Slimming ResNet by Slimming Shortcut

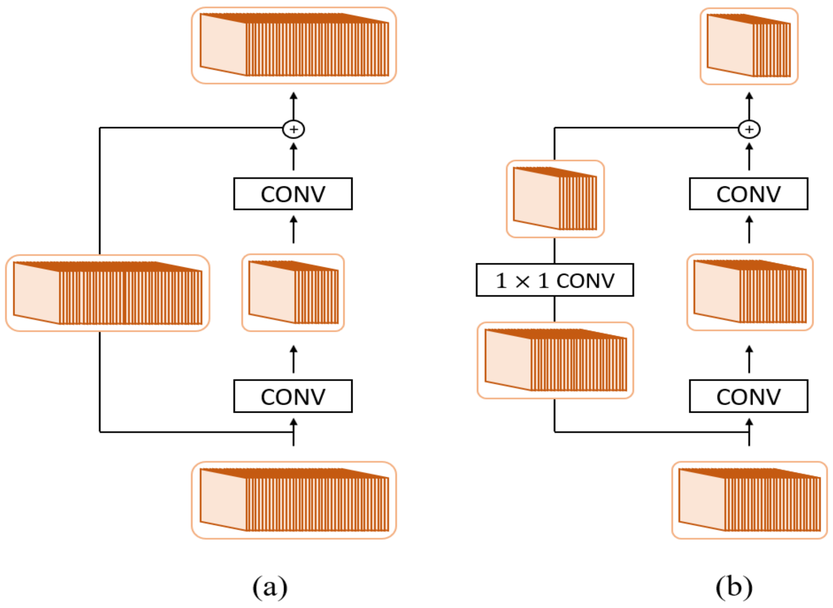
College of Engineering
School of
Electrical Engineering

Donggyu Joo Doyeon Kim Junmo Kim

ICPR 2020

Introduction

- Pruning of ResNet is mainly focused on the inside of the blocks.
- The connections in shortcut (skip-connection) makes the pruning related to the shortcut difficult.
- This induces the unstable bottleneck-like structure after pruning.



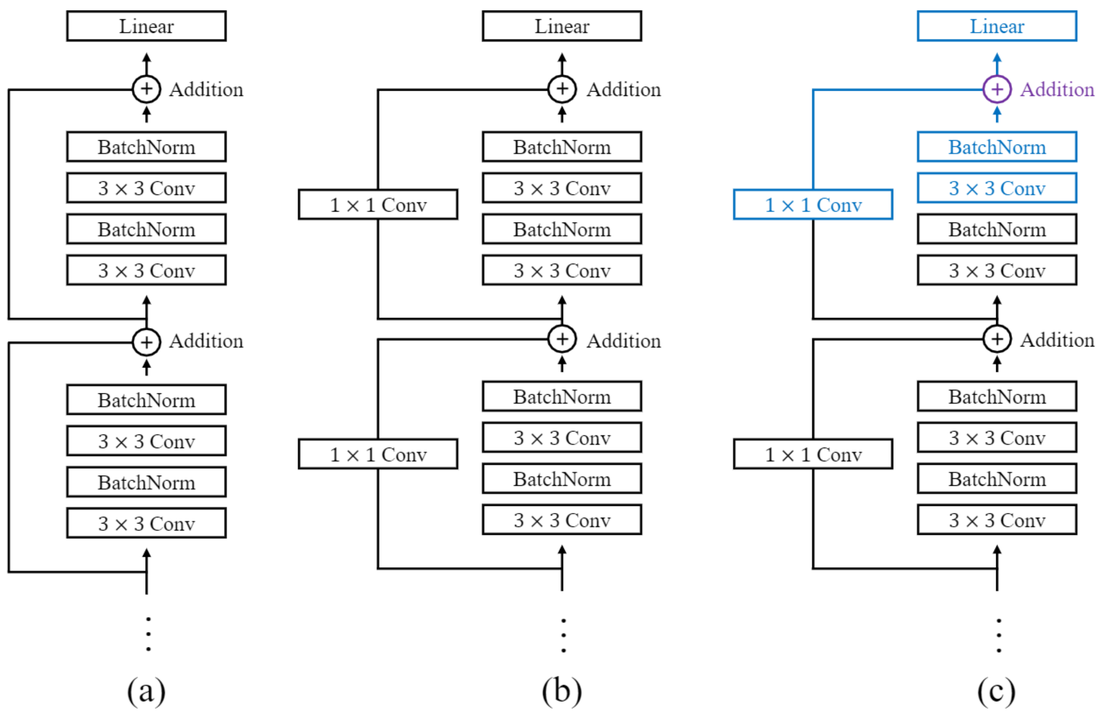
(a) Inblock Pruning: conventional pruning method only prunes the channels inside of the residual block. (b) SSPruning: channels inside of the block and shortcut can be pruned together without any restriction.

