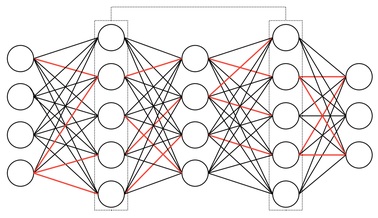
C-SWAP: Explainability-Aware Structured Pruning for Efficient NN Compression

Baptiste Bauvin [1] Loïc Baret [1] Ola Ahmad [1]

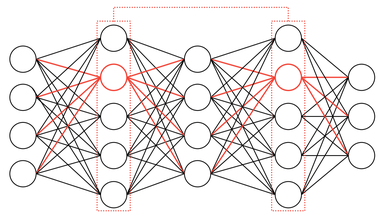
[1]**Thales**, *cortAIx Lab*, Montreal, QC, Canada

Motivation

- Neural networks are **overparametrized** ⇒ **High costs** ⇒ Issues in **edge deployment**.
- **Structured pruning** ⇒ **easy deployment** but usually **aggressive**.
- **One-shot pruning** ⇒ **efficient** but **accuracy drop**.



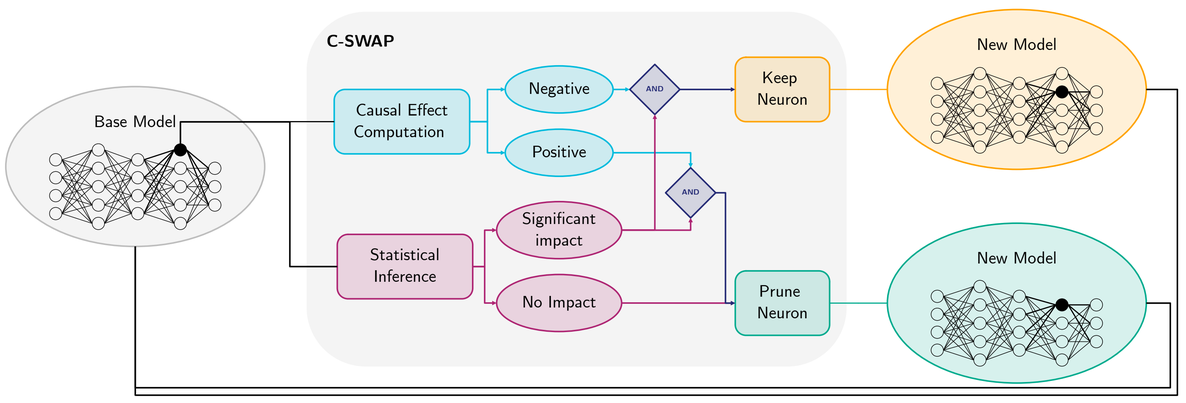
(a) Unstructured pruning (individual weights)



(b) Structured pruning (NN sub-structures)

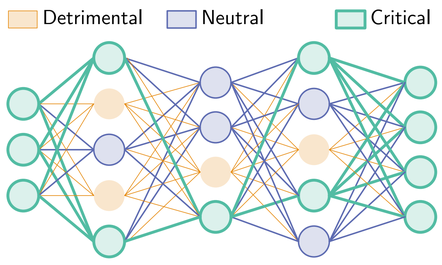
Can explainability guided structured pruning to make one-shot methods competitive without retraining?

Overview

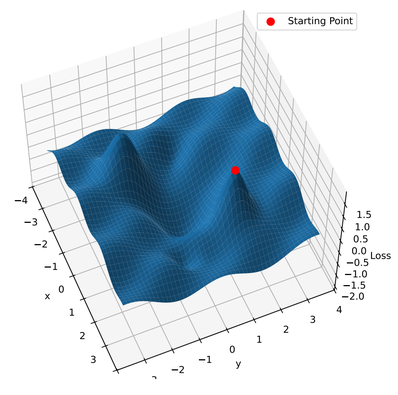


Causal Effect Computation & Statistical Inference

- Global Causal Effect: $\xi_n = \frac{1}{M}\sum_{x\in\mathcal{S}} \frac{\sigma_n^*(x)-\sigma(x)}{\sigma(x)}$, encapsulates the **positive/negative impact** of a neuron on the network.
- Significance test: T-test to compute predicate $\pi_n^{(k)}$ ⇒ True if removing neuron n has a **significant impact** on output for class k given manifold $\mathcal{S}$.
- Neuron partition:
  - If $\bigwedge_{k=1}^{C} \neg\, \pi_n^{(k)}$ ⇒ No impact ⇒ **Neutral (N)**
  - If $\xi_n < 0$ and $\bigvee_{k=1}^{C} \pi_n^{(k)}$ ⇒ Good impact ⇒ **Critical (C)**
  - If $\xi_n > 0$ and $\bigvee_{k=1}^{C} \pi_n^{(k)}$ ⇒ Bad impact ⇒ **Detrimental (D)**



Progressive Pruning



- Progressive pruning: Prune **during the analysis** if neuron not relevant enough
- Motivation: Usual ranking strategies are **efficient** but **erroneous**. Pruning is like an *optimization problem*.

