Revisiting Random Channel Pruning for Neural Network Compression: Supplementary Material

Yawei Li[1] Kamil Adamczewski[2] Wen Li[3] Shuhang Gu[4] Radu Timofte[1] Luc Van Gool[1,5]
[1]Computer Vision Lab, ETH Zürich [2]MPI-IS [3]UESTC [4]USYD [5]KU Leuven
{yawei.li, radu.timofte, vangool}@vision.ee.ethz.ch

In this supplementary material, we first explain in detail the difference between this work and the previous works. Then we provide a justification of our statement in the main paper "the performance of the channel pruned network is upper bounded by the original network". The we show how residual blocks with skip connections are pruned in Sec. 3. Finally, more experimental results are given in Sec. 4.

1. Difference with Other Works

In the main paper, we explained the main difference between our work and [4, 8]. In this supplementary, we provide a detailed comparison between our work and [4, 8].

Difference with [8]: Our work is different from [8] in the following aspects.

1) ***Aim.*** The aim of [8] is to identify the value of network pruning as discovering the network architecture whereas our aim is to propose random pruning as a neutral baseline to compare different pruning methods.

2) ***Method.*** How to select the pruning ratio is not thoroughly investigated in [8] while our work uses random pruning.

3) ***Empirical study.*** The empirical study in [8] is mostly done pairwise by comparing a network resulting from a pruning algorithm and the one trained from scratch. Comparison between different pruning criteria is not done. Our work thoroughly compares 6 pruning criteria and 1 architecture search method.

Difference with [4]: Our work is different from [4] in the following aspects.

1) ***Perspective.*** The analysis in [4] is conducted on single layers while our work evaluates the overall network performance.

2) ***Conclusion.*** The theoretical and empirical analysis in [4] mainly support the similarity between norm based pruning criteria. Yet, the empirical study does not support the similarity between importance-based, BN-based, and activation-based pruning criteria. Our study discovers comparable performances between norm-based, importance-based, sensitivity-based, and search based methods.

3) ***Enlightenment.*** The study in [4] "guides and motivates the researchers to design more reasonable criteria" while our study finds out that advanced pruning criteria behaves just comparable with the naive L1/L2 norm "calls for an optimized sampling method that improves the search efficiency".

2. Upper Bounded Performance of Channel Pruning.

In the this section, we provide the justification of the statement in the main paper "the performance of the channel pruned network is upper bounded by the original network".

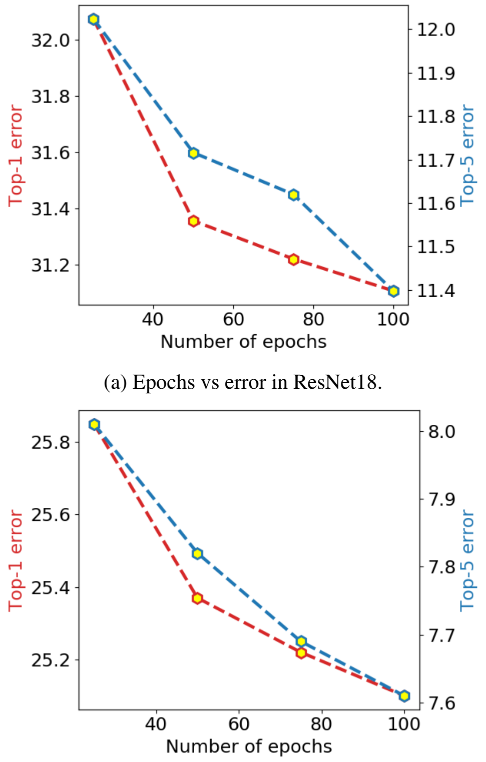
In the paper "The Lottery Ticket Hypothesis", the authors showed that some pruned networks could learn faster while reaching higher test accuracy and generalizing better than the original one [1]. Yet, the conclusion is derived for unstructured pruning. The problems of unstructured pruning and structured pruning are quite different. Unstructured pruning removes single connections in a CNN and results in irregular kernels. And it is possible that the number of kernels in the resultant sparse network is the same as the original network. The capacity of a network could be fully utilized by the sparse network. This is why unstructured pruning could easily lead to an extremely pruned network without accuracy drop while for structured pruning researchers struggle with the trade-off between accuracy drop and compression ratio. Without expanding the search space (*i.e.* changing the position of pooling layers [7], widening the network [5, 12, 13]), it is very difficult to find a pruned network with better performance. Thus, we can safely conclude that the performance of channel pruned networks is upper bounded by the original networks.

