Fire Together Wire Together: A Dynamic Pruning Approach with Self-Supervised Mask Prediction

Sara Elkerdawy[1] Mostafa Elhoushi[2] Hong Zhang[1]
Nilanjan Ray[1]
[1]University of Alberta, [2]Toronto Heterogeneous Compilers Lab, Huawei
{elkerdaw, hzhang, nray1}@ualberta.ca

1. Compute Information

Table 1 shows the compute hours for the training of the reported models in our paper using a 4 V100 GPU machine. Code is available at https://github.com/selkerdawy/FTWT

Dataset	Model	Compute hours
CIFAR-10	VGG	2.0
	ResNet56	8.0
	MobileNet	3.5
ImageNet	ResNet18	40.0
	ResNet34	55.5
	MobileNetv1	41.0

Table 1. Compute hours on a 4 V100 GPU machine

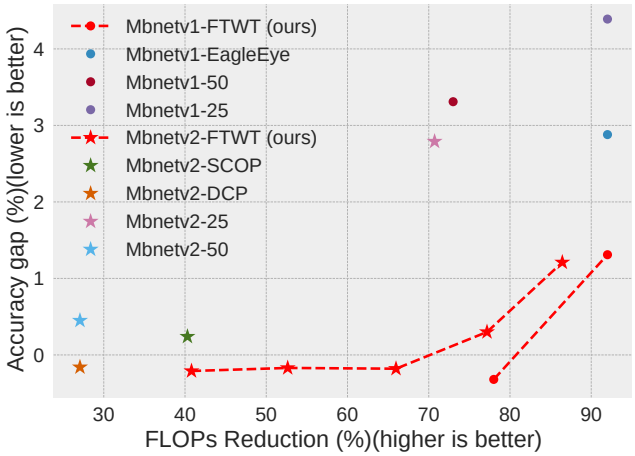

Figure 1. MobileNetV1/V2 on CIFAR10.

2. CIFAR

2.1. MobileNet

We compare our method on MobileNetv1/v2 under different pruning ratio with other pruning methods such as EagleEye [1], SCOP [4] and DCP [5]. Note that EagleEye reports the best out of two candidate models different in signature, thus double the training time. We outperform SOTA by 37% higher FLOPs reduction on similar accuracy as shown in 1.

2.2. Core Filters Visualization

We show visualization of number of core filters per layer as explained in main papers. Core filters per layer indicate the filters that are activated in all different routes. Diversity is the highest at the middle layers which explains the accuracy drop in static uniform pruning (Table 1 main paper). Static MobileNet_50 results on 3.31% drop in accuracy in comparison to our dynamic method which improved the baseline with slight increase in accuracy 0.17%.

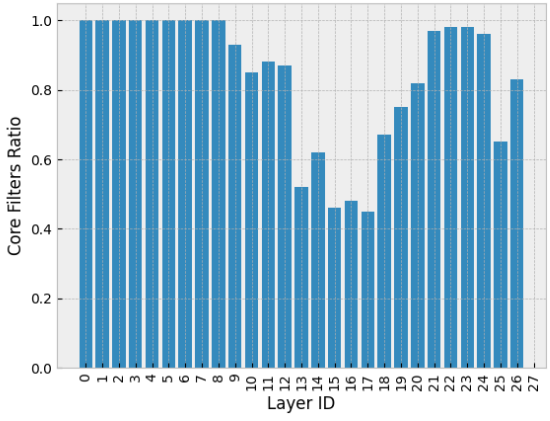

Figure 2. Number of core filters per layer in MobileNet.

2.3. Error Bars

Table 2 shows the numerical details of CIFAR experiments with the mean and standard deviation over 3 runs.

