《Efficient and portable GEMM-based convolution operators for deep neural network training on multicore processors》阅读笔记

zxp

March 9, 2024

1 论文内容

《用于多核处理器上深度神经网络训练的高效且可移植的基于GEMM的卷积算子》这篇论文是针对CPU平台，选择CPU的原因是高性能CPU空闲的时候也可以用来训练，工作是优化im2col，论文将变换集成到CEMM中减少了内存使用，为了实现算子还需要索引重新映射Reindex变换，并且指出论文提出的方案有很好的移值性（因为代码大部分用C写的，没用多少汇编）。

1.1 融合im2col变换的GEMM和索引重新映射

论文的核心是将im2col变换集成到GEMM中，即通过GEMM后得到的是im2col变换后的张量而不是常规布局的输出张量，将im2col转换的成本隐藏在GEMM实现中，这样可以省去部分下一层的im2col变换的内存开销，并且不需要对GEMM进行额外优化。靠GEMM后重新打包和索引重新映射确定元素位置。

1.2 实验

这篇论文的工作和其他的优化卷积的工作还有个不同是不只是单纯的应用在推理阶段，还能应用在训练阶段，论文实现了反向传播和权重更新。
论文将优化后的卷积集成到PyDTNN，论文作者认为这个框架提供了一个更易于访问和定制的解决方案。内存节省在24%到70%之间。

2 心得

这篇论文是优化im2col，将im2col转换的成本隐藏在GEMM实现中来减少内存开销。