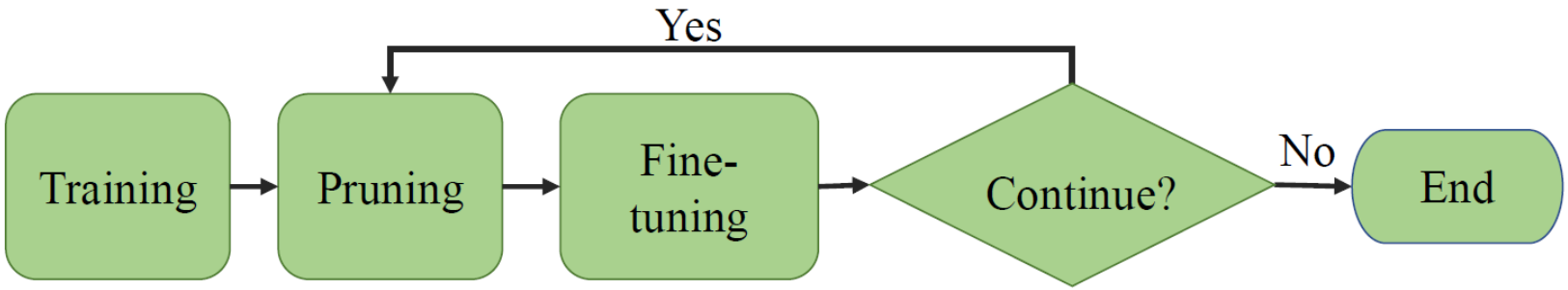
Softer Pruning, Incremental Regularization

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

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

Motivation

- Network pruning is widely used to compress Deep Neural Networks (DNNs).
- A typical three-step pruning pipeline of three phases: training, pruning and fine-tuning.

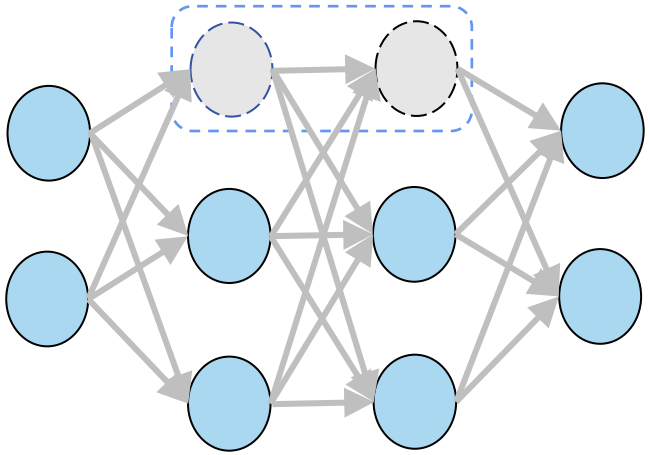


Motivation

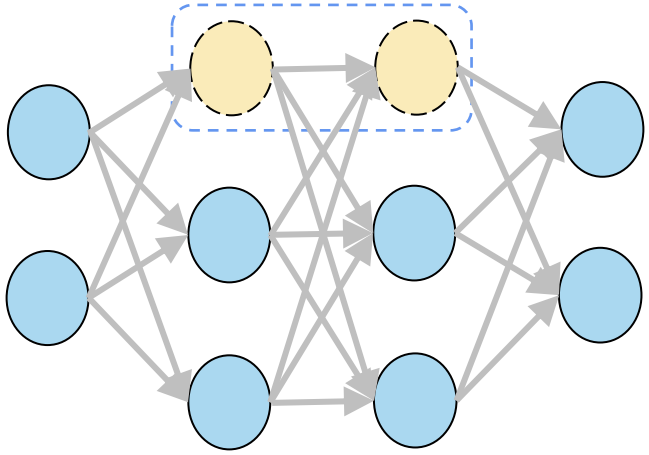
Hard Filter Pruning(HFP) method sets pruned filters to zeros and stops updating their parameters.

Soft Filter Pruning (SFP) method zeroizes the pruned filters during training while updating them in the next training epoch.

