《Optimizing Depthwise Separable Convolution Operations on GPUs》阅读笔记

zxp

April 13, 2024

1 论文内容

《gpu上深度可分离卷积操作的优化》这篇论文提出了一个方法采用动态贴图大小方案来自适应地分布GPU线程的计算数据，以提高GPU利用率并隐藏内存访问延迟。

1.1 论文中的算法

论文提出两种列和行重用算法，核心是增加数据复用，对数据从全局内存搬运到共享内存的流程进行详细设计（这篇论文还有前置的论文，前置的论文也是对数据从全局内存搬运到共享内存的流程进行详细设计，不过没这篇论文的流程好）。论文提出不要按顺序帮运数据，因为第二列会用到第一列和第三列的数据，因此先搬运第一列第三列第二列的数据就可以在缓存/寄存器中找到，理论上可以实现数据只搬运一次，论文还用CUDA中的shuffle指令读取隔壁线程寄存器的数据，以此来减少数据搬运。

1.2 分块策略

用过滤器来划分，因为过滤器纬度在CNN中是固定的。当过滤器纬度太大时会将每个线程计算的通道减半，减小到不会等待数据搬运。对GPU利用率更敏感的话小批量的卷积会表现更好。

1.3 论文作者的分析

性能提升主要归功于列和行重用算法提供的内存访问次数减少。

2 心得

论文中没有写是否计算数据从CPU搬运到GPU的时间。
和GPU不同，论文中除了强调内存性能，还强调了流处理器的利用率，论文体现他提出的算法的优势除了缓存和寄存器利用得好还说论文提出的算法分块分的好，流处理器利用率高。
感觉GPU的核心是对共享内存的利用和使用合适的分块分线程策略来对提高流处理器的运用。