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

zxp

April 6, 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

论文的核心似乎是增加数据复用，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。

2 心得

还未仔细看完。论文中没有写是否计算数据搬运的时间。
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。