《Triton: An Intermediate Language and Compiler for Tiled Neural Network Computations》阅读笔记

zxp

March 30, 2024

1 论文内容

《Triton:用于平铺神经网络计算的中间语言和编译器》这篇论文提出了Triton，一个自动优化卷积操作的编译器。基于C语言和LLVM。提出这个的原因是供应商库（cuBLAS，cuDDNN）对新提出的模型支持不够好，并且受硬件限制。专门优化的卷积操作可移值性低，并且手写微内核需要大量体力劳动。

1.1

使用论文提出的编译器需要使用论文中提出的Triton-c语言。Triton-IR会分析程序，和TVM编译器一样Triton-JIT会自动调优，生成高效的机器码。并且有很好的扩张性，可以适应各种卷积方法。

1.2 Triton-JIT会做的优化

1.2.1 数据预取

循环内部的微内核内存操作可能是有问题的，因为它们可能导致严重的延迟，在缺乏足够的独立指令时无法隐藏。
分块，可能紧密地贴合机器的计算能力和存储层次。Triton-IR的结构使其能够自动枚举和优化任何可表达程序的有效嵌套平铺配置。

1.2.2 合并内存访问

相邻的线程同时访问附近的内存位置时，内存访问被称为合并，因为内存通常是以大块的形式从DRAM中检索的。Triton-IR会排序微内核，让内存访问接近的微内核先后执行，减少内存加载次数。（就是增加访问内存的连续性吧）

1.2.3 参数调节

传统的自动调优器通常依赖手写的参数化代码模板，以在预定义的工作负载上实现良好的性能。Triton-JIT可以通过简单地连接与上述每个优化过程相关的元参数，直接从Triton-IR程序中提取优化空间。

2 心得

这篇论文是提出了一个编译器，能够自动优化卷积操作，论文结构是先描述编译器怎么使用，然后描述编译器做了什么，特别是Triton-JIT做的优化，最后是实验结果和分析。