Hard pruned nodes don't update



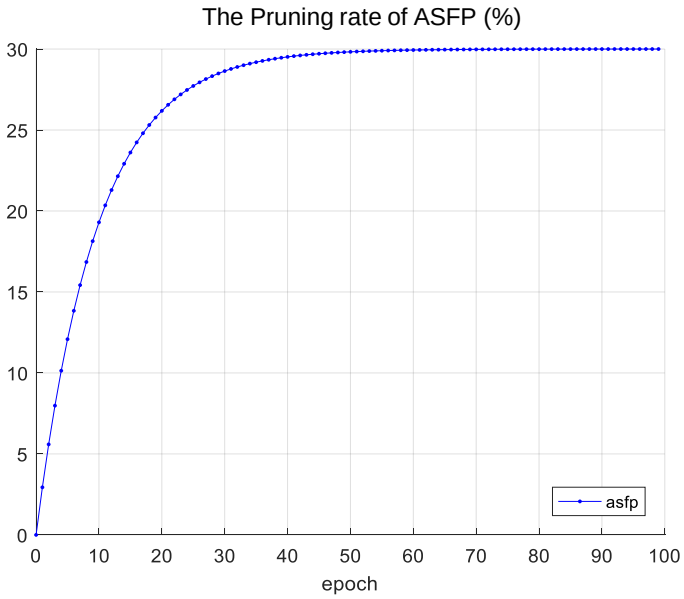
Softly pruned nodes can update



Motivation

A **drawback** of SFP is that there is a severe accuracy drop after pruning in case of large pruning rates, because the trained information of the pruned filters is completely dropped.

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.



The target pruning rate is 30%

Objective

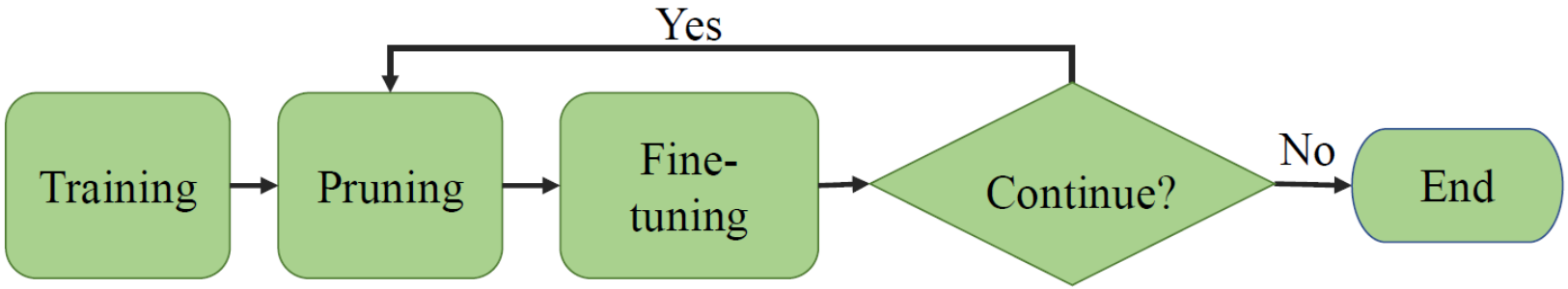
To utilize the trained pruned filters, we proposed **a SofteR Filter Pruning (SRFP)** method and its variant, Asymptotic SofteR Filter Pruning (ASRFP), simply decaying the pruned weights with a monotonic decreasing parameter.

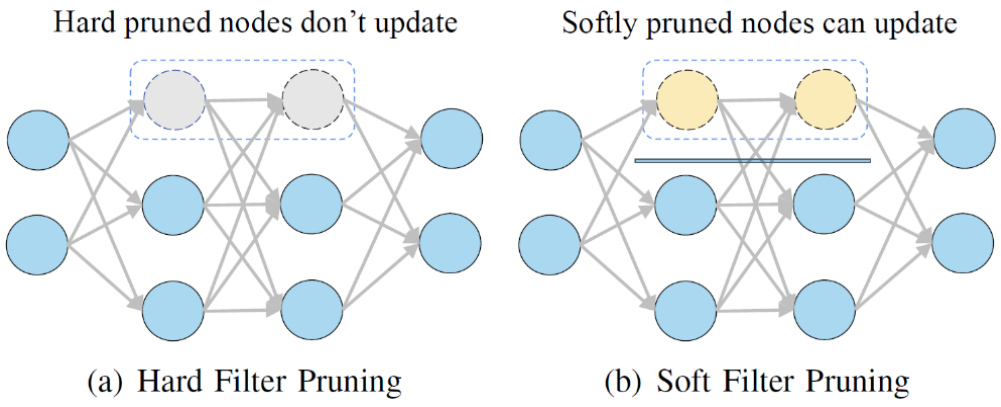
Algorithm 1: SRFP Algorithm

inputs : training set: X, pruning rate: P_i, initial decay rate: α, the model with parameters $W = \{W_i, 0 \leq i \leq L\}$.
output: The pruned model with parameters $W^* = W^{t\_max}$
Initialize the model parameter W^0
for $t = 0, ..., t\_max - 1$ **do**
  Decrease weight decay rate α
  Train model parameters $\hat{W}^{t+1}$ based on data set X and W^t
  for $i = 1, ..., L$ **do**
    Compute the ℓ_2-norm of each filter $\|\hat{W}_{i,j}^{t+1}\|_2, 1 \leq j \leq n$
    Select $n \times P_i$ filters with minimal ℓ_2-norm values
    Decay the parameters of chosen filters with α
  Get the softly pruned model parameters W^{t+1} based on $\hat{W}^{t+1}$
Get the pruned model with final parameters $W^* = W^{t\_max}$

Method

- 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.





