《Optimizing Data Layout for Training Deep Neural Networks》阅读笔记

zxp

June 15, 2024

1 论文内容

《优化训练深度神经网络的数据布局》这篇论文核心工作是提供一个仲裁框架会根据训练时提供的参数决定使用数据布局

1.1 背景

论文指出数据布局会影响训练的性能，论文提到两种数据布局NCHW和NHWC，一般认为NHWC在cpu因为适合矢量化会快一些。论文聚焦的是模型裁剪，模型裁剪后部分数据会用不到，但这些被跳过的数据还是会加载到缓存中，会引起缓存抖动，并且这对不同的数据布局影响不一样。裁剪有多种方式，不同的方式裁剪不同的纬度，裁剪宽高通道或者批次，不同裁剪方式对不同数据布局影响也不同。

1.2 论文的方案

论文给了缓存模型用来计算内存访问次数（详细见论文3.2），论文还强调不不同数据布局不影响训练精度。论文还指出不论用什么数据布局都无法消除cache中没用的部分。

2 心得

论文基于数据布局和模型裁剪，和我们相关的部分就是数据布局很重要。