《YFlows: Systematic Datafow Exploration and Code Generation for Efcient Neural Network Inference using SIMD Architectures on CPUs》阅读笔记.tex

zxp

March 16, 2024

1 论文内容

《YFlows:在cpu上使用SIMD架构进行高效神经网络推理的系统数据流探索和代码生成》这篇论文是一遍6页的会议，在CPU平台优化直接卷积，针对的是SIMD，论文提出一个工具自动生成使用了SIMD的代码的工具（主要看论文怎么使用SIMD）。

1.1 论文内容

论文首先提出在代码中实现SIMD优化的必要性，虽然编译器优化选项能使用SIMD自动矢量化，但SIMD优化过于复杂需要程序员手动书写。
论文使用的数据布局是常见的NCHW的变种，NCHW[xc]，数据分成大小为X×H×W的块，X为物理矢量寄存器的倍数，然后对通道矢量化（常见方案）。论文指出这种布局比起BHWC数据布局有数据重用的可能。
论文指出如何分配向量急寄存器取决于向量寄存器的个数和数据重用的机会（数据重用减少数据移动成本，可以利用数据布局的优势）.
论文也指出步长对数据重用带来的影响，步长越大数据重用越少。
论文提到数据流，不知道具体怎么做的，但指出将output的元素算完再写入会更好，虽然会使用很多向量寄存器但节省了在每次计算完成后对标量变量进行累加的时间。把加载到向量寄存器里的数据所参与的所有运算一次算完会更好，论文指出复用的寄存器中应该是过滤器的数据。

1.2 实验

论文用TVM做对比，使用的数据类型是int8。TVM是高度优化的高效的神经网络部署框架。