😒 Rank neurons and prune:
- ⇔ **Choose** all optimization directions **from the start.**
- ⇒ *"Ranked descent"*.

😓 Re-Analyze before each prune:
- ⇔ Find **best** direction **at each step**.
- ⇒ Usual greedy descent: **very costly**.

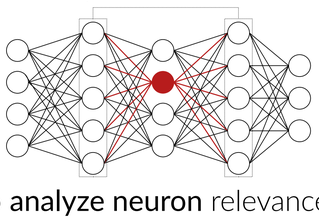
🤯 Progressive Pruning (C-SWAP):
- ⇒ Find a **adequate** direction **at each step.**
- ⇒ Relaxed greedy descent.

Substantial benefit: C-SWAP leverages **causal analysis** properties to detect **non relevant neurons during analysis**.

⇒ Robustly finds and prunes **Neutral** and **Detrimental** neurons (N & D).

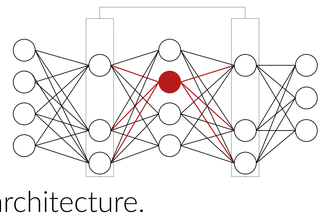
Classical XAI pruning

- For neuron in each layer:
  - Compute a XAI-based relevance score
- Rank neurons based on score
- Prune neurons up to threshold



C-SWAP

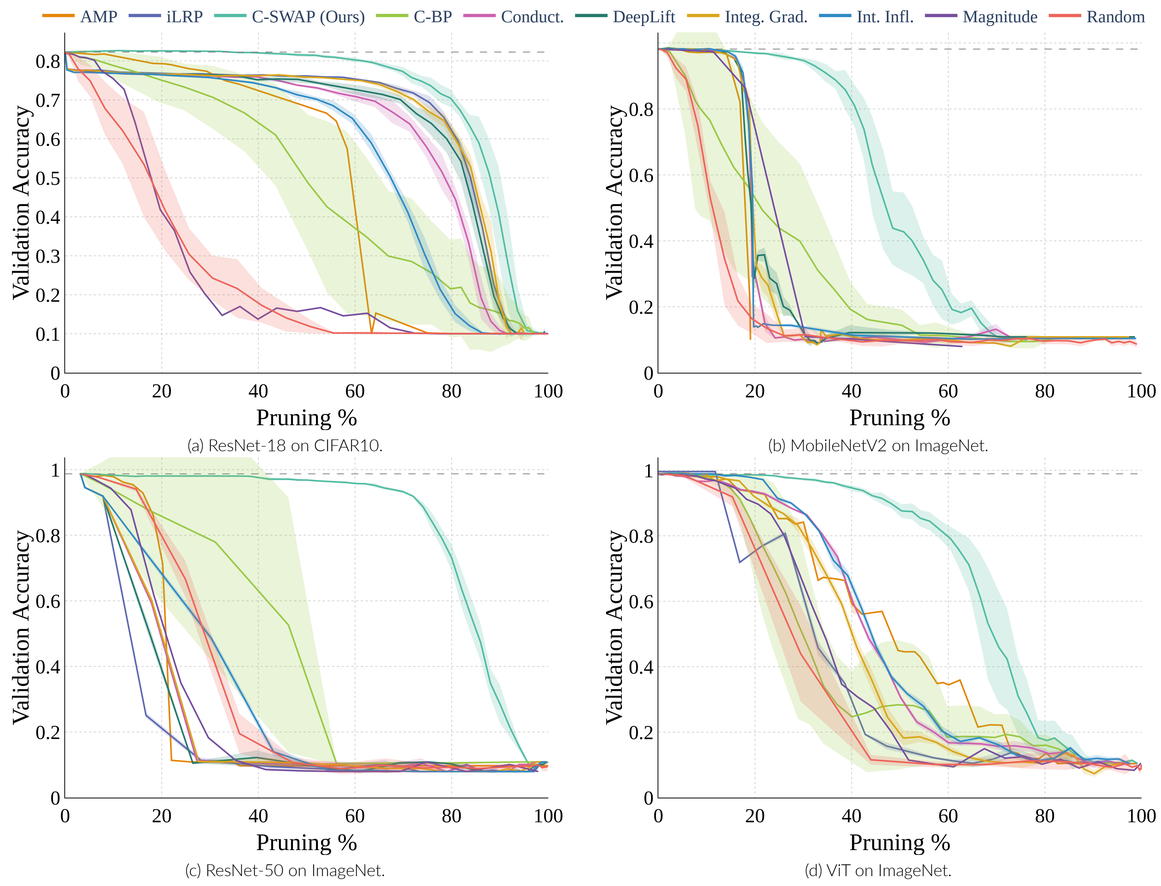
- For neuron in each layer:
  - Causal analysis : If detected as D or N: remove ; else: keep.



⇒ Allows to **analyze neuron** relevance in **up-to-date pruned** architecture.