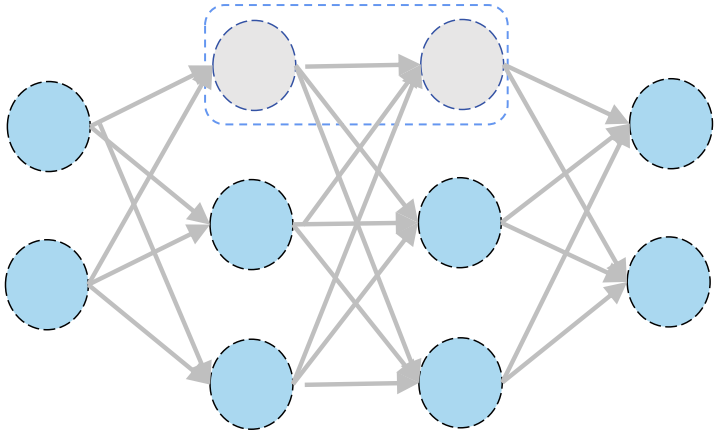
- **Comparison of hard filter pruning and soft filter pruning. The gray nodes removed by hard filter pruning could not update in the next training epoch, while the yellow nodes pruned by soft filter pruning method can still update.**

Method

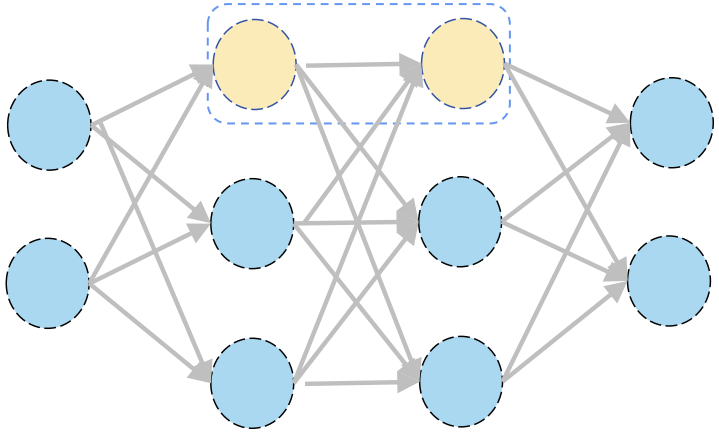
Decaying Parameter α of pruned nodes:

$$\alpha_e(t) = \alpha_0(\frac{\alpha_0}{\epsilon})^{-\frac{t}{t\_max-1}} \quad for\ 0 \leq t < t\_max,$$

