Concept-Aware Pruning via Disentangled Subspaces for Robust Convolutional Networks

Kirin Danek · **Vikram V. Ramaswamy**
Department of Computer Science, Princeton University · *{kd9132, vr23}@princeton.edu*

PRINCETON UNIVERSITY

CVPR JUNE 3-7, 2026 DENVER COLORADO

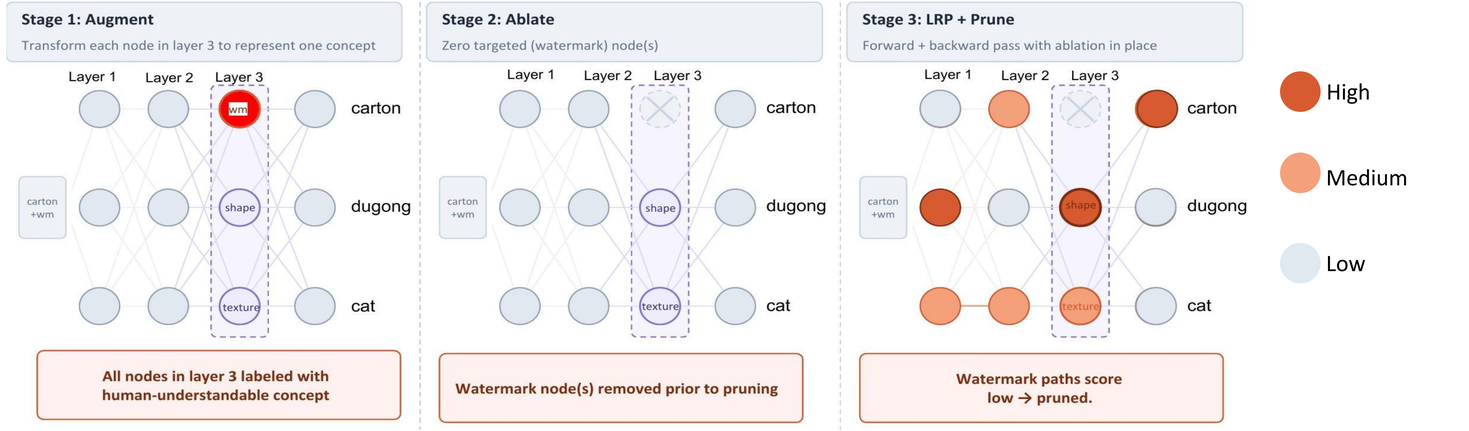
Overview

Network pruning compresses models for efficient inference, but standard pruning methods do not consider the semantic meaning of what they move, and can amplify reliance on spurious shortcuts (e.g., watermarks, backgrounds, …).

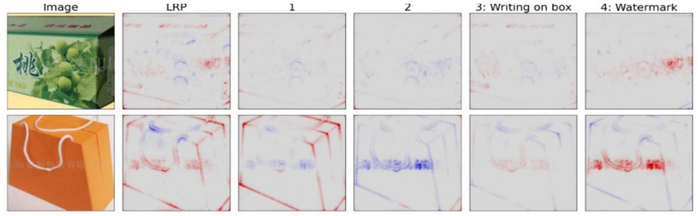
We introduce **Concept-Aware Network Pruning (CNP)**, a framework that uses concept-based explanations to make pruning *semantically targeted* — preserving robust class signal while removing filters that encode spurious concepts.

Key idea. Project layer activations onto orthogonal *concept subspaces*, zero out the spurious one, then let LRP-based pruning [2] naturally remove the filters that served it.

