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。

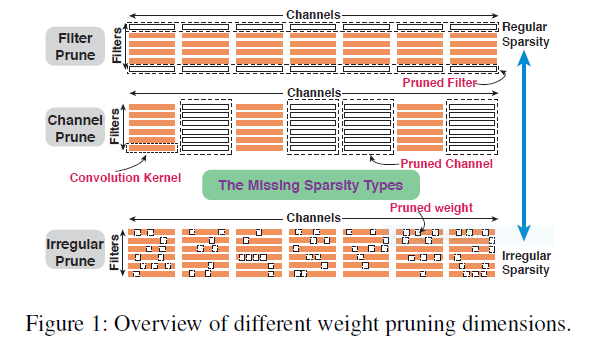
提出了需求和motivation: 现在缺少的因素在于1）找到一种新的稀疏维度使得模型既可以得到细粒度剪枝的精度优势和粗粒度剪枝的规则化优势; 2) 构建对应的“算法-编译器-硬件”框架，提高硬件效率。

PCONV主要由intra-convolution kernel剪枝（pattern pruning）和inter-convolution kernel剪枝（connectivity pruning）构成。Pattern pruning中每个卷积核中被剪掉的权重是固定的，但是被剪掉的位置（或者说形状）是不同的。好处在于控制了每个卷积核中的压缩率都是相同的，同时控制剪枝形状的种类，使其不过于复杂。

Inter-convolution kernel剪枝（connectivity pruning）在某种角度上可以被看做是剪掉了特定的卷积核，缩短了滤波器的长度。卷积核（比如拉普拉斯卷积核和高斯滤波器）在计算机视觉任务中有重要的作用，以卷积核为尺度进行剪枝也是精度得到保证的重要原因。

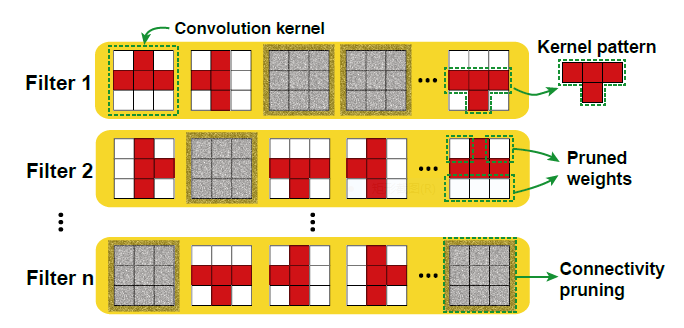
* 1. 背景

DNN模型压缩可以让模型拥有更少的参数、更小的参数体积和更快的运算速度。剪枝方法一般分为结构化剪枝和非结构化剪枝，其中结构化剪枝又可以分为通道剪枝和滤波器剪枝。



非结构化剪枝的模型压缩率有限而且一般来说不统一，但是可以保证很高的精度。然而非结构化剪枝由于其不规则的特性，对计算机硬件带来了很多障碍：1. 因为需要存储稀疏格式下的索引数据，程序在运行时代码中会嵌入很多控制和跳转指令，导致该方法很难做到指令集并行，降低了模型可以被加速和优化的空间；2. 不规则的数据分布带来了不规则的访存行为，影响了数据通路的性能表现；3. 稀疏度不均衡也会带来不同处理器、不同运算部件的负载不均。

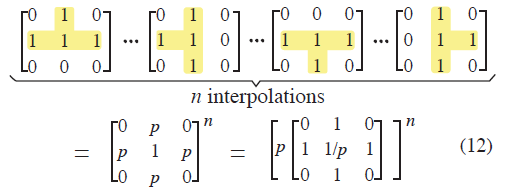
结构化剪枝用于解决非结构化剪枝中负载不均衡和系数格式导致的索引空间过大的问题。该方法可以生成更规则和更好的权重矩阵，减少了权重矩阵中的索引空间，在GPU/CPU上拥有更高的并行度和更好的性能。但是该方法会显著降低模型精度。



