WEIGHT REPARAMETRIZATION FOR BUDGET-AWARE NETWORK PRUNING

ICIP 2021

Robin Dupont - *Sorbonne Université & Netatmo*
Hichem Sahbi - *Sorbonne Université*
Guillaume Michel - *Netatmo*

Machine Learning & Deep Learning for Information Access

Pruning and sparsity

Pruning - Overview

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Pruning

- Group of methods that aim to design **lightweight architectures** *(small memory footprint or fast inference time)*

Pruning - Overview

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Pruning

- Group of methods that aim to design **lightweight architectures** *(small memory footprint or fast inference time)*
- Introduce **sparsity** by removing **redundant** or **unecessary weights** of the network

Pruning - Overview

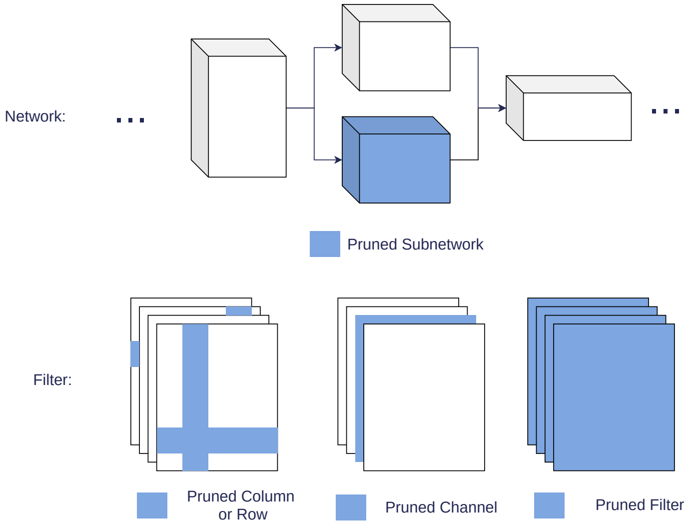
Netatmo SCIENCES SORBONNE UNIVERSITÉ

Pruning

- Group of methods that aim to design **lightweight architectures** *(small memory footprint or fast inference time)*
- Introduce **sparsity** by removing **redundant** or **unecessary weights** of the network
- Can by applied at **different granularity** *(fine-grained vs coarse-grained)*

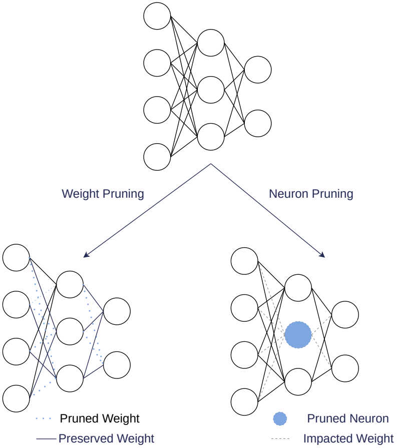
Pruning - Coarse-grained

Netatmo SCIENCES SORBONNE UNIVERSITÉ



Pruning - Fine-grained

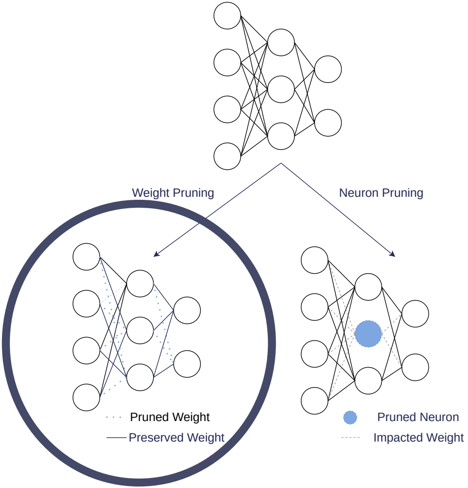
Netatmo SCIENCES SORBONNE UNIVERSITÉ



{*Optimal Brain Damage*, LeCun et al. 1989; *Optimal Brain Surgeon*, Hassibi et al. 1992; *Learning both weights and connections for efficient neural network*, Han et al. 2015}

Pruning - Fine-grained

Netatmo SCIENCES SORBONNE UNIVERSITÉ



Our method belongs to the the **fine-grained weight pruning** category.

