When XAI meets Compression & Sub-graph Discovery

Pruning By **Explaining** Revisited: Optimizing Attribution Methods to Prune **CNN**s & **Transformer**s

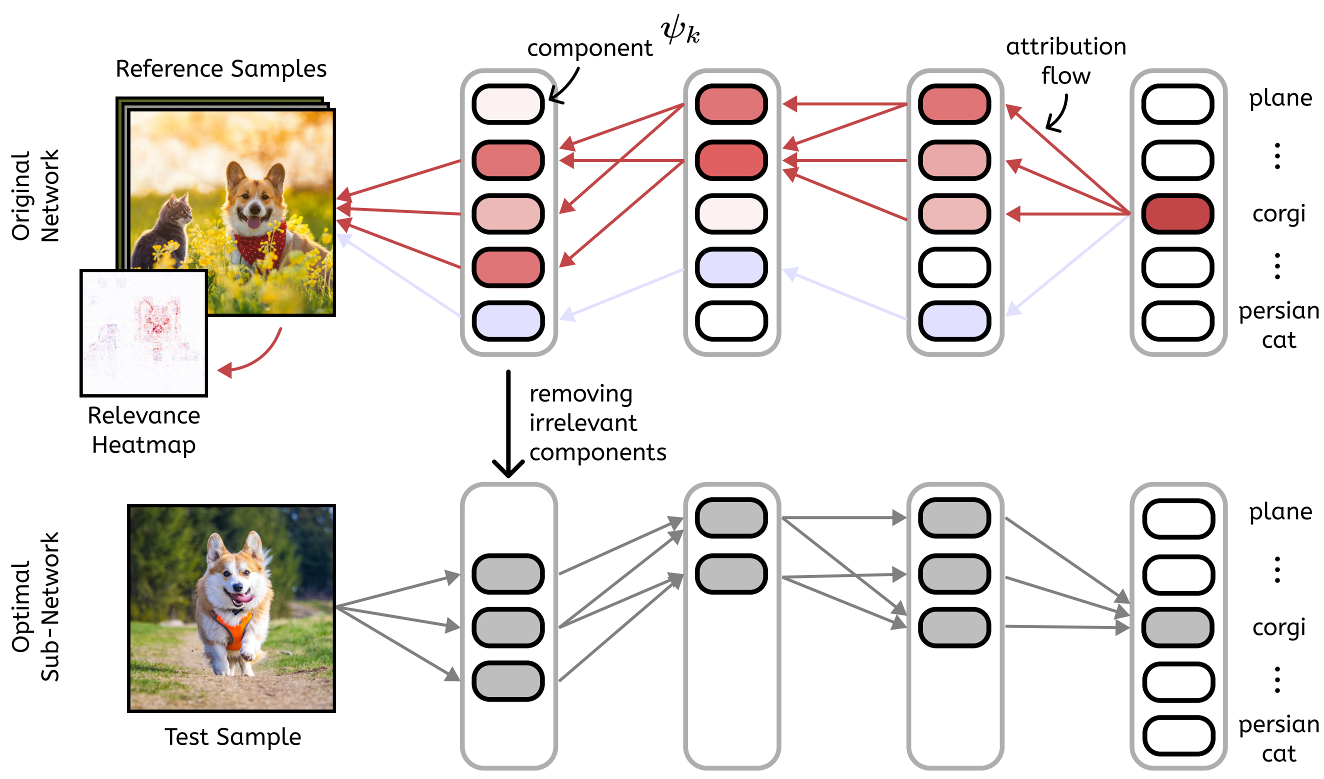
Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, Sebastian Lapuschkin

Fraunhofer HHI · BIFOLD · Technische Universität Berlin · iToBoS · TEMA

ECCV

PXP Pruning by eXplaining in PyTorch

Pruning by Explaining



⟶ Our Pruning Framework

Given **a set of reference samples** $\mathcal{X}_{\text{ref}}$ defined as:

$$\mathcal{X}_{\text{ref}} = \{x_1, x_2, \dots, x_{n_{\{\text{ref}\}}}\}$$

Importance score of a component ψ_k can be computed by:

$$\bar{R}_{\psi_k} = \frac{1}{n_{\text{ref}}} \sum_{i=1}^{n_{\text{ref}}} R_{\psi_k}(x_i)$$

But, how should we compute R_{ψ_k}? In other words, **what is a reliable pruning criterion?**

+ Use relevance scores of Layer-wise Relevance Propagation:

$$R_{i \leftarrow j}^{(l-1,l)} = \frac{z_{ij}}{z_j} R_j^l$$

What is an **advantage** of this criterion?

