FILTER PRUNING FOR COMPACT AND EFFICIENT CONVOLUTIONAL NEURAL NETWORKS

MILTON MONDAL

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

FEBRUARY 2023

ABSTRACT

Convolutional neural networks (CNNs) have become the primary tool for solving most computer vision tasks. The networks have become deeper over time to extract more abstract features for improving generalization performance. Currently, workstations and GPU servers only can use these deep CNNs due to their capacity to store a large number of parameters and execute a high number of floating point operations (FLOPs). The main objective of this thesis is to design compact and efficient CNNs. The first question that comes to our mind is whether all filters present in a network are essential for a task or not, or if they are important, are their importance the same or different? To address these questions, we observe the performance of the networks after pruning filters from them. We find that with retraining the pruned model can recover the performance degradation. A lot of filter-pruning algorithms have been proposed in recent years. However, the literature lacks a clear understanding of whether we should prune a filter at a uniform rate or non-uniformly from each layer. To find the answer, we conduct several experiments while pruning the same percent filters from different layers. We observe that the performance drop in each case is different, demonstrating that different layers are sensitive to filter pruning differently. We also discover that existing methods manually set the pruning fraction per layer by observing the layer-wise sensitivity for each layer. However, this manual task of determining the pruning fraction for each layer becomes extremely tedious for very deep models. We solve this problem by designing the filter importance in such a manner that they become comparable across the layers.

We also eliminate some of the major problems of existing filter pruning methods. A filter pruning method can be either feature-dependent or feature-independent. Existing feature-dependent methods do not consider the class label of a training example when calculating filter importance using feature map activations. However, our observations indicate that a feature can have high intensity for one class but not for others, indicating that the feature is useful for that class. Our proposed method, Global Filter Importance based Adaptive Pruning (GFI-AP) resolves this issue by defining filter importance based on the strength of class-specific feature maps. GFI-AP outperforms feature-dependent methods that do not consider class-specific feature map strength. Similarly, the problem in existing feature-independent filter pruning methods is that they determine the filter index that needs to be pruned based only on that filter weight. However, when we prune

a filter from a network, the corresponding channel of all filters in the next layer is also eliminated to maintain structural consistency. Thus, unlike existing methods, our feature-independent pruning method, Filter Pruning by Successive Layers (FPSL) prunes a filter through successive layers analysis. FPSL removes a filter from a layer so that the feature maps of the subsequent layer remain close to their unpruned state. Additionally, FPSL eliminates the need for layer-by-layer retraining, manual per-layer pruning percentage selection, and intensive hyperparameter search for finetuning. In a CNN, there is always a possibility of redundancy in the number of filters used as we do not ensure that different filters should pick up different information of the input. Thus, pruning helps in eliminating those unimportant filters from the base model. Here we propose an alternate approach Multi-band CNN (M-CNN) which modifies the architecture design so that different filters capture complementary bands of the input signal. It is an attempt to prevent the generation of redundant filters in the base model. Incorporating multi-band filtering into CNN has provided a novel means of building a compact and efficient network. The proposed M-CNN maintains the model performance but reduces the number of filters that needs to be trained by a factor of four for the first or last convolutional layer. The advancements and improvements achieved by our proposed methods over existing approaches would enable the deployment of compact and efficient deep learning models in edge devices, allowing them to be more beneficial for vision and other related tasks.