{*Optimal Brain Damage*, LeCun et al. 1989; *Optimal Brain Surgeon*, Hassibi et al. 1992; *Learning both weights and connections for efficient neural network*, Han et al. 2015}

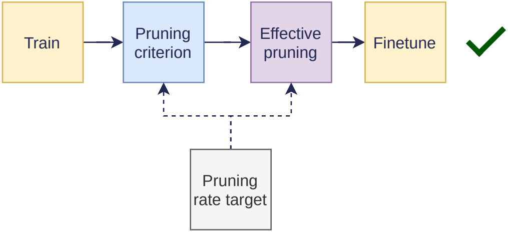
Pipelines

Pruning Pipelines

Pruning - Standard Pipeline

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Standard Pruning Pipeline

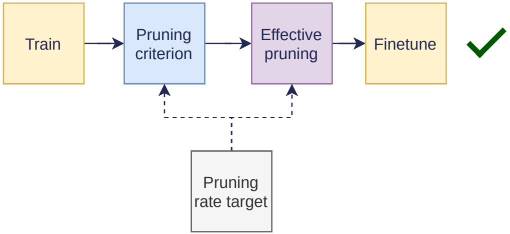


Effective Pruning: Setting pruned weights to 0 and freezing them

Pruning - Standard Pipeline

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Standard Pruning Pipeline



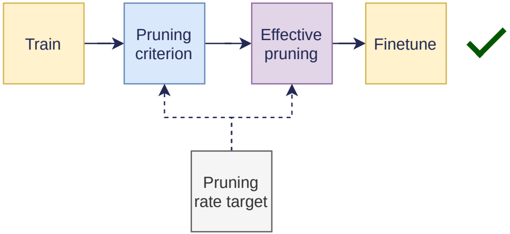
Standard pipelines requires the application of a **pruning criterion** to determine which weights will be **pruned**. This is done **after training**.

Effective Pruning: Setting pruned weights to 0 and freezing them

Pruning - Standard Pipeline

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Standard Pruning Pipeline



Standard pipelines requires the application of a **pruning criterion** to determine which weights will be **pruned**. This is done **after training**.

Initial training does **not take into account final weight budget**.

Effective Pruning: Setting pruned weights to 0 and freezing them

Pruning - Our Pipeline

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Our Pipeline

Effective Pruning: Setting pruned weights to 0 and freezing them

Pruning - Our Pipeline

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Our Pipeline

Our pipeline takes the **final pruning rate** as an input **during initial training**.

Effective Pruning: Setting pruned weights to 0 and freezing them

Pruning - Our Pipeline

Our Pipeline

Our pipeline takes the **final pruning rate** as an input **during initial training**.

Topology is optimized with the budget constraint **from the start.** This helps **preventing disconnections** in the network.

Effective Pruning: Setting pruned weights to 0 and freezing them

Our Method

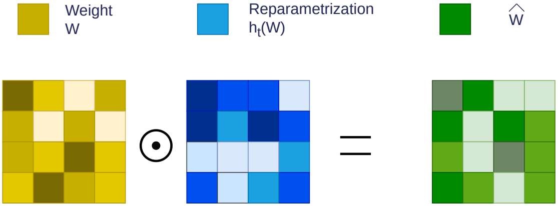
Weight Reparametrization

Weight Reparametrization

Weight Reparametrization

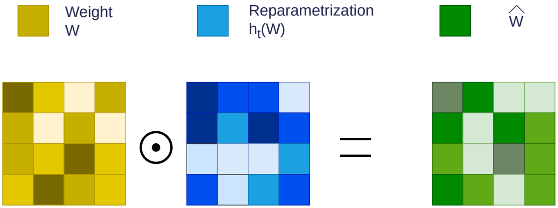
Netatmo SCIENCES SORBONNE UNIVERSITÉ

Weight Reparametrization



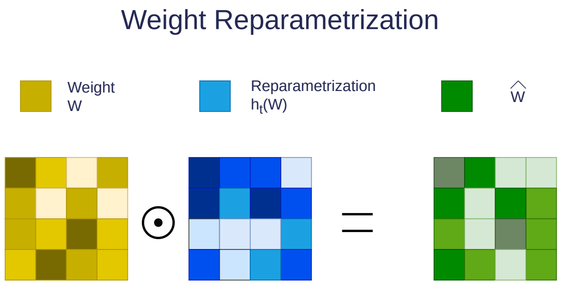
Weight Reparametrization

Weight Reparametrization



- New weights $\hat{w}$ are defined by $\hat{w} = w \odot h_t(w)$.

Weight Reparametrization



- New weights $\hat{w}$ are defined by $\hat{w} = w \odot h_t(w)$.
- w are the standard neural network weights.