+ LRP's relevance scores are intrinsically **normalized** due to their conservation property across layers.

How large should be the set of reference samples $\mathcal{X}_{\text{ref}}$?

+ The more samples used for attribution, the more stable the pruning is. However, for **CNNs**, the work of [1] has shown that **10 reference samples** per class is sufficient.

+ For **Transformers** on the other hand, our experiments conveyed that **only 1 reference sample** generates robust relevance scores for pruning.

⟶ Optimization of XAI Methods

Typically takes place to generate faithful explanations, but solutions are **not necessarily optimal for pruning**.
So, why don't we **optimize XAI for pruning** directly?

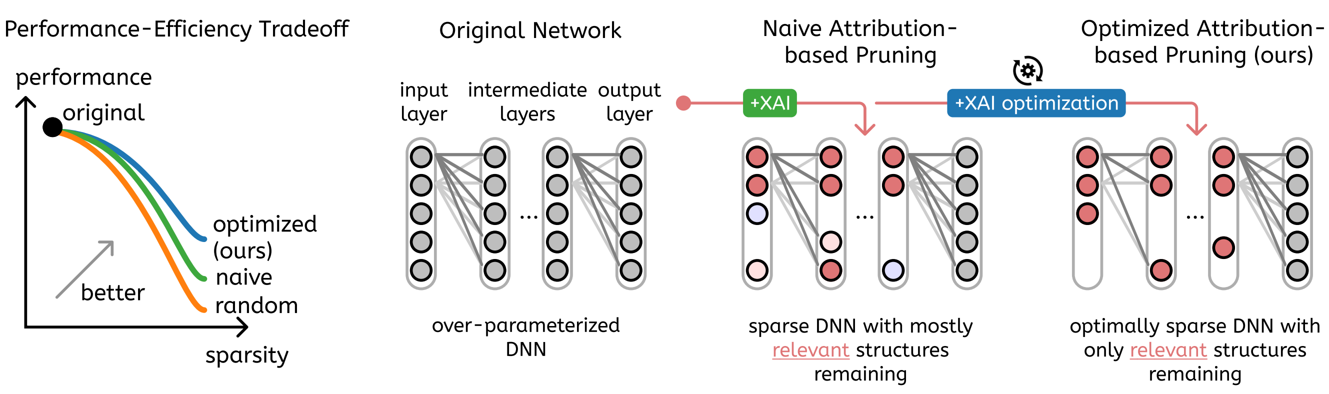