1. Method: Concept-Aware Pruning



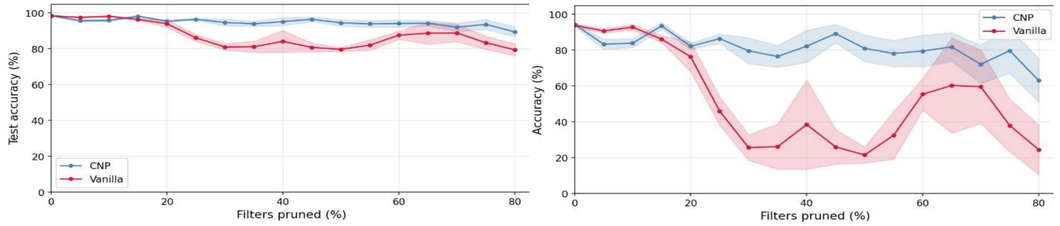
2. Identifying Subspaces with DRSA [1]



Sub	AP_{wm}	AP^{+}_{wm}	AP^{-}_{wm}	AP_{class}	AP^{wm}_{class}	AP^{nwm}_{class}
1	0.4524	0.4816	0.3131	0.9845	0.9928	0.9838
2	0.3129	0.3219	0.3080	0.5440	0.6342	0.5782
3	0.6122	0.5691	0.9210	0.8378	0.8177	0.8962
4	**0.9651**	**0.9269**	**1.0000**	**0.5041**	**0.5221**	**0.5472**

S4 isolates the watermark (AP=0.97) while S1-S3 capture valid content

3. Results: ImageNet Robustness

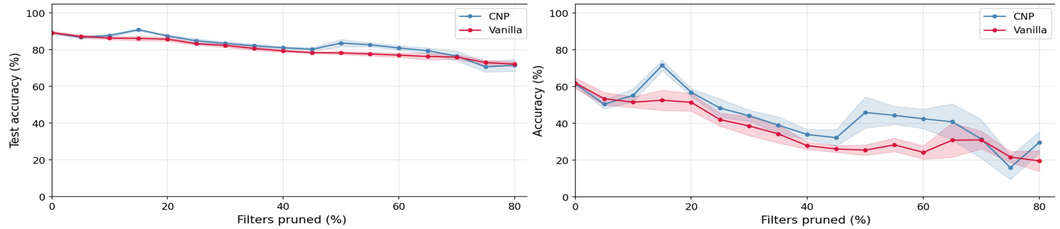


"Carton" vs. "Dugong," VGG-16 backbone, n=5 runs. (left) Balanced acc. (right) Dugong + WM only.

(3a) In ImageNet [3], class "Carton" is spuriously associated with watermarks; class "Dugong" is not.

- Ablate watermark subspace; prune with CNP vs. vanilla LRP-pruning [2].

✓ CNP recovers up to **+60% accuracy** on worst-performing subgroup with no loss on balanced accuracy.



"Crate" vs. "Packet," VGG-16 backbone, n=5 runs. (left) Balanced acc. (right) Packet + WM only.

(3b) "Crate" associated with watermarks; "Packet" not.

- Harder ImageNet pair; unpruned model is **heavily watermark-reliant** (~60% OOD acc).
- Watermark subspace **only weakly predictive** (AP ≈ 0.72 vs. 0.97 above).
- CNP slightly improves over vanilla, but still heavily relies on watermark.

4. Limitations and Future Work

- **Problem:** Concept representations might shift during intermediate fine-tuning.
  - **Future:** re-learn subspaces after each fine-tuning iteration.
- **Problem:** PCA, DRSA [1] not accurately representing/disentangling concepts.
  - **Future:** Replace with Concept Activation Vector-style intervention
- **Future:** Demographic fairness applications (gender, race, …).

References

[1] P. Chormai, et al. *Disentangled explanations of neural network predictions by finding relevant subspaces.* TPAMI 2024.

[2] S.-K. Yeom, et al. *Pruning by explaining: A novel criterion for deep neural network pruning.* 2019.

[3] J. Deng, et al. *ImageNet: A large-scale hierarchical image database.* 2009.

Acknowledgements: We acknowledge funding from the Princeton SEAS Innovation grant and Princeton SEAS Project X.