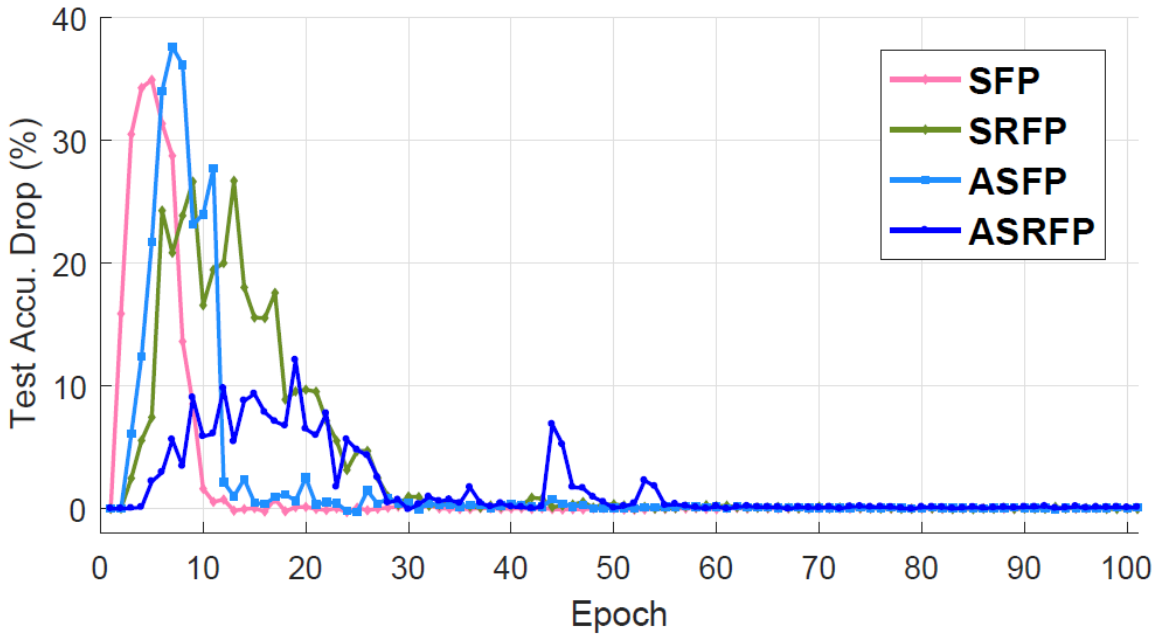
SFP zeroizes pruned nodes



SRFP decays pruned nodes with a decaying parameter

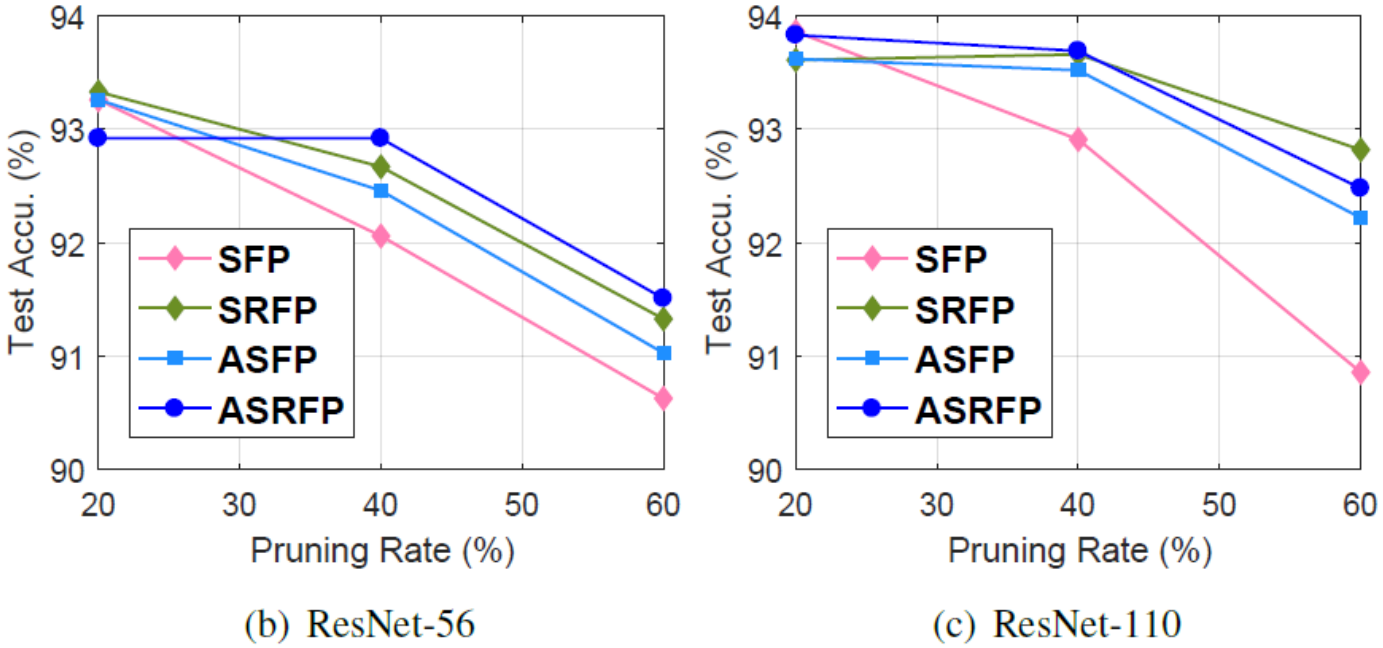


Result



- 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%.
- The Test Accuracy Drop is the difference between the Top-1 accuracy before pruning and the Top-1 accuracy after pruning, where 0 means that there is no obvious accuracy drops caused by pruning.

Result



- **Comparison of Test Accuracies of ResNet-56/110 on CIFAR-10 among SFP/ASFP/SRFP/ASRFP with the pruning rate changing.**

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

[1] Y. He, G. Kang, X. Dong, Y. Fu, and Y. Yang, “Soft filter pruning for accelerating deep convolutional neural networks,” IJCAI International Joint Conference on Artificial Intelligence, vol. 2018-July, pp. 2234–2240, 2018.

[2] Y. He, X. Dong, G. Kang, Y. Fu, C. Yan, and Y. Yang, “Asymptotic Soft Filter Pruning for Deep Convolutional Neural Networks,” IEEE Transactions on Cybernetics, vol. PP, pp. 1–11, 2019.