N° 3072

Weight Reparametrization for Budget-Aware Network Pruning

Robin Dupont[1,2], Hichem Sahbi[1], Guillaume Michel[2]

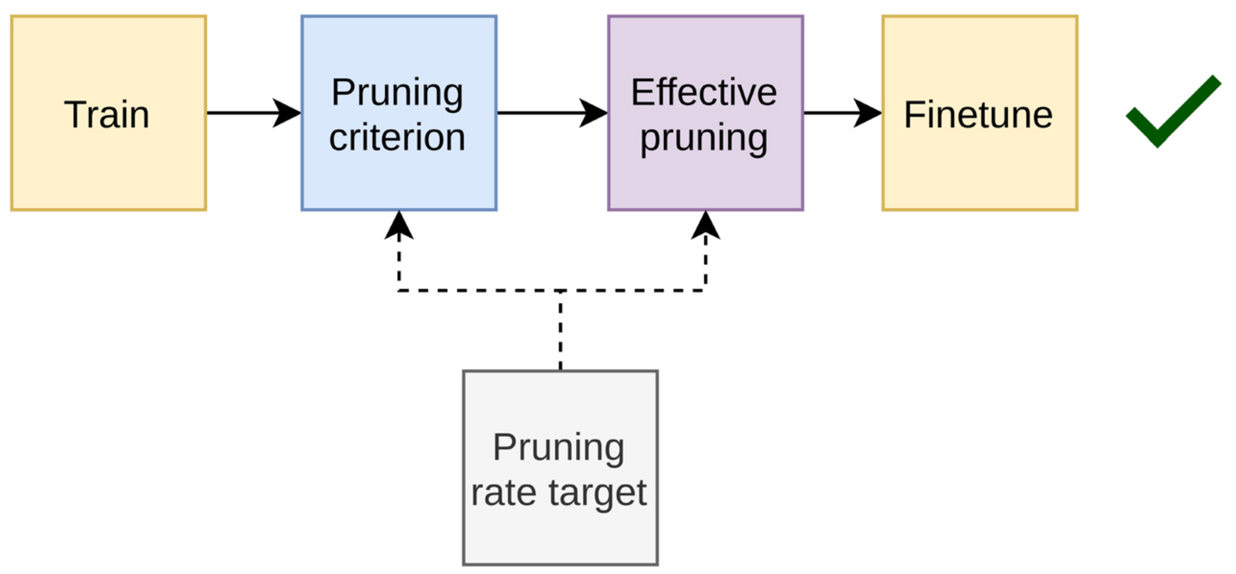
[1]Sorbonne Université - LIP6, [2]Netatmo

Summary : We propose a new weight reparametrization to allow optimization of both topology and weights at the same time, for pruning under budget constraint.

MOTIVATION AND CONTRIBUTION

STANDARD PRUNING PIPELINES

- Standard pruning techniques [1] **require a fine-tuning step**, after effective pruning, in order to compensate for the loss of accuracy.
- This step could be **cumbersome** and the resulting pruned network may be **topologically inconsistent**.