- By utilizing the additional 1×1 convolutional layer, we prune the channels related to the shortcut without limitation. We propose a novel pruning method, Slimming Shortcut Pruning (*SSPruning*) for pruning channels in shortcuts on ResNet based networks.

Method

1. Shortcut Separation

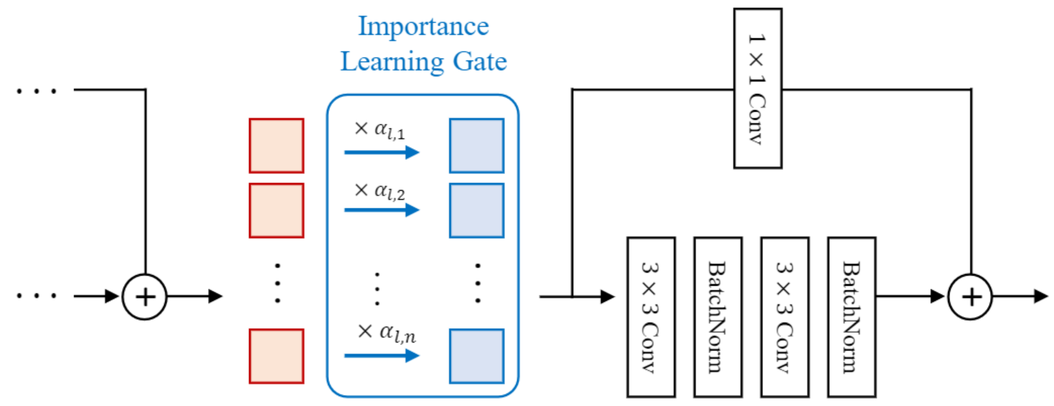
- We locate 1×1 convolutional layer to each shortcut region to separate every connected channels. This convolutional layer is initialized as an identity matrix to preserve the network computation



- (a) Original ResNet. Removing one channel in the shortcut leads to the change of most layers and filters related to them.
- (b) Shortcut Separation. Using the additional 1×1 convolutional layer, we can separate each shortcut in independent region.
- (c) A 1×1 convolution layer of the shortcut is blocking a direct connection to the bottom layers. When we prune the channels related to an addition layer (marked in purple), we do not need to modify the shortcuts of the other layers. We only need to modify the layers marked in blue.

2. Importance Learning Gate

- The importance learning gate (ILG) layer is applied after each addition layer (also between two convolutional layers in the residual block).
- ILG is composed of n parameters $\alpha_1, \alpha_2, \cdots, \alpha_n \in R$ where n is the number of channels in the corresponding location. These n parameters are initialized as 1, and independently multiplied to the output of each corresponding channel.



- We add a regularization term $f(\alpha) = \lambda|\alpha|$ to each ILG parameter to force to have smaller values for less important channels.
- While the training proceeds, for each epoch, k number of channels with the top-k smallest ILG parameter in the entire network are selected and removed from the network.

Experiments

1. Comparative Experiments

Network	Method	Baseline Acc (%)	Pruned Acc (%)	Acc ↓ (%)	Flops ↓ (%)
ResNet-32	MIL [32]	92.33	90.74	1.59	31.2
	SFP [33]	92.63	92.08	0.55	41.5
	FPGM [12]	92.63	91.93	0.70	53.2
	SSPruning-50	**92.54**	**92.80**	**-0.26**	**50.1**
	SSPruning-60	**92.54**	**92.36**	**0.18**	**60.1**
ResNet-110	PFEC [34]	93.53	93.30	0.23	38.6
	SFP [33]	93.68	93.38	0.30	40.8
	NISP [29]	-	-	0.18	43.8
	FPGM [12]	93.68	93.85	-0.17	52.3
	SSPruning-50	**93.65**	**93.99**	**-0.34**	**50.1**
	SSPruning-60	**93.65**	**93.76**	**-0.11**	**60.1**

2. Ablation on Shortcut Pruning

Experiment for SSPruning			
Flops ↓ (%)	Pruned Acc (%)		
	InBlock Only	Shortcut Only	Both (proposed)
50%	92.58	92.16	**92.80**
60%	92.08	91.57	**92.36**

Experiment for Random Selection			
Flops ↓ (%)	Pruned Acc(%)		
	InBlock Only	Shortcut Only	Both
50%	92.32	92.29	**92.51**
60%	91.84	91.85	**92.03**

Conclusion

- In this study, we propose a new pruning framework SSPruning which enables the pruning of shortcuts in the ResNet.
- This approach reveals the potential of the shortcut pruning which can be widely used with various methods.