Filter Pruning via Automatic Pruning Rate Search(Supplementary Material)

Qiming Sun[1], Shan Cao[1], and Zhixiang Chen[2]

[1] Shanghai University, Shanghai 200444, China
cshan@shu.edu.cn
[2] The University of Sheffield, Sheffield, S1 4DP, UK

1 Supplementary Experiments

1.1 Ablation Experiment

Compared with other network structures, VGG-16 has greater redundancy. Our ablation experiments provide a clearer perspective of the impact of APRS based on VGG-16 with CIFAR-10. The settings of the ablation experiment are the same as the main text. The first ablation experiment uses random cropping to remove parameters. Under similar parameter reduction and FLOPs with random pruning in Table 1, APRS achieves desirable top-1 accuracy (94.51% vs 93.27%), which is higher than the baseline. In the second ablation experiment we use a reverse order for search pruning rate, named Anti-APRS. It seems noticeable that Anti-APRS performs very poorly in all cases compared to APRS. The accuracy of Anti-APRS is only 88.73% when removing 55.83% of the parameters. The accuracy drop will be more serious when more parameters are

Table 1: Ablation studies of sensitivity indicators. Anti-APRS means pruning in the reverse order recommended by APRS, Random is achieved by randomly assigning channel sensitivity. ↓% refers the percentage of the removed part.

Method	Acc(%)	Parameters(↓%)	FLOPs(↓%)
VGG-16	93.87	0.0	0.0
Random	93.27	39.23	33.87
Anti-APRS	89.63	36.79	35.57
APRS(ours)	**94.51**	**38.66**	**37.97**
Random	92.44	56.58	52.32
Anti-APRS	88.73	55.83	49.28
APRS(ours)	**93.72**	**55.23**	**46.72**
Random	90.42	74.78	55.67
Anti-APRS	87.09	73.56	51.54
APRS(ours)	**93.47**	**73.02**	**52.90**

Table 2: Comparison results on CIFAR-10 dataset. Acc(%)↓ refers Top-1 accuracy decrease. PR represents the pruning rate.

Method	Top-1(%)	Acc(%)↓	Parameters(PR)	FLOPs(PR)
VGG-16-L1 [1]	93.25	0.0	15.00M(0.0%)	313.00M(0.0%)
L1 [1]	93.40	-0.15	5.4M(64.0%)	206.00M(34.2%)
APRS+L1	**93.79**	**-0.54**	**5.4M(64.0%)**	**173.28M(44.6%)**
VGG-16-HRank [2]	93.87	0.0	14.98M(0.0%)	313.73M(0.0%)
HRank [2]	92.34	1.53	2.64M(82.1%)	108.61M(65.3%)
APRS+HRank	**94.23**	**-0.36**	**2.64M(82.1%)**	**98.12M(68.7%)**
ResNet-110-L1 [1]	93.53	0.0	1.72M(0.0%)	253.00M(0.0%)
L1 [1]	93.55	-0.02	1.68M(2.3%)	213.00M(15.8%)
APRS+L1 [1]	**93.79**	**-0.26**	**1.68M(2.3%)**	**200.08M(20.9%)**
ResNet-110-HRank [2]	93.50	0.0	1.72M(0.0%)	252.89M(0.0%)
HRank [2]	94.23	-0.73	1.04M(39.4%)	148.70M(41.2%)
APRS+HRank	**94.29**	**-0.79**	**1.04M(39.4%)**	**127.84M(49.5%)**

pruned (Anti-APRS obtains a hign accuracy of 87.09% with around 73% parameters reduction). Compared with Anti-APRS, APRS has advantages in all fields (46.72% *vs.* 49.28% in FLOPs reduction, 55.23% *vs.* 55.83% in parameters reduction, 93.72% *vs.* 88.73% in top-1 accuracy).

Table 1 confirms that using the overall Wasserstein ranking and considering the joint effect of inter-layer sensitivity leads to higher performance. From a certain angle, APRS handles the redundancy in the parameters more flexibly, and allocates the overall pruning power to each layer in a targeted manner, so as to obtain a more efficient model.

1.2 Combination of APRS and other hierarchical pruning methods.

APRS can provides us an optimized pruning rate, which we use to get hign performance on CIFAR-10 [3] and ImageNet [4]. APRS can not only be used as a separate filter pruning algorithm, but also can be combined with other algorithms to achieve good performance. We applied the pruning rate searched in APRS on top of other filter pruning algorithms, and the results are shown in Table 2. L1 [1] numbers the filters and greedily prunes the channels one by one. The rank of the output feature map is found to be an important indicator in filter pruning [2], and the pruning rate of each layer is set. Although both methods achieve remarkable pruning results, APRS makes them perform better again. For VGG-16 on CIFAR-10, when the optimized pruning rate is applied on the basis of HRank [2], the performance will be improved by 1.89% (94.23 *vs.* 92.34%), as shown in Table 2. In addition, APRS+L1 [1] obtains 93.79% top-1 accuracy with ResNet-110 on CIFAR-10 (improved by 0.24% with less FLOPs compared to L1 [1]).

References

1. Han, S., Pool, J., Tran, J., Dally, W.J.: Pruning Filters for Efficient ConvNets. In: ICLR. (2017)
2. Lin, M., Ji, R., Wang, Y., Zhang, Y., Zhang, B., Tian, Y., Shao, L.: HRank: Filter Pruning using High-Rank Feature Map. In: CVPR. (2020)
3. Alex Krizhevsky, Geoffrey Hinton, e.a.: Learning multiple layers of features from tiny images (2009) Technical report, Citeseer.
4. Russakovsky, O., Jia Deng, H.S., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. In: IJCV. (2015)