Rethinking Group Fisher Pruning for Efficient Label-Free Network Compression

Paper ID: 693

Jong-Ryul Lee[1] and Yong-Hyuk Moon[1,2]

[1] Electronics and Telecommunications Research Institute (ETRI)
[2] University of Science and Technology (UST)

BMVC 2022

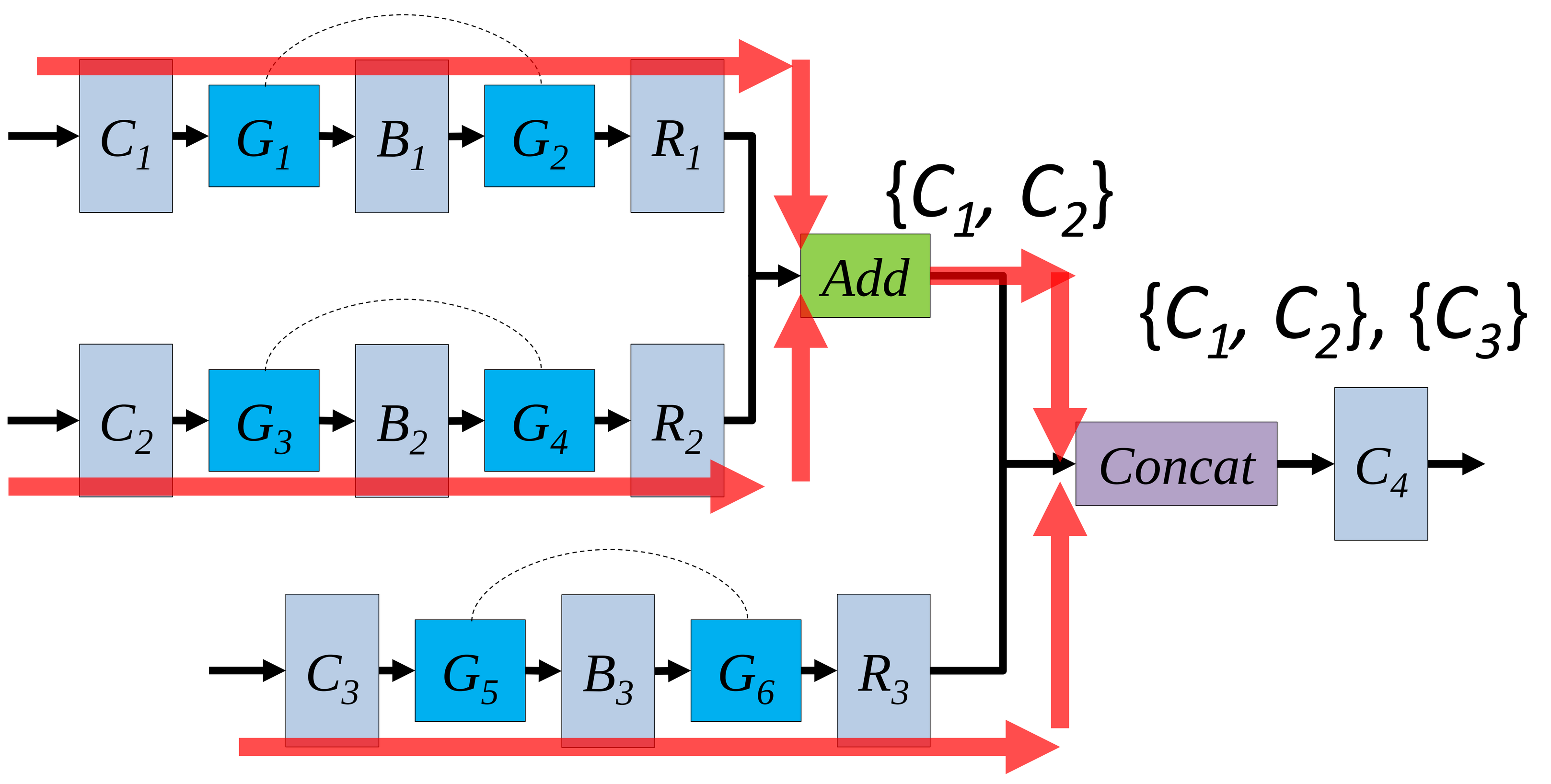
Introduction

- Group Fisher Pruning
  - ✓ A powerful gradient-based channel pruning method for convolutional neural networks
  - ✓ Limitation 1: Not support concatenation
  - ✓ Limitation 2: Too expensive cost for pruning channels