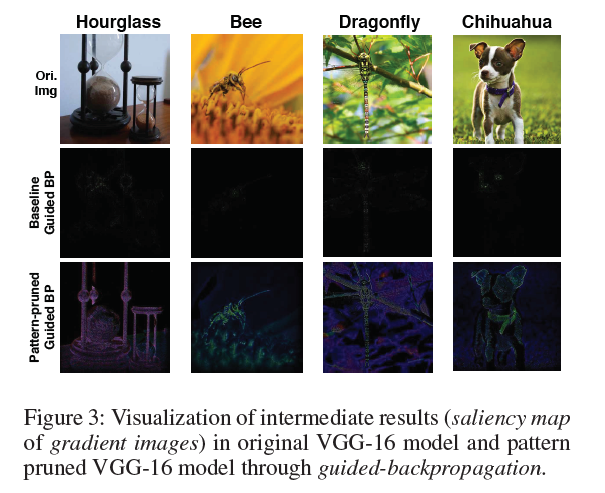
在计算机视觉领域中，特定模式的卷积核模式往往有不同的作用，比如高斯滤波器等。通过对权重进行模式化剪枝，可以使其有不同的作用。Connectivity pruning则是直接剪掉了滤波器中的一些卷积核，破坏了输入通道和输出通道之间的原有连接，缩短了滤波器的长度。

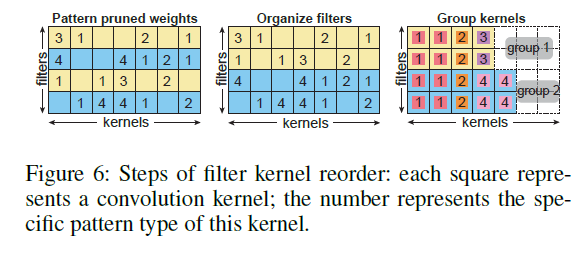
* 1. motivation

稀疏卷积核的模式本身就存在发掘图像中有效信息的潜力，所以根据数学理论设计了几种卷积核模式。

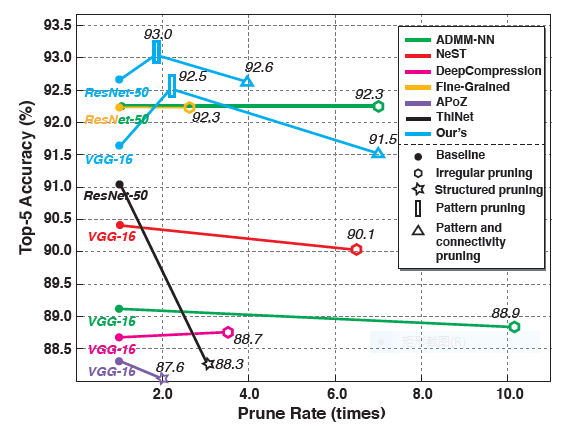


如图展示的高斯滤波器，在卷积核若干次的累加后形成的效果。





使用编译器对已经被剪枝处理过的参数进行重排序处理，使其可以规整的被并行计算。



1. paper

An Efficient Hardware Accelerator for Structured Sparse Convolutional Neural Networks on FPGAs.

IEEE TRANSCATIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS 2020

Chaoyang Zhu

Zhejiang U

* 1. 引言和摘要

使用FPGA做稀疏网络的计算加速，通过跳过无效的0数据计算来降低功耗，通过数据重用来降低对访存的需求。相比于之前的FPGA加速器取得了很好的加速比。

目前CNN网络在部署时面临很严重的问题：在CPU上运行速度过慢、效率不高；在GPU上运行功耗过大，使用场景有限。因此越来越多的CNN模型选择在FPGA或其他硬件平台上上部署。

剪枝会带来很多零元素，进而会导致很多无效计算。这些无效计算一方面是增加了计算的功耗，另一方面影响了计算效率拖慢了运行速度。一些人完成了在FPGA上设计稀疏矩阵乘器件的工作，极大的提高了计算效率，但是由于一些剪枝方法在数据的稀疏分布上并不均匀，因此稀疏矩阵乘法部件又面临着负载不均衡、资源利用率较低的问题。