FILTER PRUNING VIA SOFTMAX ATTENTION

ICIP 2021, Anchorage, Alaska, USA

Sungmin Cho, Hyeseong Kim and Junseok Kwon

School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea

INTRODUCTION

Problem

- ✓ Too many parameters and Flops are needed for models
- ✓ Not available in small device

Network Pruning

- ✓ A kind of model compression methods.
- ✓ Focus on filter (conv) pruning

Goal

- ✓ Reduce parameters and Flops
- ✓ Prevent dropping Accuracy

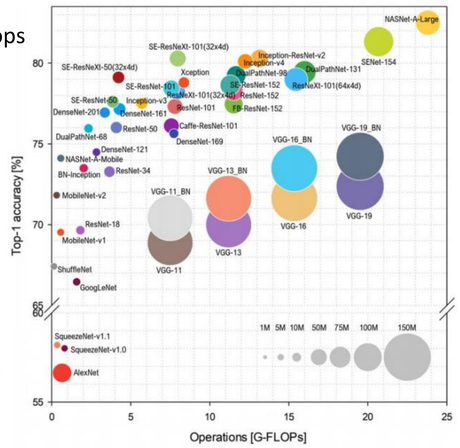


Figure [1] parameters, flops top-1 accuracy on imagenet

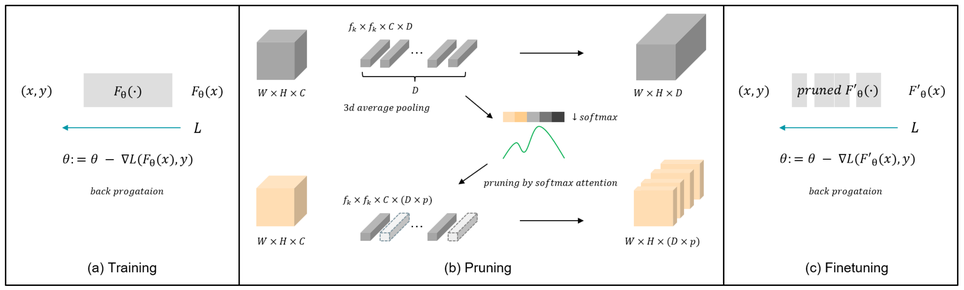
OVERVIEW

Pruning process

- ✓ Training – Pruning – Finetuning
- ✓ In E-scratch method, finetunning

Proposed Method

- ✓ Relative depthwise seperable Conv
- ✓ Softmax Attetntion pruining



SOFTMAX ATTENTION

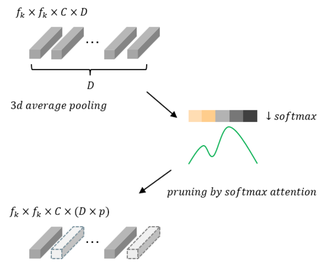
Softmax attention pruning

$$F_i \in \mathbb{R}^{f_k \times f_k \times C \times D}$$

$$F^i_{avgpool} = \frac{1}{N} \sum_j^{f_k} \sum_k^{f_k} \sum_l^{C} F_i \in \mathbb{R}^D$$

$$||F^i_{scores}||_1 = ||\sigma\left(F^i_{avgpool}\right)||$$

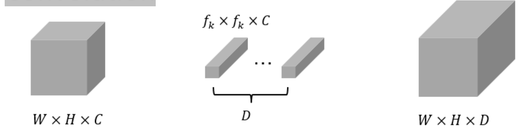
$$||F^i_{scores}||_1 \quad > \quad p \in \{0.5, 0.75, 0.875\}$$



- ✓ Let Fi be the i-th convolution filters (weights)
- ✓ Averpooling and take softmax function
- ✓ Find out effectively channel importance considering all other channels

RELATIVE DEPTHWISE SEPARLABLE CONVOLUTION

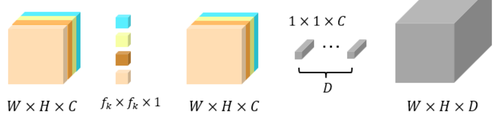
Convolution



- ✓ Basic conv
- ✓ $W \times H \times C \times D$

DWC

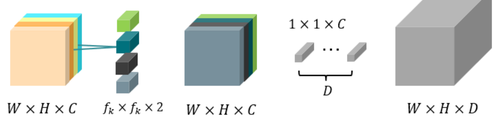
depthwise conv *pointwise conv*



- ✓ Depth wise separable conv [2]
- ✓ $W \times H \times D + C \times D$
- ✓ Consists of DepthW, PointW
- ✓ Not sufficient Intersection of channels as pointwise convolution