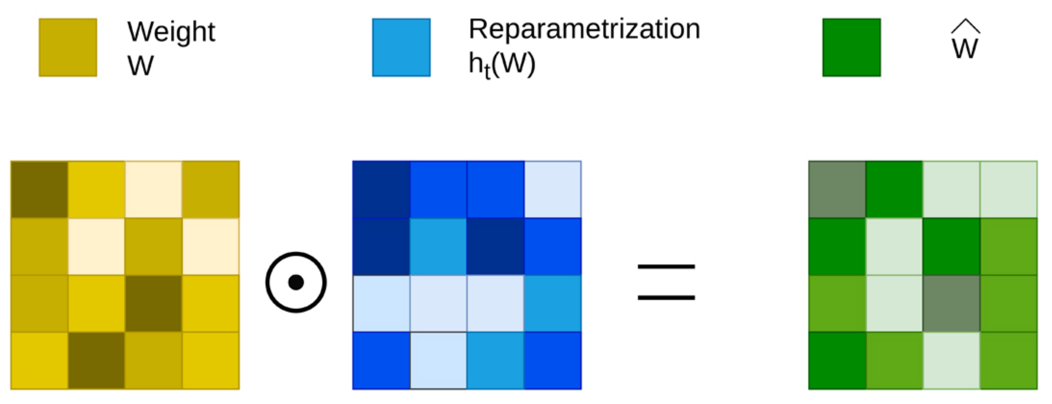


OUR PRUNING PIPELINE

- Our proposed method, in this paper, is **end-to-end** and **does not require** any **fine-tuning** after the effective pruning. The **pruning criterion is embedded** in the reparametrization.
- Our reparametrization also allows **controlling the budget** through a custom loss, thus **optimizing both the topology and the weights** for a given targeted budget.
- Besides, it **prevents disconnections** in the network topology.

OUR METHOD

WEIGHT REPARAMETRIZATION



Reparametrized weights are called **apparent weights** denoted $\hat{w}$. They are defined by $\hat{w} = w \odot h_t(w)$.

REPARAMETRIZATION FUNCTION

The reparametrization function h_t acts as a **regularizer** that **soft-prune the smallest weights**. The soft pruning is later enforced through the effective pruning step.

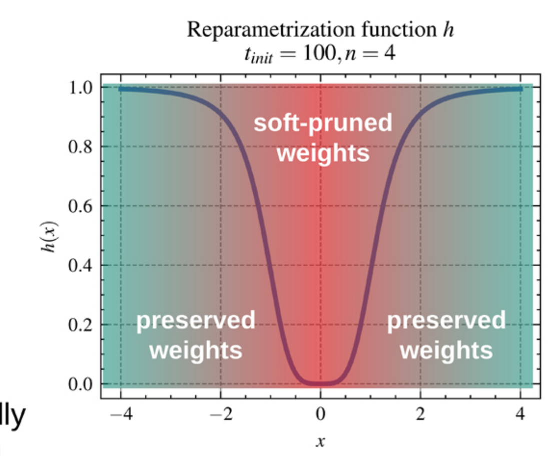
C1 and ***C2*** ensure $0 \leq h_t(x) \leq 1$

$$h_t(x) = C_1\left(\exp\left\{-\frac{1}{(tx)^n + 1}\right\} - C_2\right)$$

t controls the bandwidth of the pit. It is a learnt parameter

n controls the sharpness of the falling and raising edges

+1 to numerically stabilize $h_t(x)$



BUDGET LOSS

The budget loss drives **sparsity**. It is normalized by $C_{initial}$ for **better conditionning**. The budget loss is combined with the classification loss with a **mixing coefficient** λ>0 that controls its **relative importance**

Current cost (sum of weight reparametrizations)

