ID: 0421312032

A Review on
“Incremental Filter Pruning via Random Walk for Accelerating Deep Convolutional Neural Networks”

Introduction

Accelerating Deep Convolutional Neural Networks (CNNs) has become an emerging research field. Among the different ways proposed for accelerating networks, filter pruning has been considered as a potential solution because of its advantage in tremendous speedup and memory reduction for the network model as well as an intermediate feature map.

Problem of Statement

Previous filter pruning research has made significant progress in compressing and speeding up Deep CNNs. However, they continue to face the problem of model capacity reduction due to the following factors.

1. **Violent pruning:**
   Previous research works have used a violent technique in which all filters are pruned simultaneously, resulting in significant accuracy loss over Deep CNNs, particularly when applied to the Image-Net large scale classification datasets.
2. **Filter degradation:**
   Previous research simply changed the pruned filter to zero and then retrained it in a different way, which caused the loss of filter learning ability. The activation function (ReLU) is not activated. As a result, the model capacity is decreased, which negatively impacts the performance.

It must be referenced that a less complex and more accurate attainable solution is still required for welcoming the robust pattern recognition of Deep CNNs.

Objectives

The research paper's major objectives are to use an incremental method for solving the problem of violent pruning and to use a random walk for solving the problem of filter degradation.

The Proposed Method

In this paper, a novel filter pruning method called Incremental Filter Pruning through Random Walk (IFPRW) is proposed. The IFPRW employs an incremental filter pruning approach shown in Figure1. The incremental of IFPRW is carried out through pruning a percent of filters every time. After that, a "Fixed Training" operation is performed which involves retraining the remaining filters while keeping the trimmed filters fixed.

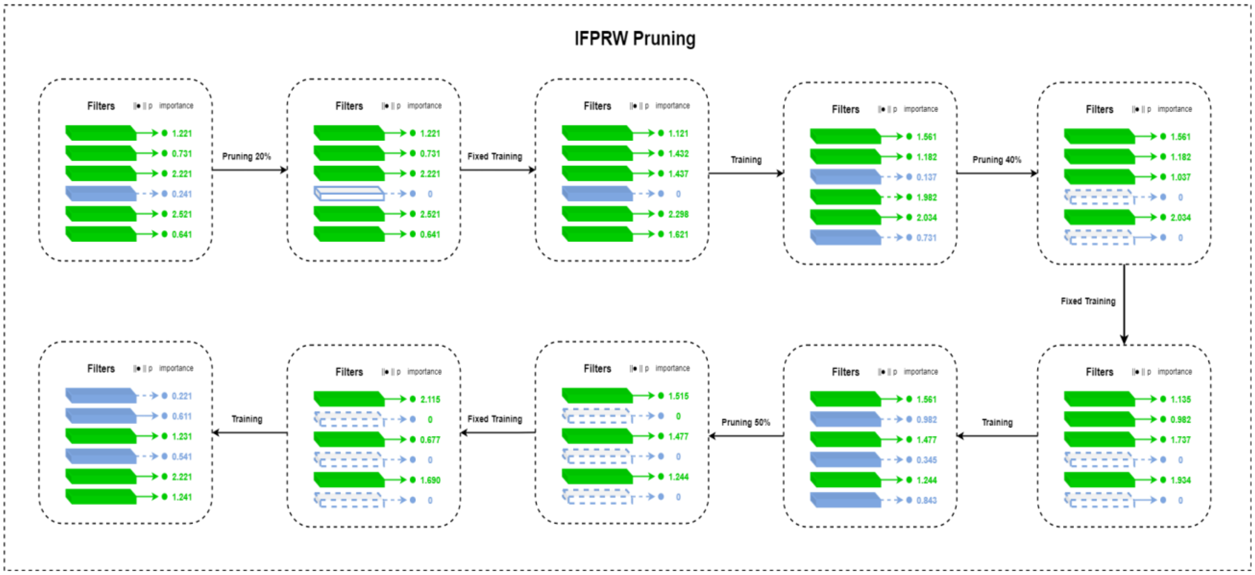

Figure 1. IFPRW Pruning on Deep CNNs model

The incremental method makes use of the remaining filters to compensate for the accuracy loss caused by the pruned filters. It allows the compressed Deep CNNs to have a wider model capacity. The random walk of the proposed IFPRW is accomplished by employing random walk on the pruned filter which permits the degraded filter to be activated with a specific probability that can enlarge the optimization space and consequently obtain better accuracy than others.

Outcome and Contribution

This research work validates the performance of the acceleration approach on 2 benchmark datasets (CIFAR-10 and ILSVRC-2012) between Soft Filter Pruning (SFP) and IFPRW. On ResNet-110, IFPRW decreases FLOPs by more than 46% for the CIFAR-10 dataset, with a 0.28 percent relative accuracy improvement. Furthermore, on ResNet-101, IFPRW eliminates more than 54 percent FLOPs for the ILSVRC-2012 dataset with just a 0.7 percent top-5 accuracy drop, demonstrating that IFPRW surpasses current filter pruning algorithms and generate a more compact model.

Hence, the large-scale experimental analysis of the proposed IFPRW verifies that in comparison to violent pruning, incremental pruning has a higher model capacity and so performs better. This also proves that its random walk technique can solve the filter degradation problem while increasing overall model capacity indirectly. These are the main contributions of the study paper.