Reparametrization Function

Reparametrization Function

Reparametrization function

Netatmo SCIENCES SORBONNE UNIVERSITÉ

C_1 and ***C_2*** ensure
$0 \le h_t(x) \le 1$

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

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 *$h_t(x)$*

Reparametrization function

Netatmo · SCIENCES SORBONNE UNIVERSITÉ

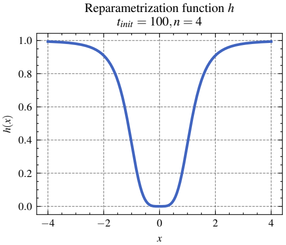
C_1 and C_2 ensure $0 \leq h_t(x) \leq 1$

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

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 $h_t(x)$



Reparametrization function

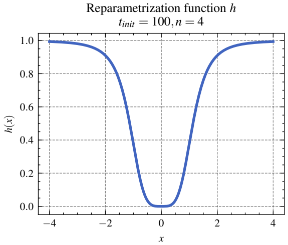
C_1 and C_2 ensure
$0 \le h_t(x) \le 1$

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

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 $h_t(x)$



- *t* is a **learnt parameter**, optimized with SGD. *t* is initialized to 100.

Reparametrization function

Netatmo SCIENCES SORBONNE UNIVERSITÉ

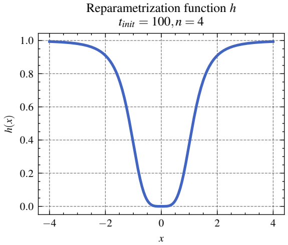
C_1 and **C_2** ensure $0 \le h_t(x) \le 1$

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

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

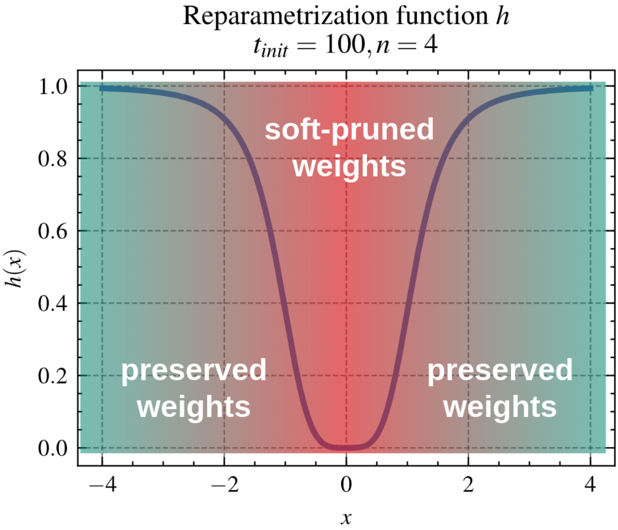
n controls the sharpness of the falling and raising edges

+1 to numerically stabilize $h_t(x)$



- t is a **learnt parameter**, optimized with SGD. t is initialized to 100.
- n is fixed to 4.

Reparametrization function



Budget Loss

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Budget Loss

Budget Loss

Budget: ℓ_0-norm is not differentiable. h is used as a surrogate.

Budget Loss

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Budget: ℓ_0-norm is not differentiable. *h* is used as a surrogate.

Current cost
(sum of weight reparametrizations)

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

Target cost

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

Initial cost

Budget Loss

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Budget: ℓ_0-norm is not differentiable. h is used as a surrogate.

Current cost
(sum of weight reparametrizations)

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

Target cost

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

Initial cost

- **Current cost** is computed at **each step**.

Budget Loss

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Budget: ℓ_0-norm is not differentiable. *h* is used as a surrogate.

Current cost
(sum of weight reparametrizations)

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

Target cost

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

Initial cost

- **Current cost** is computed at **each step**.
- **Target cost** is the **number of parameters** that will be **kept**.

Budget Loss

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Budget: ℓ_0-norm is not differentiable. h is used as a surrogate.

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

- **Current cost** is computed at **each step**.
- **Target cost** is the **number of parameters** that will be **kept**.
- **Initial cost** is the sum of weight reparametrizations **before** the first step.

Budget Loss

Budget: ℓ_0-norm is not differentiable. h is used as a surrogate.

Current cost
(sum of weight reparametrizations)

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

Target cost

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

Initial cost

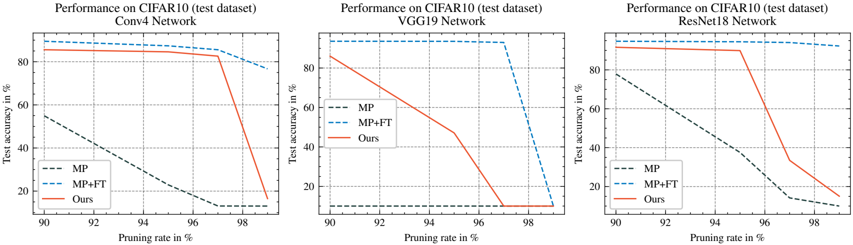
- **Current cost** is computed at **each step**.
- **Target cost** is the **number of parameters** that will be **kept**.
- **Initial cost** is the sum of weight reparametrizations **before** the first step.
- Total loss is $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{budget}}$, with $\lambda > 0$

Results

CIFAR 10

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Results on CIFAR 10 for **Conv4**, **VGG19** and **ResNet18** networks



