《Computing Large 2D Convolutions on GPU Efficiently with the im2tensor Algorithm》阅读笔记

zxp

March 2, 2024

1 论文内容

《利用im2tensor算法在GPU上高效地计算大型2D卷积》这篇论文是针对GPU的，提出来一个名为im2tensor的算法。但主要是在介绍GPU，还用了一节介绍怎么使用roolfine。

1.1 GPU

CUDA让GPU拥有处理常规计算的能力，CUDA编程基于SIMT（单指令，多线程），使用时编写单个线程，指定运行的线程数，线程被划分成可指定大小的线程块。GPU处理矩阵乘法特别迅速，现在卷积大部分依赖于卷积乘法。
近年的GPU中引入了张量核心，这是种专门用于矩阵乘法的额外硬件。

1.2 im2tensor

论文关于提出的算法描述的并不多，描述中是按张量的对角线分开分别计算然后求和，论文认为这样会更好的使用GPU的用于矩阵计算的核心。论文指出填充可以让输入张量获得最佳的纬度。
论文还通过计算复杂度的方式得出输入和输出张量的通道应该尽可能的小。

1.3 roofline

论文给了用roofline的方向，要性能好计算强度要大。要增加算数强度需要增加计算数量又要尽量减少数据搬运，得读取输入张量更多的行。

2 心得

这篇论文是基于GPU的，主要在介绍GPU编程，论文提出的im2tensor算法就只提了按对角线分块，关于roofline的部分主要说了增大算数强度，大概意可能是增加计算数量需要读取更多的数据，尽可能少的数据算更多次，需要多重复使用核的数据。