卷积神经网络剪枝方法综述

1. Paper

PCONV: The Missing but Desirable Sparsity in DNN Weight Pruning for Real-Time Execution on Mobile Devices.

Statistics

Xiaolong Ma

Northeastern University.

* 1. 引言与摘要

DNN剪枝方法主要有两种：non-structured fine-grained和structured coarse-grained。细粒度剪枝稀疏度更高而且精度保持的很好，但是对硬件不友好；结构化剪枝对硬件友好但是当剪枝率高时精度损失很大。文章提出了PCONV方法，对两种粒度的剪枝做了妥协：在粗粒度剪枝的框架内用了细粒度剪枝的模式。Sparse Convolution Patterns (SCP) kernel内进行剪枝，每个filter的稀疏度都保持均衡并且很好的保持了精度。

介绍了几个人的几篇论文，仍然是一些经典的方法和套路。

提出了需求和motivation: 现在缺少的因素在于1）找到一种新的稀疏维度使得模型既可以得到细粒度剪枝的精度优势和粗粒度剪枝的规则化优势; 2) 构建对应的“算法-编译器-硬件”框架，提高硬件效率。