⟶ CNNs and Transformers in Pruning

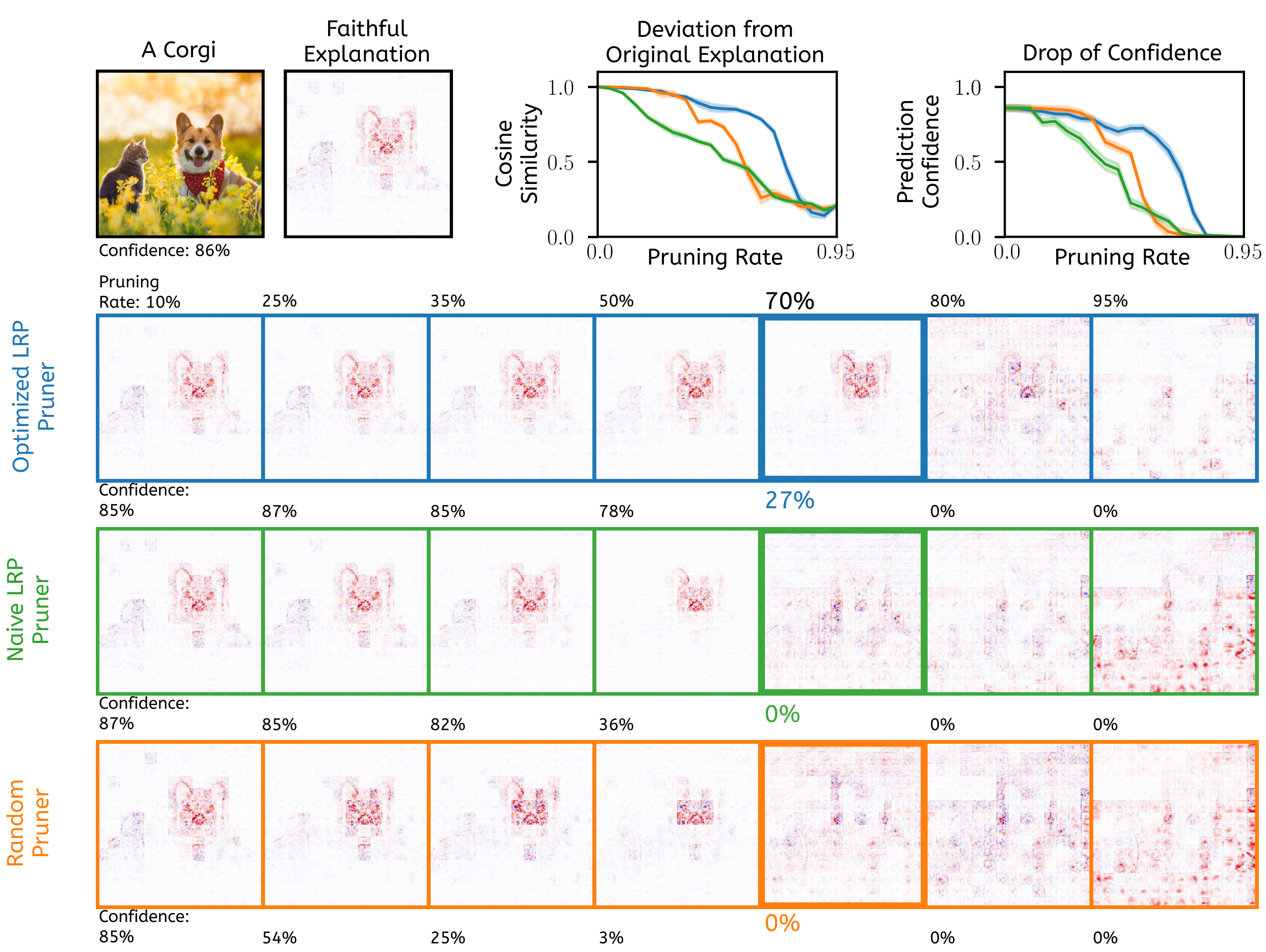
CNNs have quite **sufficient amount of parameters** and thus not much can be pruned from them based on the default task (i.e., ImageNet 1000-class classification). **Explanations that faithfully attribute CNNs**, perform well on pruning as well.

Transformers are typically more **overparameterized** than CNNs, which induce more pruning rates while keeping high performance given the default task. Unlike CNNs, **a faithful explainer** of Transformers **does not guarantee stable pruning**, thus **encouraging extra optimization** of explainer.

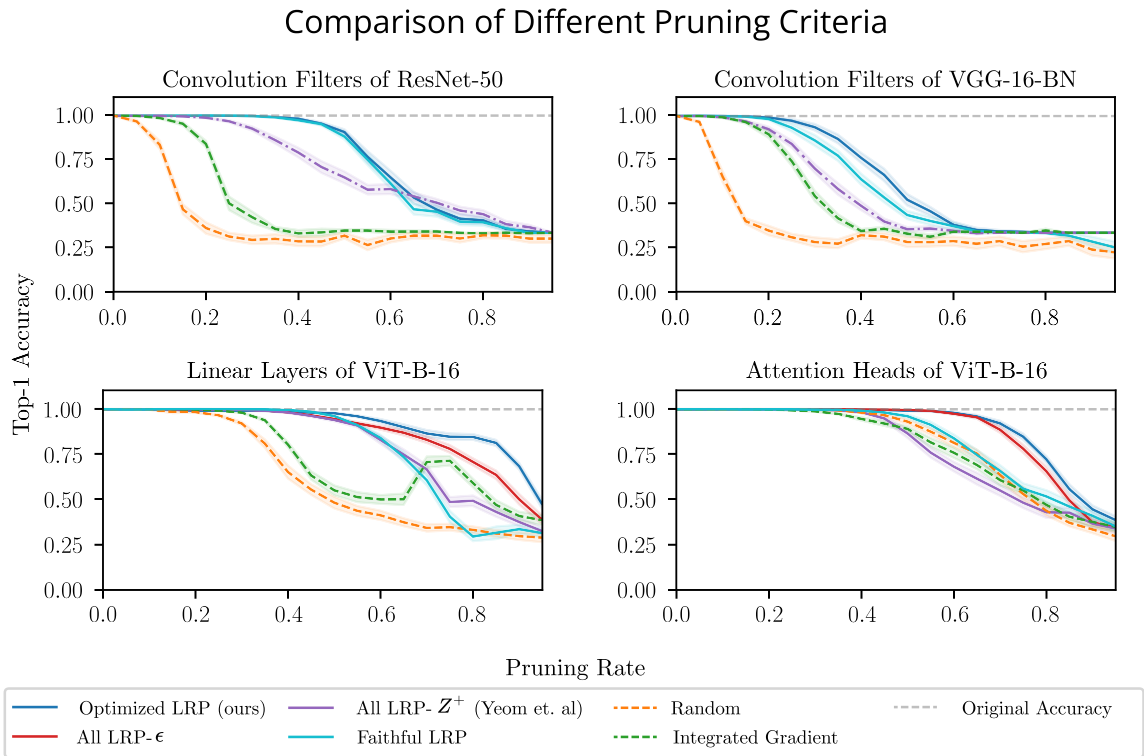
Overall, **LRP-Epsilon** [2, 3, 4] is a **promising explainer for pruning** across different architectures.