- Our contributions
  - ✓ A formal algorithm to handle DenseNet-style skip connections for pruning channels
  - ✓ Effectively reducing Group Fisher Pruning's cost
  - ✓ Connecting knowledge distillation with Group Fisher Pruning for label-free channel pruning

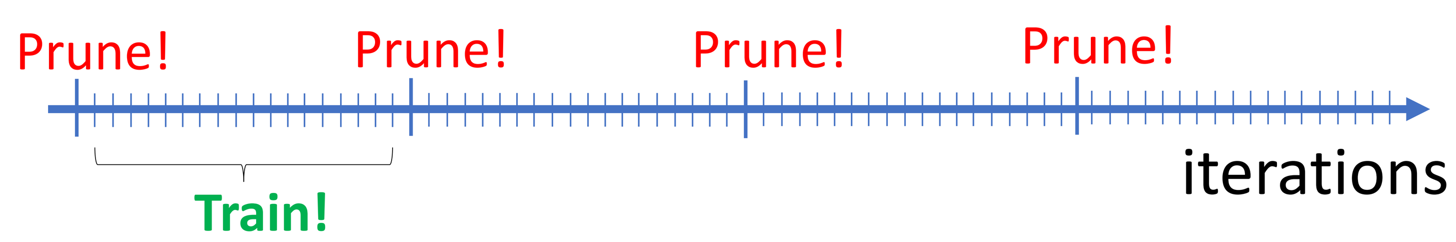
Handling Skip Connections

- The output channels of C1 are coupled with those of C2.
- The output channels of C1 aren't coupled with those of C3.
- By keeping predecessor convolutional layers, our algorithm finds groups of layers (gates) sharing coupled output channels.

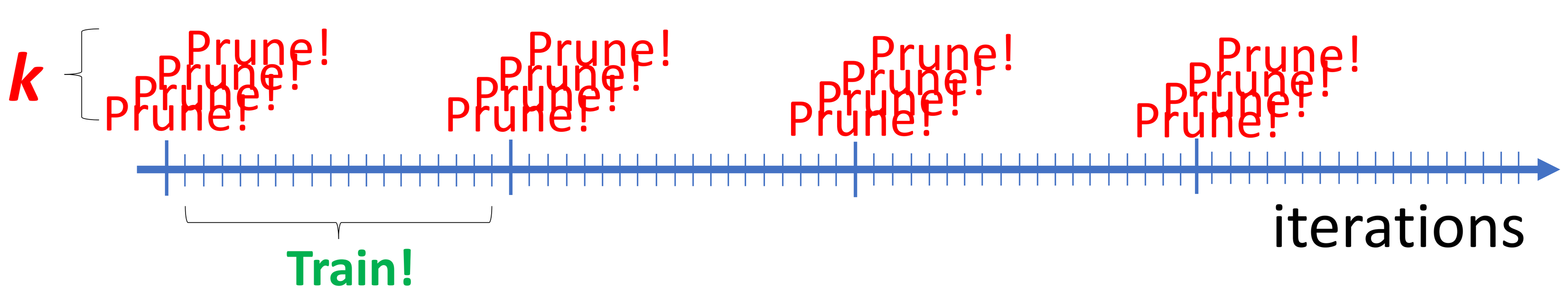


Making It Efficient

- Group Fisher Pruning
  - ✓ Removes a single channel for each pruning step



- Our method
  - ✓ The number of removed channels at a time -> k.
  - ✓ For each pruning step, our method removes the top-k least important channels based on the score function.



