Computing Large 2D Convolutions on GPU Efficiently with the im2tensor Algorithm

JX-Ma

2024/3/1

1 笔记

本文提出了一种新的 2D 卷积算法 Im2tensor, 介绍了算法的独特之处，即使只有一个核，它也能表现出矩阵-矩阵乘法。文章前面提出了一些基本卷积的定义，后面介绍了一下 GPU 编程，gpu 最初是为了在计算机屏幕上高效地生产和显示图像而设计的。他们首先通过光栅化和像素着色等硬件固定功能来实现这一目标。随着人们对这种强大处理器的兴趣不断增长，gpu 变得更加灵活，对一般计算也更加开放。2007 年，英伟达发布了 CUDA 语言，使 gpu 成为一个方便的计算平台。这也使得卷积也可以在 GPU 平台上进行，从而在很大程度上加快卷积的速度。CUDA 主要针对于并行计算，我们在编写程序时需要指定线程数，线程也能被分为很多个线程块。在 cpu 上实现并行化编程只需要指定线程数。张量核心可以用于计算矩阵的乘加运算，在 cpu 上实现的 simd 并行计算只支持向量的乘加操作，这也意味着在 GPU 上实现分块并不是在一维的角度上进行分块，而是在二维空间上分块。
im2tensor 算法，把卷积核按照对角线分为不同的子张量，然后按着对角线并行化处理这些卷积后的结果。之后介绍了对于输入张量维度不同，在使用填充时需要找到合适的填充大小，这里计算算术强度是并没有把写入内存计入算术强度的计算中。之后在对实验结果进行分析，并对算法的准确性进行评估。

2 总结

本文主要讲的是在 GPU 上的优化，在 GPU 上实现编程的方式和 cpu 上有很大的差别，GPU 上更适用于并行化编程，对于线程也能在分为线程块，在 cpu 上最多实现向量-向量乘加，而 GPU 上则可以实现矩阵-矩阵乘加。对于在 cpu 还是 gpu，都可以借助 roofline 模型来达到最优的性能。