据量，特征映射小于32性能特别差，不适合小的特征映射，但现在CNN模型中超过一半的卷积层纬度都小于32.
对GPU的介绍，“gpu支持的片内内存包括寄存器和共享内存，片外内存主要是全局内存。寄存器是最快的，全局内存是最慢最大的。在整个计算过程中，从全局存储器到片上存储器的数据加载时间是最关键的，而隐藏全局存储器的延迟是加速最重要的一点。”

1.2 论文考虑的方式

论文说隐藏内存访问可以用两种方法，1，流水线，算数强度足够大，从内存中预取数据的时间小于计算时间。2，先从全局内存中读取大量数据（卢帅师兄的优化方式是结合了两种，不过是先读取大量数据到共享内存，隐藏的是从共享内存到寄存器的时间）。论文认为单通道卷积中计算的次数不够隐藏从内存读取的时间就用的第二种。

论文还从用的GPU的带宽出发计算了GPU的数据搬运能力，算出数据传输所需要的线程。
论文提出划分给每个流处理器的步骤：1，划分特征映射和过滤器，使分配给每个SM的数据的总大小小于共享内存的大小。2. 评估可对每个SM中的数据执行的FMA操作的次数。3，如果FMA大于搬运时间，使用第一种方法，基于数据预取。4. 如果没有，则重新划分特征映射和过滤器，使传输到所有流处理器的数据的总大小大于共享内存，使用第二种方法。
数据映射，三种情况，考虑过滤器，考虑输入张量，或者都考虑。（后面的分析指出按通道分流处理器会导致将结果加入output的时候需要访问全局内存会很慢）
论文提出了一种步进固定块方法，设置共享内存为32的倍数

2 心得

论文核心是隐藏内存访问的两种方法，对于单通道卷积用数据映射，对多通卷积用共享内存来隐藏全局内存的访问延迟