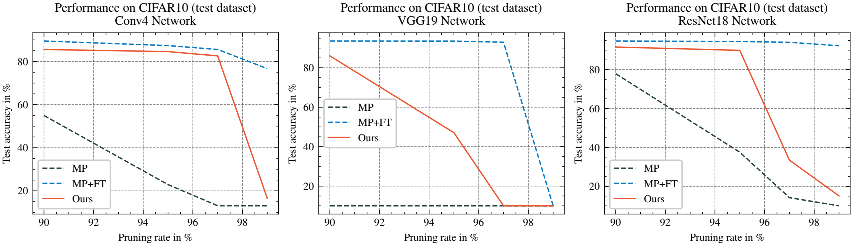
MP: Magnitude Pruning, MP+FT: Finetuned Magnitude Pruning

{Frankle et al. 2019; Simonyan at al. 2015; He et al. 2016}

CIFAR 10

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Results on CIFAR 10 for **Conv4**, **VGG19** and **ResNet18** networks



MP: Magnitude Pruning, MP+FT: Finetuned Magnitude Pruning

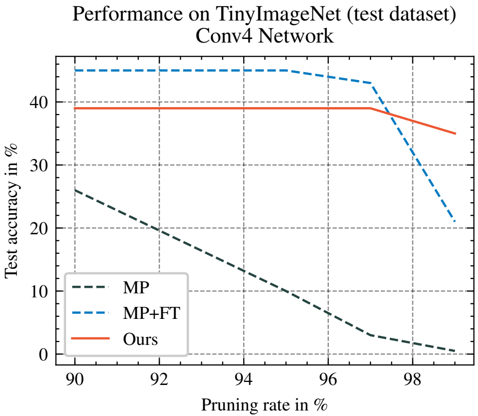
Our method **does not need finetuning**.

{Frankle et al. 2019; Simonyan at al. 2015; He et al. 2016}

TinyImageNet

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Results on TinyImageNet for **Conv4** network

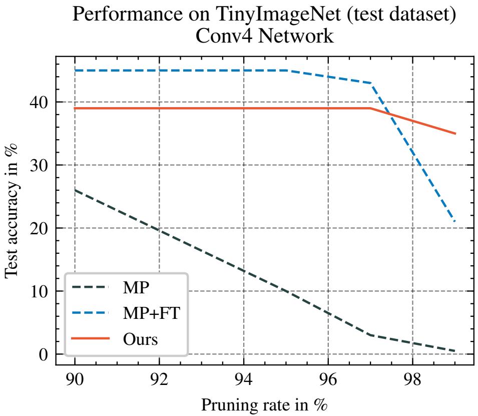


MP: Magnitude Pruning, MP+FT: Finetuned Magnitude Pruning

TinyImageNet

Netatmo SCIENCES SORBONNE UNIVERSITÉ

Results on TinyImageNet for **Conv4** network



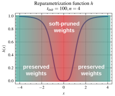
MP: Magnitude Pruning, MP+FT: Finetuned Magnitude Pruning

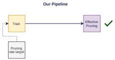
💡 Our method **does not need finetuning**.

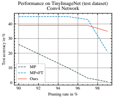
SUM UP

Key points

Key points

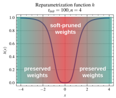




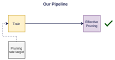


- Our reparametrization acts as a **regularizer** and a **saliency indicator**, which **induce sparsity** by **soft-pruning** the smallest weights.

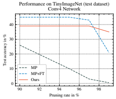
Key points



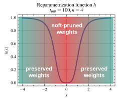
- Our reparametrization acts as a **regularizer** and a **saliency indicator**, which **induce sparsity** by **soft-pruning** the smallest weights.



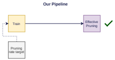
- It allows to optimize **both topology and weights** under **bugdet constraints**.



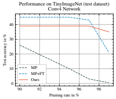
Key points



- 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.

Budget Loss

Perspectives

Perspectives

Netatmo SCIENCES SORBONNE UNIVERSITÉ

- Evaluate our method on **larger** and **more complex datasets.**

Perspectives

- Evaluate our method on **larger** and **more complex datasets.**

- Improve performances to outperform **finetuned** magnitude pruning more consistently.

Perspectives

Netatmo SCIENCES SORBONNE UNIVERSITÉ

- Evaluate our method on **larger** and **more complex datasets.**
- Improve performances to outperform **finetuned** magnitude pruning more consistently.
- Assess the **impact of the reparametrization** function and test other functions.

Thank you! 👍

Robin DUPONT
Sorbonne Université & Netatmo

Hichem SAHBI
Sorbonne Université

Guillaume MICHEL
Netatmo

✉️ robin.dupont@netatmo.com