Learning to Prune in Training via Dynamic Channel Propagation

Shen, Shibo
Zhejiang University

Introduction: In this paper, we propose a novel network training mechanism called "**dynamic channel propagation**" to prune the neural networks during the training period. In particular, we pick up a specific group of channels in each convolutional layer to participate in the forward propagation in training time according to the significance level of channel, which is defined as channel utility. The utility values with respect to all selected channels are updated simultaneously with the error back-propagation process and will adaptively change. Furthermore, when the training ends, channels with high utility values are retained whereas those with low utility values are discarded. Hence, our proposed scheme trains and prunes neural networks simultaneously.

Key words: deep learning compression, channel pruning, dynamic channel propagation

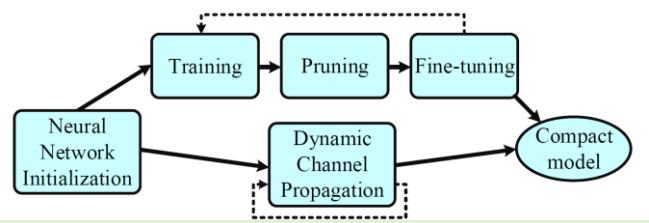


Fig. 1: The comparison of traditional pruning procedure and our proposed approach. The dash lines represent iterative pruning to achieve a high compression rate.

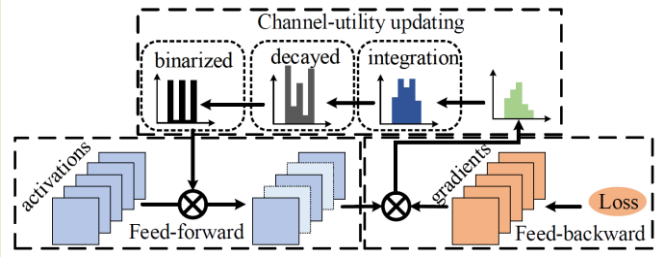
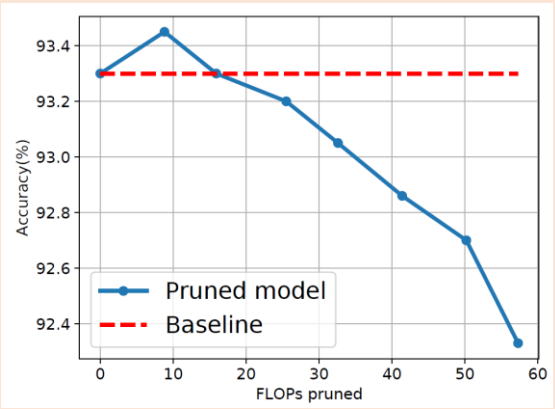
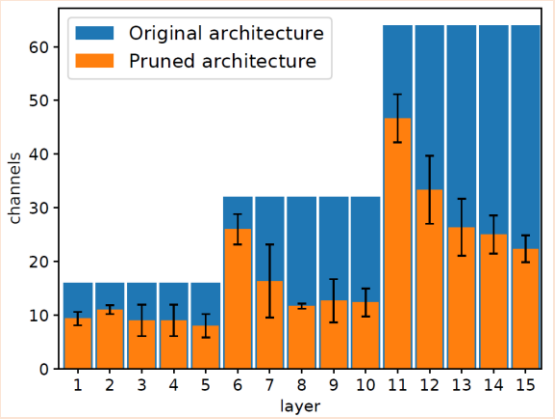


Fig. 2: A high level view of the proposed training process, which contains channel selection and utility updating in the forward and backward pass, respectively.

Data set	Architecture	Top-1	Top-5	FLOPs pruned
CIFAR-10	VGG-16	93.50%	-	73.3%
CIFAR-10	ResNet-32	92.60%	-	50.2%
ImageNet	ResNet-50	74.25%	92.05%	41.1%





Experimental results: (a) The accuracies with respect to different datasets and network architectures (b) The pruned results with respect to various pruning rates (c) The channel distribution of the pruned model for ResNet architectures. Our code could be found at *https://github.com/shibo-shen/Dynamic-Channel-Propagation*

Conclusion: We present a novel neural network training algorithm that picks up the most important channels from the redundant convolutional layers and simplify the general channel pruning procedure by selectively adjusting the parameters of the critical kernel filters in the training phase. Our scheme is light-weight and can be easily incorporated into traditional training process of neural networks. In addition, the proposed scheme can determine a good sub-architecture as long as given a compression rate.