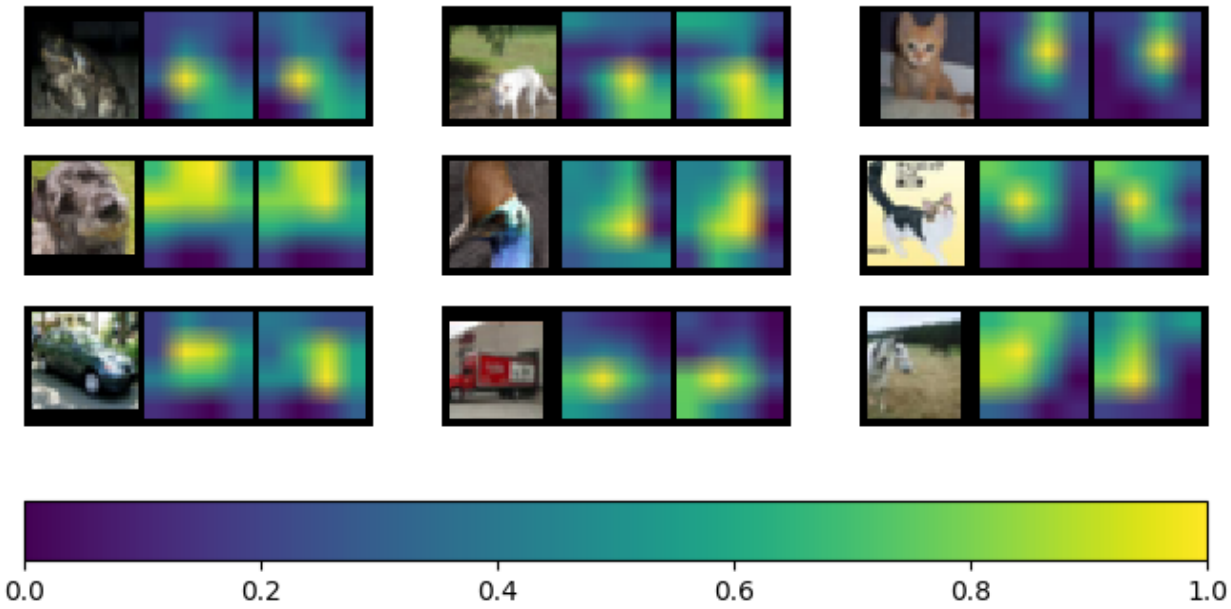

Figure 3. Heatmap visualization of random input samples from CIFAR for the 10th layer in MobileNetV1. Each triplet represents input image, baseline heatmap, pruned heatmap. FLOPs reduction in the layer is $\approx$ 70%, yet the pruned heatmap highly approximate the heatmap with fully activated filters.

Run	VGG16-BN		VGG16-BN		ResNet56		MobileNet	
	Acc. (%)	FLOPs (%)	Acc. (%)	FLOPs (%)	Acc. (%)	FLOPs (%)	Acc. (%)	FLOPs (%)
	93.26	73.10	93.66	65.41	92.61	66.45	91.01	78.03
	93.21	73.00	93.49	64.37	92.63	66.43	91.16	79.52
	93.11	73.47	93.51	65.27	92.65	66.41	91.03	78.04
Mean	93.19	73.19	93.55	65.01	92.63	66.43	91.06	78.53
Std	0.07	0.24	0.09	0.56	0.02	0.02	0.08	0.8

Table 2. Mean and standard deviation across different runs on CIFAR.

2.4. Heatmap Visualization

We visualize the heatmap of a highly pruned layer in comparison to the baseline model. Figure 3 shows comparison between heatmap from the baseline with all filters activated and heatmap of dynamically selected filters. As can be seen, dynamic pruning approximates the baseline with high attention on foreground objects. This shows that even with 70% pruning ratio in that layer, we are able to approximate the behavior of the original model.

3. ImageNet

3.1. Joint vs Decoupled

As training ImageNet models are expensive, we show few experiments to compare the results with joint and decoupled training modes in Table 3. Similar to results on CIFAR, decoupled training outperforms joint training under similar FLOPs reduction.

Model	Joint	Decouple	FLOPs reduction (%)
Resnet34	70.06	71.71	37
	71.52	72.79	52

Table 3. Joint vs decoupled training on ResNet34 ImageNet

Broader Impact

Neural Network pruning has the potential to increase deployment efficiency in terms of energy and response time. However, obtaining these pruned models are yet to be optimized for a better overall computational consumption and more environment friendly. Moreover, pruning require careful understanding of deployment scenarios such as questioning out-of-distribution generalization [2] or altering the behavior of networks in unfair ways [3]. We showed results on out-of-distribution shift to tackle the first part. We did not investigate fairness of the model's pre-

diction as both datasets (i.e CIFAR and ImageNet) are balanced. Although, we prune based on the mass of the heatmap equally for all samples. We hypothesize this trait can give our method an advantage over fixed pruning ratio which might hurt some input samples over some others.

References

[1] Bailin Li, Bowen Wu, Jiang Su, and Guangrun Wang. Eagleeye: Fast sub-net evaluation for efficient neural network pruning. In *European conference on computer vision*, pages 639–654. Springer, 2020. 1

[2] Lucas Liebenwein, Cenk Baykal, Brandon Carter, David Gifford, and Daniela Rus. Lost in pruning: The effects of pruning neural networks beyond test accuracy. *arXiv preprint arXiv:2103.03014*, 2021. 2

[3] Michela Paganini. Prune responsibly. *arXiv preprint arXiv:2009.09936*, 2020. 2

[4] Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chunjing Xu, Chao Xu, and Chang Xu. Scop: Scientific control for reliable neural network pruning. *Advances in Neural Information Processing Systems*, 33:10936–10947, 2020. 1

[5] Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. Discrimination-aware channel pruning for deep neural networks. *Advances in neural information processing systems*, 31, 2018. 1