RDWC

depthwise conv With channel interaction *pointwise conv*



- ✓ Realtive depthwise separable conv
- ✓ $W \times H \times D \times 2 + C \times D$
- ✓ Enlarge channel intersection, becom propagation info to entire channels

EXPEREMENTS & RESULTS

Ablation Study

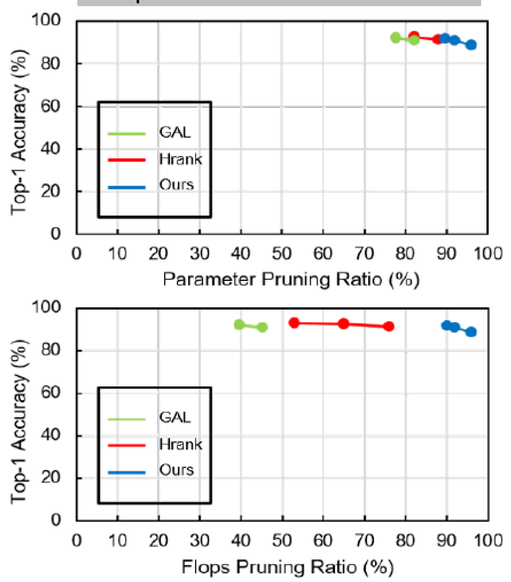
	Flops(PR)	# of params	Top-1%
VGG5(baseline)	(0%)	(0%)	**90.91**
RDWC	(-87.01%)	(-85.04%)	90.61
RDWC + pruning ($p = 0.5$)	**(-96.26%)**	**(-94.84%)**	90.00

- ✓ Prunes lots of flops and parameters
- ✓ The accuracy drop is very small
- ✓ Outperform other state-of-the-art methods in terms of Flops, parameters, top-1 Acc.

Comparison

dataset	model	pruning rate	Flops(PR)	# of params(PR)	Top-1%	Acc Drop
MNIST	vgg5	-	161.61M (0%)	3.65M(0%)	99.72%	0 %
	Spinalvgg5	-	-	3.63M (-0.55%)	99.72%	0 %
	Ours	0.875	16.43M(-89.83%)	452.04K(-87.62%)	**99.74%**	+0.02%
	Ours	0.75	12.41M(-92.32%)	351.94K(-90.36%)	99.72%	0%
	Ours	0.5	**6.05M(-96.26%)**	**188.21K(-94.84%)**	99.70%	-0.02%
Fashion-MNIST	vgg5	-	161.61M (0%)	3.65M(0%)	94.63%	0 %
	Spinalvgg5	-	-	3.63M (-0.55%)	**94.68%**	+0.05%
	Ours	0.875	16.43M(-89.83%)	452.04K(-87.62%)	93.94%	-0.69%
	Ours	0.75	12.41M(-92.32%)	351.94K(-90.36%)	93.64%	-1.01%
	Ours	0.5	**6.05M(-96.26%)**	**188.21K(-94.84%)**	93.05%	-1.58%
CIFAR10	vgg16	-	313.73M(0%)	14.99M(0%)	93.48%	0%
	GAL	-	171.89M(-45.2%)	2.67M(-82.20%)	91.23%	-2.25%
	Hrank	-	73.70M(-76.5%)	1.78M(-88.13%)	90.73%	-2.75%
	Ours	0.875	30.69M(-90.22%)	1.56M(-89.59%)	**91.67%**	-1.81%
	Ours	0.75	23.06M(-92.65%)	1.19M(-92.06%)	90.80%	-2.68 %
	Ours	0.5	**11.06M(-96.47%)**	**0.59M(-96.06%)**	88.69%	-4.79 %

Comparison with state of the art



REFERENCE

[1] Bianco, Simone, et al. "Benchmark analysis of representative deep neural network architectures." *IEEE Access* 6 (2018): 64270-64277.

[2] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

IEEE ICIP 2021