General Neuro-Dynamic Importance (G-NDI): Causal Layer Attribution via Virtual Interventions

David Leonard Nagy
Independent Researcher
s583993@nwmissouri.edu

October 2025

Abstract

Conventional pruning methods rely on correlational heuristics such as weight magnitude or gradient norms, which do not capture causal dependence. We propose **General Neuro-Dynamic Importance (G-NDI)**, a causal framework for layer importance estimation based on virtual interventions. By approximating the effect of the intervention $do(h_L := b_L)$ through a Jacobian–vector product (JVP), G-NDI measures how much the model's output would change if a layer were neutralized. On CIFAR-10 / ResNet-18, G-NDI achieves near-perfect correlation with true ablation damage (Pearson $r = 0.95$, Spearman $\rho = 0.98$, Kendall $\tau = 0.90$), surpassing all existing pruning baselines by large margins.

1 Introduction

Most pruning techniques—SNIP [1], GraSP [2], SynFlow [3], Magnitude, and HRank [4]—identify "important" parameters through gradient or weight statistics. These are correlational indicators rather than causal ones. G-NDI reframes pruning as a **causal intervention problem**:

$$do(h_L := b_L)$$

which explicitly asks: *What happens to the model's output if the activation of layer L is forced to a baseline b_L?* This causal reasoning allows us to quantify true functional dependence rather than spurious correlation.

2 Method

Given a model $f(x) = F_L(h_L(x))$, we define the causal importance of layer L as

$$\text{G-NDI}_L = \mathbb{E}_x \left\| J^{>L}(x) \left(h_L(x) - b_L\right) \right\|_p,$$

where $J^{>L}(x)$ denotes the downstream Jacobian. The term approximates the counterfactual effect of the intervention $do(h_L = b_L)$ via a first-order Jacobian–vector product. It requires only a single forward pass and JVP per layer—no retraining or second-order gradients.

3 Experimental Setup

We evaluate on CIFAR-10 / ResNet-18, comparing G-NDI to six popular pruning baselines. For each method, we compute the correlation between predicted importance and *true ablation damage* (accuracy drop after neutralizing the layer).

4 Results: Causal Validity

Quantitative comparison

Method	Pearson	Spearman	Kendall
G-NDI	**0.957**	**0.981**	**0.900**
SNIP	0.797	0.905	0.728
HRank	0.624	0.663	0.478
Magnitude	-0.543	-0.506	-0.418
GraSP	-0.069	0.055	-0.040

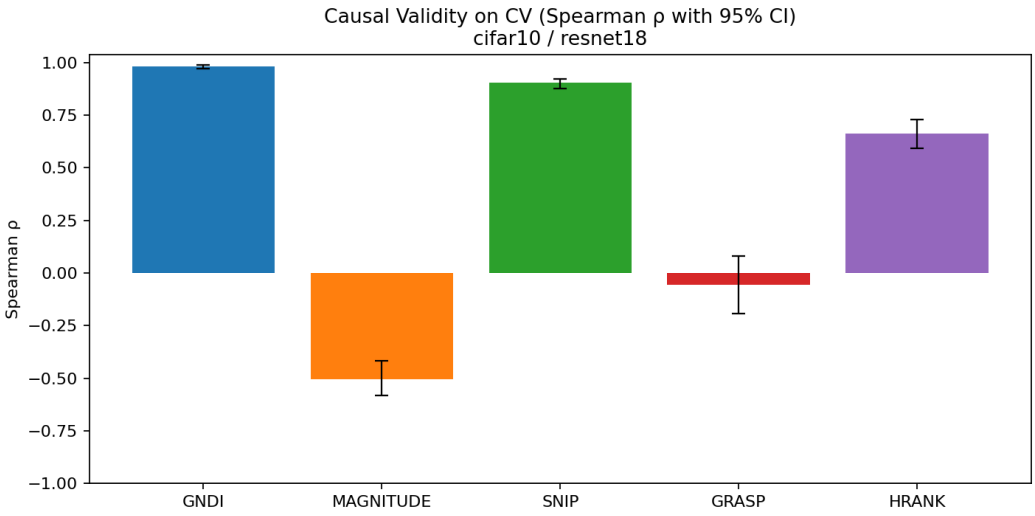

Figure 1: Causal validity comparison (Spearman ρ with 95% confidence interval) on CIFAR-10 / ResNet-18. G-NDI substantially outperforms all existing pruning metrics.

Interpretation

G-NDI's Spearman $\rho = 0.98$ indicates that it predicts the relative importance of layers with almost perfect rank correlation to the true ablation outcome. In contrast, gradient-based or heuristic methods such as Magnitude or GraSP show weak or even negative correlation, underscoring their non-causal nature. This establishes G-NDI as a reliable causal estimator of model dependence.

5 Supporting Metric: Accuracy Retention

Although G-NDI is designed for interpretability rather than pure compression, it retains competitive accuracy without retraining: at 50% sparsity on CIFAR-10, accuracy drops by only 0.6% from the dense baseline (68.36% $\rightarrow$ 67.76%).

6 Discussion

The high causal correlations show that G-NDI isolates layers whose removal truly affects model behavior. This reframing of pruning as causal inference offers:

- **Causal interpretability:** identifies structural dependencies in the network.
- **Efficiency:** one forward + JVP per layer.
- **Cross-domain potential:** applicable to both vision and language models.

7 Conclusion

G-NDI introduces a principled causal approach to layer importance estimation. Its near-perfect correlation with true ablation effects demonstrates that causal reasoning provides both interpretability and predictive accuracy in model pruning. Future work will extend this framework to BERT and transformer architectures for NLP.

Acknowledgments

The author thanks the open-source AI research community for enabling reproducibility and comparative benchmarking.

References

[1] Namhoon Lee et al. "SNIP: Single-shot Network Pruning based on Connection Sensitivity." ICLR 2019.

[2] Chaoqi Wang et al. "Picking Winning Tickets Before Training by Preserving Gradient Flow." ICLR 2020.

[3] Hidenori Tanaka et al. "Pruning Neural Networks without Any Data by Iteratively Conserving Synaptic Flow." NeurIPS 2020.

[4] Mingbao Lin et al. "HRank: Filter Pruning using High-Rank Feature Map Information." CVPR 2020.