Feather: An Elegant Solution to Effective DNN Sparsification - *Supplementary Material*

Athanasios Glentis Georgoulakis
athglentis@gmail.com
George Retsinas
gretsinas@central.ntua.gr
Petros Maragos
maragos@cs.ntua.gr

School of Electrical and Computer Engineering
National Technical University of Athens
15773 Athens, Greece

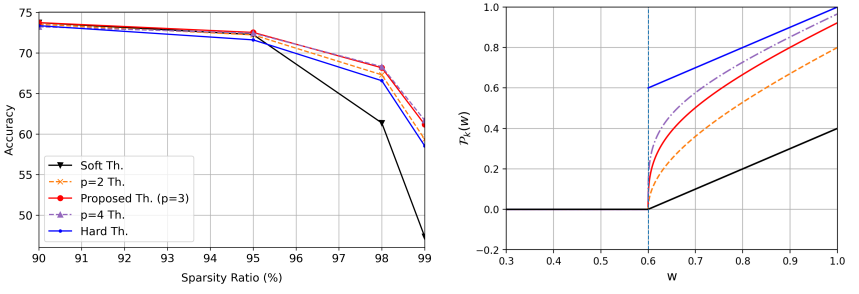
1 Training Hyperparameters

Dataset	CIFAR-100	ImageNet
Epochs	160	100
Batch Size	128	256
Weight Decay	5e-4	5e-5
Optimizer	SGD	SGD
LR	0.1	0.2
LR-Scheduler	Cosine	Cosine+Warmup
Momentum	0.9	0.9
Label Smoothing	-	0.1

Table 1: Training hyperparameters used for all our experiments on CIFAR-100 and ImageNet datasets.

Table 1 summarizes the training hyperparameters used for our experiments on CIFAR-100 [3] and ImageNet [6] datasets. The chosen hyperparameters are selected based on standard practices for the particular datasets and are kept the same regardless the network architecture or the target sparsity ratio (in contrast to *e.g.* [2, 8] where the Weight Decay is adjusted among different runs, based on the target sparsity ratio). By adopting commonly used hyperparameters and keeping them unchanged among all our experiments we opted to show that our method is able to achieve SoA results without the need of fine-tuning and complicated training configurations.

2 Impact of Threshold's p-value



(a) (b)

Figure 1: A study of the effect of the p value of the proposed family of thresholds on the final sparse model accuracy. Results from ResNet-20 trained on CIFAR-100 (a) and the corresponding thresholds used (b).

The proposed threshold with $p = 3$ is compared with the ones with $p = 2$ and $p = 4$ in Figure 1. We observe that $p = 3$ is preferable to $p = 2$ based on the resulting final accuracy while $p = 4$ results to no further improvement (Figure 1(a)). Due to that, $p = 3$ is chosen to give a fine balanced threshold between the two extremes, Hard and Soft thresholds respectively, although, as shown, good results are obtainable even with values of p near 3. A reasonable explanation for the slight under-performance using $p = 2$ is that the resulting threshold still leads to a considerable amount of shrinkage (Figure 1(b)), thus induces more bias between the thresholded weights and their dense counterparts. Notably, even for $p = 2$ the results are favorable compared to those obtained by using the Hard and Soft Thresholds, further validating the robustness of our family of threshold operators.

3 Stability of the Sparsity Mask *vs*. Gradient Scaling

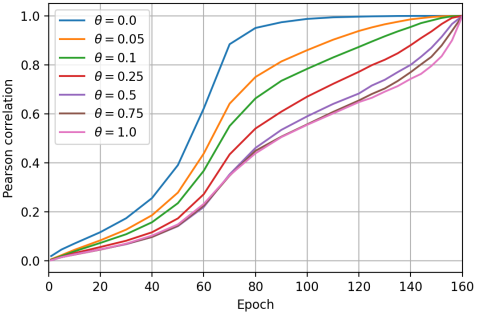
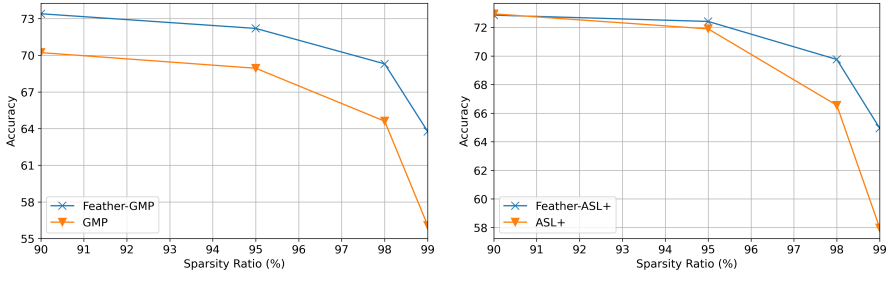


Figure 2: Plot of Pearson correlation coefficients between the sparsity mask obtained at the end of each epoch and the mask at the end of training, for varying values of the gradient scaling parameter θ. Results from ResNet-20 trained on CIFAR-100.

Figure 2 empirically validates that the gradient scaling parameter $\theta \in [0,1]$ influences the stability of the sparsity mask, *i.e.* the mask that indicates which parameters are pruned and which are active during sparse training. Specifically, for each experiment, using a specified (constant) value for θ, the Pearson correlation coefficients between the mask at the end of every training epoch and the final mask, obtained at the end of training, are shown. Experiments with θ near zero result to curves that converge to 1 more rapidly, compared to the ones from experiments with θ close to unity. This indicates that when using θ near zero (or at the extreme case $\theta = 0$) the sparsity mask (and thus the sparsity pattern) is stabilized earlier during the training process, compared to when using larger values of θ. Based on our empirical analysis, the suitable amount of stability for the sparsity mask relates to the sparsity target; The higher the requested final sparsity the more beneficial is to keep the mask more stable (up to a reasonable extent) to avoid destabilizing the highly pruned network. We note that the mask's stability is also studied in [7], where a soft top-k mask is computed by solving a regularized Optimal Transportation problem in order to regulate its stability, although our approach using gradient scaling (combined with the proposed threshold operator) is considerably less computationally expensive while resulting to favorable final accuracies.