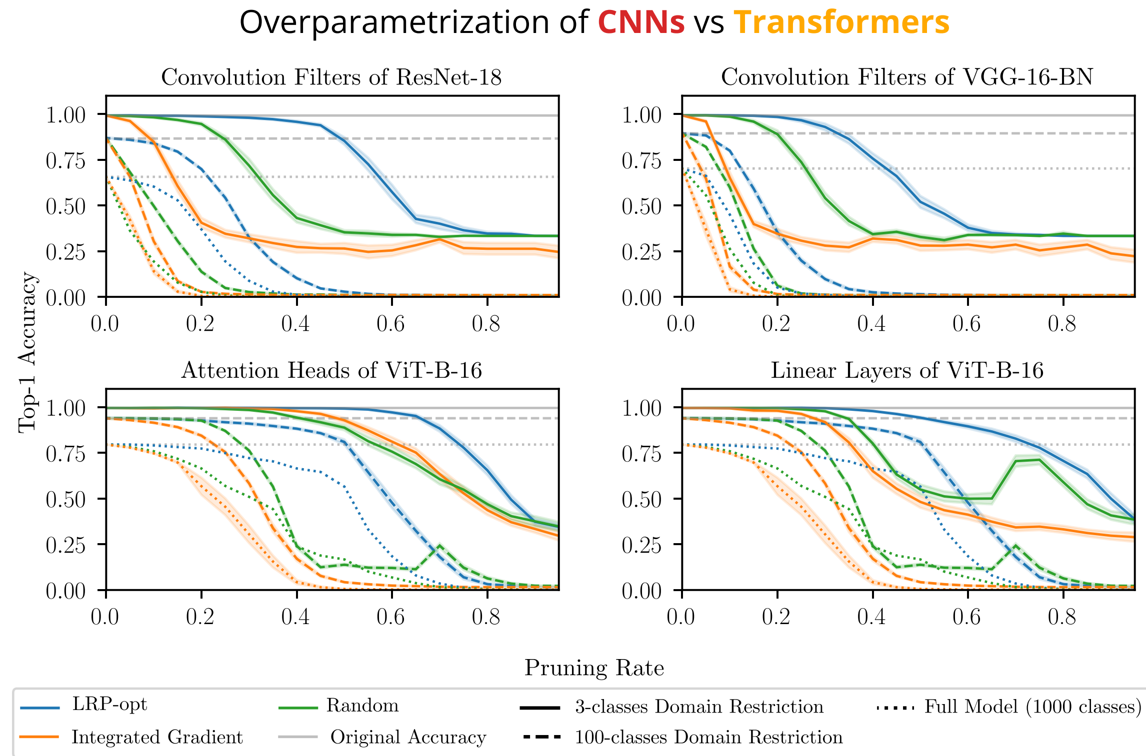
Optimize XAI for Pruning





Experiments





References

[1] Yeom et al. **Pruning by explaining: A novel criterion for deep neural network pruning**. Pattern Recognition 115, 107899 (2021)

[2] Bach et al. **On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation**. PloS one 10(7), e0130140 (2015)

[3] Montavon et al. **Layer-wise relevance propagation: an overview**. Explainable AI: interpreting, explaining and visualizing deep learning pp. 193–209 (2019)

[4] Achtibat et al. **AttnLRP: Attention-aware layer-wise relevance propagation for transformers**. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 135–168. PMLR (21–27 Jul 2024)

arXiv