Softer Pruning, Incremental Regularization

Linhang Cai*, Zhulin An†, Chuanguang Yang† and Yongjun Xu*

‘Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
‘University of Chinese Academy of Sciences, Beijing, China

Introduction

Network pruning is widely used to compress Deep Neural Networks (DNNs). The Soft Filter Pruning (SFP) method zeroes the pruned filters during training while updating them in the next training epoch. Thus the trained information of the pruned filters is completely dropped.

A drawback of SFP is that there is a severe accuracy drop after pruning in case of large pruning rates. Asymptotic Soft Filter Pruning (ASFP) is a variant of SFP to gradually increases the pruning rate towards the objective pruning rate to reduce the information loss caused by setting pruned filters to zeros while pruning.

Objectives

To utilize the trained pruned filters, we proposed a SofteR Filter Pruning (SRFP) method and its variant, Asymptotic SofteR Filter Pruning (ASRFP), simply decaying the pruned weights with a monotonic decreasing parameter.

Methods

A typical three-step pruning pipeline of three phases: training, pruning and fine-tuning. Our SRFP or ASRFP is used in the pruning phase to remove those filters chosen to be pruned smoothly using weights that gradually decay to zero, while the conventional pruning operation simply sets pruned filters to zeros.

Results

Different Test Accuracy Drops of ResNet-34 on ILSVRC-2012 among SFP/ASFP/SRFP/ASRFP with the training epochs increasing when the pruning rate is 30%.

Conclusion

We propose a pruning method SRFP and its variant ASRFP, softening the pruning operation of SFP and ASFP.

Our methods perform well across various networks, datasets and pruning rates, also transferable to weight pruning.

In theory, our methods do the L2-norm regularization on pruned nodes.

SRFP, ASRFP and ASFP pursue better results while slowing down the speed of convergence.

References
