International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01854-2

Generalized Gradient Flow Based Saliency for Pruning Deep Convolutional Neural Networks

Xinyu Liu[1] · **Baopu Li**[2] · **Zhen Chen**[3] · **Yixuan Yuan**[1]

Received: 10 June 2022 / Accepted: 12 July 2023

Abstract
Model filter pruning has shown efficiency in compressing deep convolutional neural networks by removing unimportant filters without sacrificing the performance. However, most existing criteria are empirical, and overlook the relationship between channel saliencies and the non-linear activation functions within the networks. To address these problems, we propose a novel channel pruning method coined gradient flow based saliency (GFBS). Instead of relying on the magnitudes of the entire feature maps, GFBS evaluates the channel saliencies from the gradient flow perspective and only requires the information in normalization and activation layers. Concretely, we first integrate the effects of normalization and ReLU activation layers into convolutional layers based on Taylor expansion. Then, through backpropagation, the derived channel saliency of each layer is indicated by of the first-order Taylor polynomial of the scaling parameter and the signed shifting parameter in the normalization layers. To validate the efficiency and generalization ability of GFBS, we conduct extensive experiments on various tasks, including image classification (CIFAR, ImageNet), image denoising, object detection, and 3D object classification. GFBS could feasibly cooperate with the baseline networks and compress them with only negligible performance drop. Moreover, we extended our method to pruning scratch networks and GFBS is capable to identify subnetworks with comparable performance with the baseline model at an early training stage. Our code has been released at https://github.com/CUHK-AIM-Group/GFBS.

Keywords Gradient flow · Model pruning · Network architecture · Normalization · Image classification

1 Introduction

Building on top of the state-of-the-art performance of deep convolutional neural networks (DCNNs) in computer vision, researchers have been focusing on deploying them in resource-constrained environments, such as mobile devices and autonomous drones. Yet, the deployment tends to be hampered by large storage and heavy computation required by DCNNs (Cai et al., 2018). Therefore, it is of vital importance to obtain DCNNs with small model sizes and low run-time computation costs. To address the problem, channel pruning methods, also known as filter pruning methods, is one of the predominant and effective approaches for compressing the large DCNNs (LeCun et al., 1989; Jaderberg et al., 2014; Li et al., 2016). Existing channel pruning methods often start with a well trained large-scale DCNN, then use a certain criterion to identify the least important channels and remove them. These methods aim to minimize the accuracy drop of the original network to the greatest extent, and can enjoy the benefits of existing hardware or Basic Linear Algebra Subprograms (BLAS) libraries for practical memory saving and inference acceleration.

Although many channel pruning approaches were proposed and achieved remarkable compression performance,

Communicated by Arun Mallya.

Xinyu Liu and Baopu Li have contributed equally to this work.

✉ Yixuan Yuan
yxyuan@ee.cuhk.edu.hk

Xinyu Liu
xinyuliu@link.cuhk.edu.hk

Zhen Chen
zhen.chen@cair-cas.org.hk

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China

[2] Oracle Cloud Infrastructure (OCI), Redwood City, USA

[3] Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China

Published online: 02 August 2023