Criterion	Top-1 Error (%)	Top-5 Error (%)	FLOPs [G] / Ratio (%)	Params / Ratio (%)
VGG, CIFAR10				
Baseline	5.67	0.58	313.80 /100.00	14.73M /100.00
L1	6.1	0.69	160.50 /51.15	5.05M /34.32
L2	6.06	0.67	150.60 /47.99	6.20M /42.11
GM	5.99	0.52	154.60 /49.27	4.13M /28.04
TE	6.51	0.61	157.00 /50.03	5.84M /39.63
ES	6.21	0.64	157.20 /50.10	7.06M /47.90
KL	6.19	0.66	161.50 /51.47	6.52M /44.26
ResNet20, CIFAR10				
Baseline	7.48	0.61	41.20 /100.00	272.5k /100.00
L1	9.03	0.48	20.90 /50.73	170.1k /62.43
L2	8.65	0.55	20.60 /50.00	169.9k /62.37
GM	8.69	0.6	20.90 /50.73	188.9k /69.31
TE	8.96	0.46	20.90 /50.73	164.1k /60.23
ES	8.5	0.63	25.60 /62.14	207.8k /76.24
KL	8.77	0.46	20.10 /48.79	165.2k /60.64
ResNet56, CIFAR10				
Baseline	5.58	0.26	126.80 /100.00	855.8k /100.00
L1	6.72	0.79	63.60 /50.16	503.6k /58.85
L2	6.52	0.76	64.70 /51.03	471.4k /55.08
GM	6.39	0.77	65.40 /51.58	504.0k /58.89
TE	6.86	0.59	65.70 /51.81	442.4k /51.69
ES	6.59	0.67	65.80 /51.89	545.6k /63.75
KL	7.12	0.67	65.20 /51.42	443.3k /51.80
ResNet20, CIFAR100				
Baseline	31.53	9.87	41.20 /100.00	278.3k /100.00
L1	33.41	10.42	20.80 /50.49	176.2k /63.29
L2	33.39	10.62	21.00 /50.97	175.9k /63.20
GM	33.32	10.35	20.60 /50.00	183.8k /66.03
GW	34.24	10.92	20.00 /48.54	168.8k /60.65
ES	33.81	10.13	21.00 /50.97	176.3k /63.34
KL	33.32	10.62	21.20 /51.46	187.5k /67.35
ResNet56, CIFAR100				
Baseline	27.59	9.24	126.80 /100.00	861.6k /100.00
L1	30.15	9.34	63.20 /49.84	470.8k /54.64
L2	29.48	9.43	65.80 /51.89	513.6k /59.61
GM	29.2	9.35	62.30 /49.13	559.4k /64.92
TE	29.01	9.33	65.50 /51.66	534.3k /62.01
ES	29.49	9.16	64.20 /50.63	554.4k /64.34
KL	29.23	9.3	65.10 /51.34	568.0k /65.92

Table 1. Benchmarking channel pruning criteria on CIFAR10 and CIFAR100 image classification under the scheme of random pruning.

3. Pruning Residual Blocks

Pruning a normal convolutional layer is straightforward. But when it comes to the residual blocks in MobileNetV2 [11] and ResNet [2], some special measures should be taken. For the residual blocks in MobileNetV2 and ResNet, there is a skip connection that adds the input of the block to the output of the block so that the block learns a residual component. Since the input and output of residual blocks are connected, the number of output channels of several residual blocks are the same. When pruning the residual block, their output channels should be pruned together. For both of the pruning settings explained in the main paper, *i.e.* pruning pre-trained network and and pruning from scratch, we set the same pruning ratio for the convolutional layers that are connected by skip connection.



(a) Epochs vs error in ResNet18.

