《Optimizing Memory Efficiency for Deep Convolutional Neural Networks on GPUs》阅读笔记

zxp

June 1, 2024

1 论文内容

《优化GPU 上深度卷积神经网络的内存效率》这篇论文首先指出卷积层内存效率的重要性，数据布局是影响性能的重要原因，这篇文章是研究不同数据布局在GPU上的表现。论文中的cuda-convnet选择CHWN数据布局，并使用直接卷积

1.1 论文提到的数据布局

1.1.1 批次作为最低纬度

论文将批次放在最里面的原因是批次通常为16的倍数选择有限，而GPU上的每个线程块为32个线程，比较适合线程块访问内存。论文指出批次作为最低维度有两种，chwn和hwcn，因为影响最大的是最内层，批次聚合在一起，其余纬度的数据重用一样保留了，所以这两种布局表现差不多。
N作为最低纬度受到N的大小影响明显，当批量大小N超过64时性能比NCHW更优，当c小于32的时候CHWN更优，论文指出有其他论文的工作证明通道大的话矩阵乘法的数据重用更对（这样说的话，论文这里写的c指的应该是output的通道，不同通道依赖的是不同过滤器的批次，通道大input的元素重复使用的多，在矩阵乘法中也是这样，转换后的input矩阵被重复使用的次数多）subsubsectionNCHW和NHWC Caffe和cuDNN使用的策略，NCHW为其中矩阵乘法和FFT的默认数据布局，矩阵乘法几乎使用任何尺寸的输入。FFT卷积通用性就很低，一些奇怪的维度数（过滤器过小）上会产生很大的内存开销，将纬度裁成32x32产生的额外开销可能会大于算法的优势。在FFT中CHWN和在直接卷积一样部分有优势。

1.2 论文作者的结论

数据布局对性能的影响很大，论文还提出了自动选择合适的数据布局的方式。

2 心得

论文中提出在GPU上CHWN在部分情况下比NCHW有优势