Toward Label-free Pruning

- Exploiting knowledge distillation with the output probability distribution and intermediate output tensors.
- Anchor layers (**L**): Specially selected layers providing such intermediate output tensors for knowledge distillation

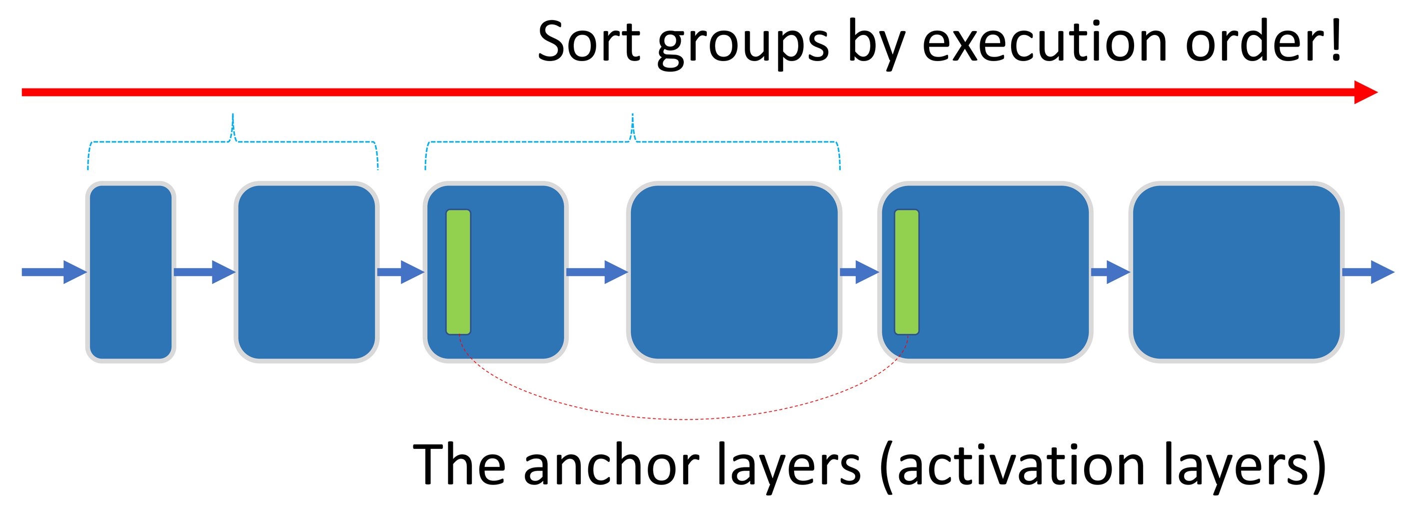
$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{N}\sum_{n=1}^{N} -\mathbf{p}_i \log \frac{\mathbf{p}_i}{\mathbf{q}_i}$$

KL divergence between the outputs of the teacher and a pruned model

$$\mathcal{L}_{\mathrm{prune}} = \mathcal{L}_{\mathrm{KL}} + \frac{1}{N}\sum_{n=1}^{N}\sum_{L\in\mathbf{L}} \left(\mathcal{X}_i^L - \mathcal{X}_i^{L_j}\right)^2$$

MSE loss between the output tensors of the anchor layers

- Anchor Layer Selection
  - ✓ Sort groups sharing coupled channels in execution order
  - ✓ Divide them into same sized partitions
  - ✓ The act. layer of the first common descendant convolution layer for each partition is selected as an anchor layer.



Results

- ImageNet and CIFAR-100 results

	ImageNet		CIFAR-100	
	Top-1	#FLOPs(B)	Top-1	#FLOPs(B)
ENetB0	77.19	0.40	86.46	0.40
GF	70.02	0.22	81.49	0.22
$\mathrm{CURL_D}$	69.37	0.23	**82.47**	0.22
$\mathrm{HRank_D}$	**71.37**	0.22	80.08	0.22
$\mathrm{BTS_{HEU}}$	68.66	0.22	78.25	0.22
$\mathrm{BTS_{ALL}}$	68.10	0.21	80.55	0.21
$\mathrm{BTS_{FCD}}$	**71.89**	0.22	**81.79**	0.22
DNet121	74.76	2.85	84.39	2.85
GF	67.13	1.67	**83.11**	1.68
$\mathrm{CURL_D}$	69.45	1.69	82.13	1.65
$\mathrm{HRank_D}$	**69.59**	1.74	81.37	1.65
$\mathrm{BTS_{FCD}}$	**70.52**	1.69	**82.91**	1.65

- Pruning cost

	GFP	CURL	HRank	Ours
ImageNet	1,319	32	5	**3**
CIFAR-100	1,457	35	5	**4**

Corresponding author: Yong-Hyuk Moon (yhmoon@etri.re.kr)