$$C(\{\mathbf{w}_1, \dots, \mathbf{w}_L\}) = \sum_{i=1}^{L} h(\mathbf{w}_i)$$

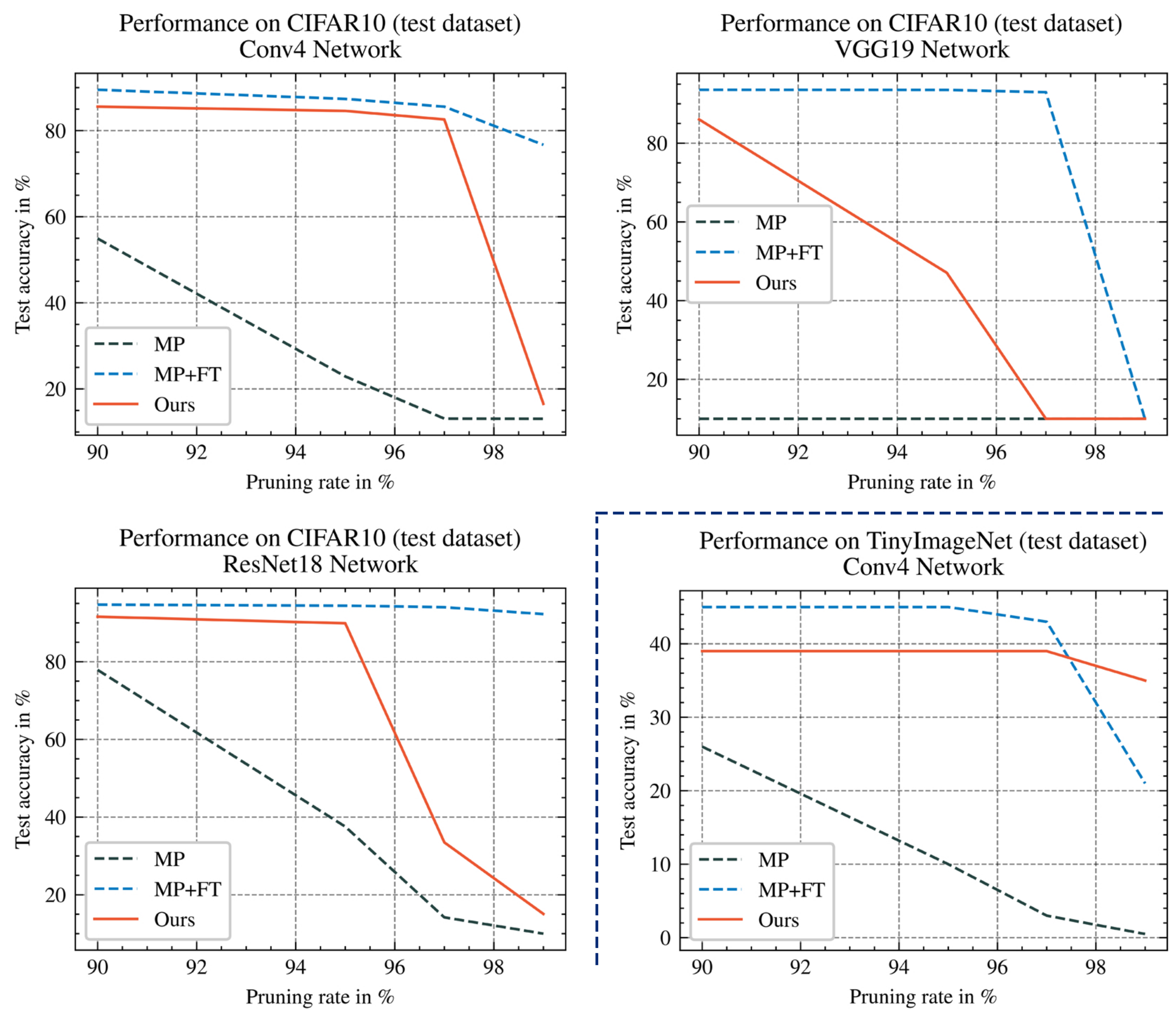
Target cost

$$\mathcal{L}_{\text{budget}} = \left(\frac{C(\{\mathbf{w}_1, \dots, \mathbf{w}_L\}) - C_{\text{target}}}{C_{\text{initial}}}\right)^2$$

Initial cost

RESULTS

Results are shown for **Conv4**[2], **VGG19**[3] and **ResNet18**[4] networks on **CIFAR10** and **TinyImageNet** (only Conv4). Three methods are compared: Ours (**which does not require fine tuning**), Magnitude pruning (MP) and finetuned MP (MP+FT).



RESULTS

- Our reparametrization acts as a **regularizer** and a **saliency indicator**, which **induce sparsity** by **soft-pruning** the smallest weights.
- It allows to optimize **both topology and weights** under **bugdet constraints.**
- Our method significantly **overperforms magnitude pruning without finetuning**, and performs better than finetuned magnitude pruning for **very high pruning rates** on more complex datasets

PERSPECTIVES

- Test on **larger** and **more complex datasets.**
- Improve performances to **consistently** outperform MP+**FT**.
- Try other **reparametrization functions**.

REFERENCES

[1] Song Han et al., "Learning both weights and connections for efficientneural network," in NIPS, 2015.
[2] Jonathan Frankle and Michael Carbin, "The lotteryticket hypothesis: Finding sparse, trainable neural net-works," in ICLR, 2019.
[3] Karen Simonyan and Andrew Zisserman, "Very deepconvolutional networks for large-scale image recogni-tion," in ICLR, 2015.
[4] Kaiming He et al., "Deep Residual Learning for Image Recognition," in CVPR, 2016.