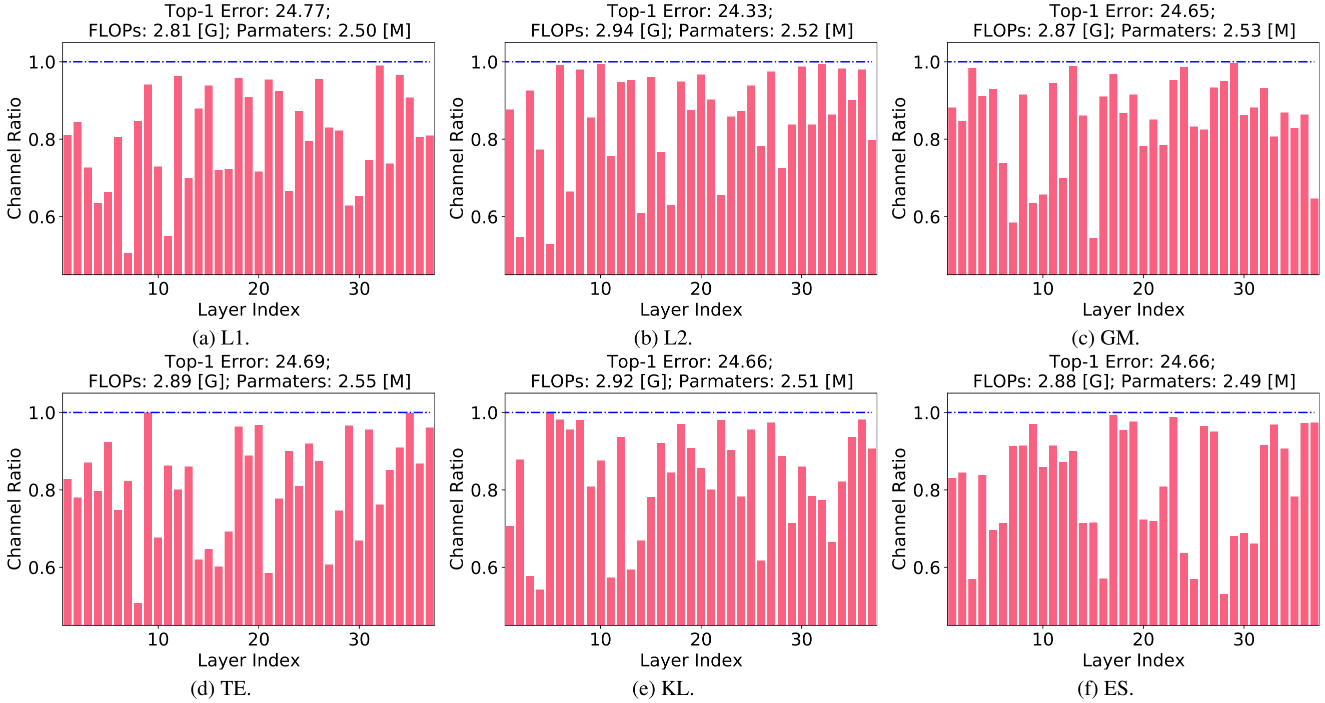
(b) Epochs vs. error in ResNet50.

Figure 1. The influence of the random sample size and fine-tuning epochs on the prediction accuracy.

Special treatments should also be taken when computing the importance score according to different pruning criteria. **I. L1/L2/GM.** For the convolutional layers that are connected by skip connection, their individual importance scores are first computed and then added up. The summation result is used as the final importance score. **II. TE.** As in the original paper, gates with weights equal to 1 and dimensionality equal to the number of output channels are append to the Batch Normalization layers. The importance score are first computed based on the gates and then added for the layers that are skip-connected. **III. ES.** The maximum empirical sensitivity is computed for layers that are connected by skip connections. **IV. KL.** To compute the KL divergence for the output probability of the pruned and original networks, masks that selects the output channels should be added to the convolutional layers. For the convolutional layers that are skip-connected, we set the same mask for them so that the same KL divergence score can be computed for all of them.

4. More Experimental Results

More experimental results are shown in this section. The results for CIFAR image classification are summarized in Table 1. Besides the results in the main paper, results of ResNet20 on CIFAR10 and ResNet56 on CIFAR100 are



(a) L1. (b) L2. (c) GM.

(d) TE. (e) KL. (f) ES.

Figure 2. Percentage of remaining channels of the pruned ResNet50 network. The network pruned by different methods are reported. The pruning ratio is 70%. The Top-1 error, FLOPs, and number of parameters are also reported in the figure.

also included. As in the main paper, a couple of pruning criteria are compared including the traditional L1 and L2 norm of the filters (L1, L2), and the recent method based on geometric median (GM) [3], Taylor expansion (TE) [10], KL-divergence importance metric (KL) [9] and empirical sensitivity analysis (ES) [6]. The additional results strengthen the conclusion in the main paper. That is, under the scheme of random pruning, the pruning criteria for selecting different channels are less important.