Classification Results

- Implemented many **XAI**-based **ranking and pruning** criteria with **no finetuning**.
- Tested on small (**ResNet18**, **MobileNet**), big (**ResNet50**) and transformers (**ViT**) architectures.



(a) ResNet-18 on CIFAR10. (b) MobileNetV2 on ImageNet. (c) ResNet-50 on ImageNet. (d) ViT on ImageNet.

Figure 3. Validation accuracy as a function of percentage of parameters removed for four architectures. Dashed line represents original model performance.

- **SAUCE** score computes the **area under the pruning curve**.
- C-SWAP **safeguards** the **critical** substructures of the network.
- Prunes **more** with **less impact** on performance.
- **XAI**-based **ranking then pruning** & baselines **depletes performance**, particularly in **complex architectures**.

	ResNet18 (C10)	ResNet50 (IN)	MobileNet (IN)	ViT (IN)	Aver.
Random	25.0 ± 2.13	30.8 ± 1.28	19.8 ± 0.81	34.2 ± 1.72	27.4
OMP	26.0 ± 0.0	24.0 ± 0.0	26.3 ± 0.0	40.1 ± 0.0	29.1
AMP	49.6 ± 0.0	24.6 ± 0.0	17.3 ± 0.0	42.7 ± 8.16	33.6
DeepLIFT	64.1 ± 0.67	21.2 ± 0.19	28.6 ± 0.25	-	37.9
C-BP (ours)	49.5 ± 5.44	42.3 ± 9.73	28.7 ± 2.43	38.5 ± 5.19	39.7
Conductance	61.5 ± 0.61	22.4 ± 0.26	27.1 ± 0.07	48.1 ± 0.13	39.8
Integ. Grad.	65.5 ± 0.14	22.4 ± 0.26	27.6 ± 0.15	44.0 ± 0.35	39.9
iLRP	65.6 ± 0.18	18.2 ± 0.18	-	36.9 ± 0.25	40.3
Inter. Infl.	55.2 ± 0.45	29.1 ± 0.23	27.6 ± 0.09	49.3 ± 0.09	40.3
C-SWAP (Ours)	$\mathbf{72.1} \pm 0.39$	$\mathbf{80.3} \pm 0.28$	$\mathbf{49.4} \pm 1.15$	$\mathbf{68.1} \pm 1.67$	**67.5**

Table 1. **SAUCE** measuring the strength of each pruning criterion.

Extension: Semantic Segmentation

- DDRNet on cityscapes: Focused on **dense** NN on a **generic** problem.
- Adapted score function to **segmentation** ⇒ $\sigma(.) = \text{IoU}(.)$
- C-SWAP much more **relevant** than baselines.
- **Segmentation** task is **not easy** to prune for baselines.

	SAUCE
Random	10.3 ± 1.2
AMP	15.0 ± 0.0
OMP	17.9 ± 0.0
C-SWAP	$\mathbf{36.9} \pm 0.44$

Table 2. SAUCE scores for DDRNet on cityscapes.

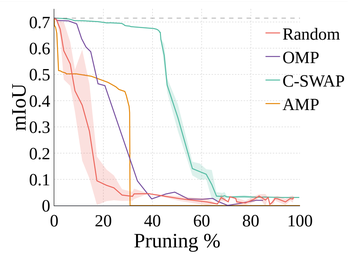
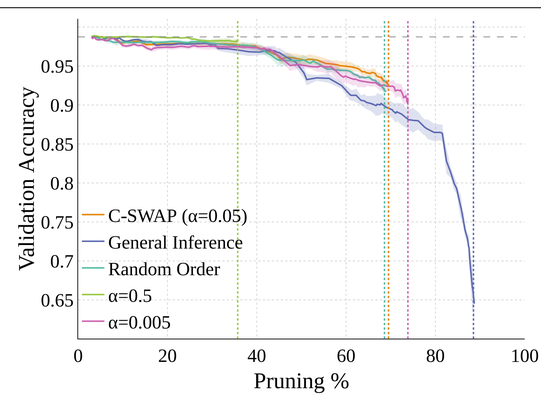


Figure 4. DDRNet on cityscapes.

Ablation Study



- C-SWAP: Our proposed method
- General inference: Statistical inference over **all samples**, instead of on **each class**.
- Random order: Exploring neurons in **random order** instead of magnitude-based order
- α variations: Modifying the **sensitivity** of the statistical inference.

🪶 Prune only **Neutral** and **Detrimental** neurons.

- **General inference** detects too many neurons as *Neutral* or *Detrimental*.
- **Random order** has limited impact on the method.
- α **variations** impact how neurons are classified as *Critical*.

Conclusion and Future Work

✅ C-SWAP novel **one-shot structured pruning** method based on **progressive pruning** and leveraging the **robustness** of **causal inference**.

✅ Achieves the compression to an **optimal trade-off** between network **complexity and performance**.

✅ Shows the **critical role** of **XAI** in advancing DNN **compression**.

🔍 Importance of individual neurons might be **too fine-grained** for very large models,

🔍 Mitigation by computing relevance based on **neuron groups** for large layers.