4 Feather Improves Pruning Backbones



(a) (b)

Figure 3: Feather improves the accuracy of common sparse training backbones: (a) GMP, a uniform layer-wise sparsity pruning backbone (b) the ASL+ framework. Results from ResNet-20 trained on CIFAR-100.

Combining the Feather module with existing backbones results to more accurate networks, as shown in Figure 3. In 3(a) Feather is used to improve the accuracy of GMP [10], a layer-wise magnitude pruning backbone that prunes all layers[1] to the same (uniform) amount of sparsity, gradually increasing the pruning ratio. Our module significantly improves the resulting accuracy when combined with the very simplistic GMP backbone. Furthermore, in 3(b) we compare the accuracy of the sparse models obtained with Feather combined with ASL+ [5] and the ones using only ASL+, showing that our module leads to accuracy improvements for the challenging sparsity ratios (95% and above).

[1]With the exception of the first convolutional layer, which was left dense when using GMP in our experiments due to having a very small number of parameters.

5 MobileNetV1 on ImageNet

Ratio	89%	94.1%
MobileNetV1 (4.21M Params): 71.95		
GMP [10]	61.80	-
STR [4]	62.10	-
ProbMask [9]	65.19	60.10
ST-3 [8]	66.67	61.19
Feather-Global	**68.13**	**63.63**

Table 2: Top-1 accuracy of MobileNetV1 on ImageNet.

In Table 2 we provide additional experiments on ImageNet [6] using the MobileNetV1 [2] architecture. More specifically, we compare the accuracies obtained by using Feather combined with the global pruning backbone with the ones from GMP [10], STR [4], ProbMask [9] and ST-3 [8] which report results for the 89% and 94.1% sparsity ratios, using the same number of epochs (100) and data augmentation as in our experiments. Our approach surpasses the previous SoA by 1.46% and 2.44% Top-1 accuracy at the 89% and 94.1% sparsity ratios respectively, a result that further validates Feather's effectiveness and generalization abilities on large datasets with different model architectures.

6 Accuracy *vs*. FLOP Measurements

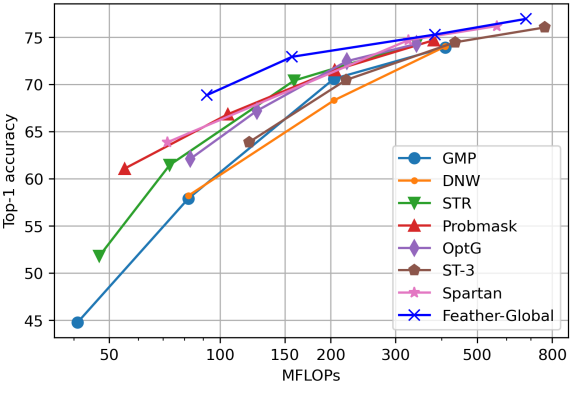


Figure 4: Top-1 accuracy *vs*. FLOPs of ResNet-50 on ImageNet.

The Feather module, combined with the global pruning backbone, leads to favorable Top-1 accuracy results over the ones from the baselines under similar FLOPs requirements of the sparsified ResNet-50 [1], as shown by the frontier curve in Figure 4. We note that the per-layer sparsity distribution obtained by the global pruning backbone by default does not prioritize FLOPs reduction, while layer-wise methods such as GMP [10] and STR [4] tend to result to sparse models with minimum FLOPs for a given sparsity ratio, although at a cost of considerable accuracy drops.

While extended analysis on optimizing FLOPs for a given sparsity target is not the scope of this work, to further showcase the efficacy of Feather we experimented with biasing the global pruning backbone towards pruning earlier layers more aggressively, as suggested in [8]. With the FLOPs-biased global pruning backbone, training the ResNet-50 on ImageNet at 99% sparsity, Feather resulted in a model with **67.2**% **Top-1 accuracy**, now requiring only **42MFLOPs**. Therefore, the superior accuracy of our sparse model was greatly preserved, still achieving the best accuracy (by a 3.3% margin) among the baselines at the 99% ratio, now for considerably fewer FLOP requirements, matching those of GMP (41MFLOPs), the baseline resulting to the fewer FLOPs, although having accuracy more than 20% higher. Having showcased Feather's great potential at obtaining models with superior accuracy and FLOPs, we leave further experimentation (possibly with more sophisticated backbones) as future work.

References

[1] 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] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[3] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 2009.

[4] Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, and Ali Farhadi. Soft threshold weight reparameterization for learnable sparsity. In *Proceedings of the International Conference on Machine Learning*, pages 5544–5555, 2020.

[5] George Retsinas, Athena Elafrou, Georgios Goumas, and Petros Maragos. Online weight pruning via adaptive sparsity loss. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3517–3521, 2021.

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[7] Kai Sheng Tai, Taipeng Tian, and Ser Nam Lim. Spartan: Differentiable sparsity via regularized transportation. *Advances in Neural Information Processing Systems*, 35: 4189–4202, 2022.

[8] Antoine Vanderschueren and Christophe De Vleeschouwer. Are straight-through gradients and soft-thresholding all you need for sparse training? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3808–3817, 2023.

[9] Xiao Zhou, Weizhong Zhang, Hang Xu, and Tong Zhang. Effective sparsification of neural networks with global sparsity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3599–3608, 2021.

[10] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.