The influence of fine-tuning epochs on the final accuracy of the pruned network is shown in Fig. 1. The result for ResNet-50 is shown in Fig. 1b. The result for ResNet-18 is shown in Fig. 1a. When the number of fine-tuning epochs is increased from 25 to 100, the Top-1 and Top-5 error of ResNet-50 drops by 0.75% and 0.4%, respectively. For ResNet-18, the Top-1 error rate and Top-5 error rate drop by 0.97% and 0.62%, respectively. This shows the significant influence of fine-tuning epochs.

In Fig. 2, the ratio of remaining channels for each of the convolutional layer is plotted. The original network is ResNet50 for ImageNet classification and the overall pruning ratio is 70%. The Top-1 error, FLOPs, and number of parameters are also reported in the figure. In Fig. 3, the accuracy distribution of the random pruned networks with respect to FLOPs is shown. Note that the networks are only updated by minimizing the squared difference between the features maps of the pruned and original network. Fine-tuning has not been conducted during this step. As can be seen, both good sub-networks with low error rate and less accurate sub-networks can be sampled. And the aim is to search the sub-networks with higher accuracy. Similar to Fig. 3, the accuracy distribution with respect to the number of parameters is shown in Fig 4.

References

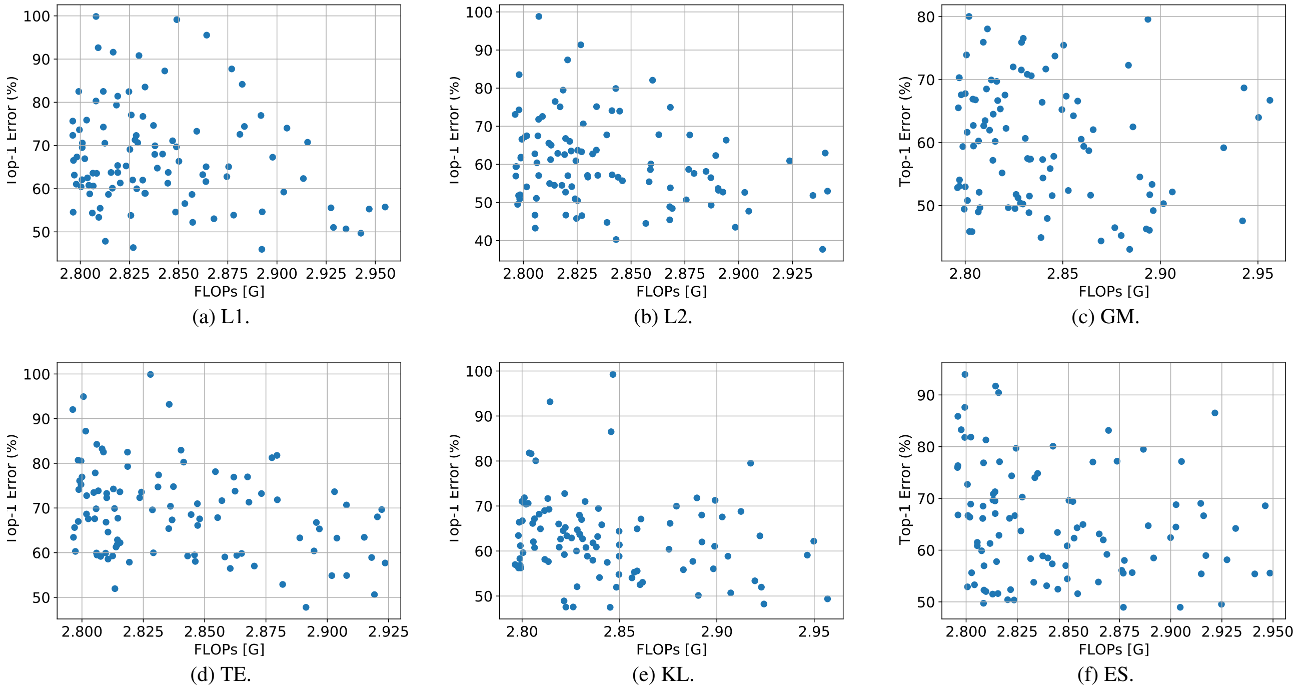
[1] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[3] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2019. 3

[4] Zhongzhan Huang, Wenqi Shao, Xinjiang Wang, Liang Lin, and Ping Luo. Rethinking the pruning criteria for convolutional neural network. In *Advances in Neural Information Processing Systems*, 2021. 1

[5] Yawei Li, Wen Li, Martin Danelljan, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. The heterogeneity hypothesis: Finding layer-wise differentiated network architectures. In *Proceedings of the IEEE/CVF Conference*



(a) L1.

(b) L2.

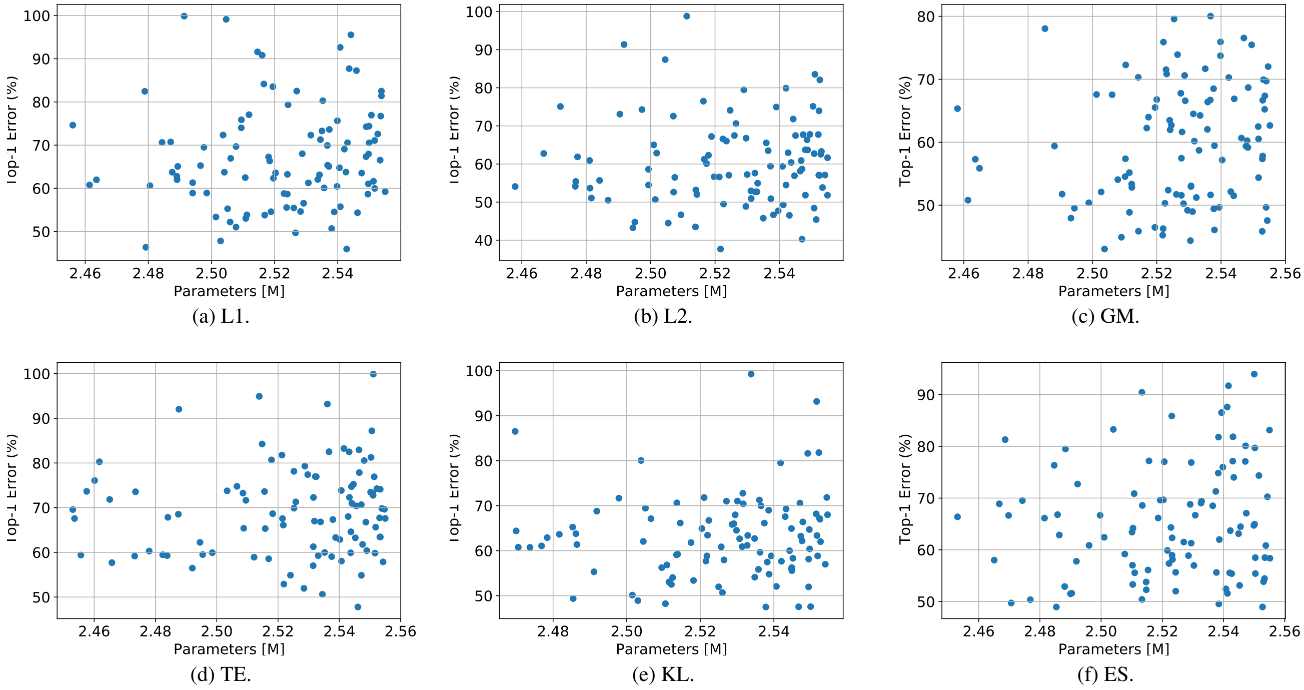
(c) GM.

(d) TE.

(e) KL.

(f) ES.

Figure 3. Accuracy distribution of network samples with respect to FLOPs for different pruning criteria. The original network is ResNet50 trained for ImageNet classification. The network pruning ratio is 70%.



(a) L1.

(b) L2.

(c) GM.

(d) TE.

(e) KL.

(f) ES.

Figure 4. Accuracy distribution of network samples with respect to the number of parameters for different pruning criteria. The original network is ResNet50 trained for ImageNet classification. The network pruning ratio is 70%.

on Computer Vision and Pattern Recognition, pages 2144–2153, 2021. 1

[6] Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, and Daniela Rus. Provable filter pruning for efficient neural networks. *arXiv preprint arXiv:1911.07412*, 2019. 3

[7] Zechun Liu, Haoyuan Mu, Xiangyu Zhang, Zichao Guo, Xin Yang, Tim Kwang-Ting Cheng, and Jian Sun. MetaPruning: Meta learning for automatic neural network channel pruning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[8] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *Proceedings of International Conference on Learning Representations*, 2019. 1

[9] Jian-Hao Luo and Jianxin Wu. Neural network pruning with residual-connections and limited-data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1458–1467, 2020. 3

[10] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11264–11272, 2019. 3

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 2

[12] Jiahui Yu and Thomas Huang. AutoSlim: Towards one-shot architecture search for channel numbers. *arXiv preprint arXiv:1903.11728*, 2019. 1

[13] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. *arXiv preprint arXiv:1812.08928*, 2018. 1