Do Sparse Subnetworks Exhibit Cognitively Aligned Attention? Effects of Pruning on Saliency Map Fidelity, Sparsity, and Concept Coherence

Sanish Suwal*
Rochester Institute of Technology
Rochester, NY
ss4657@rit.edu

Dipkamal Bhusal*
Rochester Institute of Technology
Rochester, NY
db1702@rit.edu

Michael Clifford
Toyota InfoTech Labs
Mountain View, CA
michael.clifford@toyota.com

Nidhi Rastogi
Rochester Institute of Technology
Rochester, NY
nxrvse@rit.edu

Abstract

Prior works have shown that neural networks can be heavily pruned while preserving performance, but the impact of pruning on model interpretability remains unclear. In this work, we investigate how magnitude-based pruning followed by fine-tuning affects both low-level saliency maps and high-level concept representations. Using a ResNet-18 trained on ImageNette, we compare post-hoc explanations from Vanilla Gradients (VG) and Integrated Gradients (IG) across pruning levels, evaluating sparsity and faithfulness. We further apply CRAFT-based concept extraction to track changes in semantic coherence of learned concepts. Our results show that light-to-moderate pruning improves saliency-map focus and faithfulness while retaining distinct, semantically meaningful concepts. In contrast, aggressive pruning merges heterogeneous features, reducing saliency map sparsity and concept coherence despite maintaining accuracy. These findings suggest that while pruning can shape internal representations toward more human-aligned attention patterns, excessive pruning undermines interpretability. Code is available at https://github.com/sanishsuwal7/Neurips-CogInterp/.

1 Introduction

Humans often rely on a sparse subset of cues to make decisions, focusing attention on the most salient and semantically meaningful aspects of a scene. In contrast, deep neural networks are often criticized for relying on hard-to-interpret features. Neural network pruning is a widely studied approach for removing unnecessary weights or structure from overparameterized models [9, 11]. By eliminating low-magnitude weights, pruning can drastically reduce model size and computation with minimal accuracy loss. In some cases, models like ResNet-50 can be pruned by 80–90% and still retain performance comparable to their dense counterparts [7], suggesting that much of a trained network's capacity may be redundant.

While pruning's impact on efficiency is well understood, its effect on interpretability is far less explored. In high-stakes domains such as healthcare or autonomous driving, models must not only be

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models.

accurate but also explainable in a way that aligns with human reasoning. Unfortunately, post-hoc saliency methods often produce noisy, unfaithful explanations that misrepresent the model's actual decision process [1, 2].

In this work, we investigate how pruning affects interpretability across two levels of abstraction: **a) Low-level attribution maps**, evaluated in terms of sparsity: *the concentration of attribution on a small set of relevant input features* [3], and faithfulness: *how accurately explanations reflect the features that influence predictions* [16]. **b) High-level concept representations**, assessed qualitatively by examining how the most important extracted concepts change in appearance, and coherence across pruning levels, using concept extraction method such as CRAFT [5].

We hypothesize that pruning acts as a structural regularizer, eliminating redundant pathways and forcing the model to rely on a smaller, more essential set of discriminative features. This, in turn, could yield more focused saliency maps and cleaner, semantically distinct concepts. However, aggressive pruning may compress multiple discriminative features into fewer activation patterns, reducing concept coherence even if low-level sparsity improves.

To test these hypotheses, we conduct a systematic study using a ResNet-18 [10] trained on ImageNette [14]. We apply global magnitude pruning with iterative fine-tuning following the lottery ticket hypothesis framework [7], and evaluate post-hoc explanations using Vanilla Gradients (VG)[17] and Integrated Gradients (IG)[19]. At the concept level, we use CRAFT [5] to measure qualitative changes in semantic coherence of discovered concepts.

Our results show that light-to-moderate pruning improves saliency-map focus and faithfulness while preserving distinct, semantically meaningful concepts. In contrast, aggressive pruning blurs concept boundaries, merging heterogeneous visual patterns despite maintaining accuracy. These findings suggest that appropriate amount of pruning can shape internal representations toward more human-aligned attention, but excessive pruning undermines the quality of learned concepts, highlighting a nuanced trade-off between sparsity, faithfulness, and semantic coherence.

2 Related Work

The Lottery Ticket Hypothesis [7] showed that sparse subnetworks can be retrained to match full-model accuracy. Later, Frankle and Bau [6] used Network Dissection to find that heavy pruning of ResNet-50 preserves most human-recognizable concepts. Hooker et al. [12] noted that compressed models may forget certain examples, though they did not assess explanation quality. Weber et al. [22] observed reduced noise in GradCAM maps after moderate pruning in VGG-16, but without quantitative rigor or fine-tuning. Tan [21] reported that pruning without fine-tuning can collapse explanations despite stable predictions. Suwal et al. [20] studied the impact of adversarial training and pruning on saliency maps of vehicular datasets. Our work differs by quantitatively evaluating post-hoc explanations under magnitude pruning with fine-tuning, using ROAD and Gini metrics, and extending analysis to concept-level changes, using automatic concept extraction method, CRAFT [5].

3 Methodology

We study how iterative magnitude-based pruning and fine-tuning affect both pixel-level explanations (saliency maps) and high-level concept representations. Our base model is ResNet-18 [10] trained on the ImageNette dataset [14], a 10-class subset of ImageNet [4] designed for fast benchmarking.

Table 1: Model performance on ImageNette test-set across several pruning levels

Pruning %	0	10	20	30	50	70
Accuracy	84.15	83.08	84.31	85.58	84.99	83.99

3.1 Pruning Strategy

The Lottery Ticket Hypothesis [7] shows that, within a randomly initialized, dense neural network, there exist sparse subnetworks, that when trained in isolation from their original initialization, can match or exceed the accuracy of the full network in the same number of training iterations. A winning

ticket is thus a subnetwork whose initial weights and structure are sufficient to learn the target task without relying on the excess parameters in the original model.

Following this framework, we start with a ResNet-18 [10] model randomly initialized with parameters, θ, and train it to convergence on the ImageNette dataset [14]. After training, we perform global unstructured magnitude pruning to remove a fixed proportion $p\%$ of the smallest-magnitude weights across all layers (excluding biases), producing a binary pruning mask M. The surviving weights $\theta \cdot M$ are then reset to their original initialization values from θ, yielding a winning ticket candidate, which is fine-tuned for preserving model accuracy. We repeat the prune–fine-tune cycle for n iterations, progressively increasing weight-sparsity while maintaining high predictive performance.

Classification accuracy on the ImageNette test set remains within 1–2 percentage points of the unpruned baseline up to 70% weight removal (Table 1), consistent with lottery ticket hypothesis findings that substantial weight-sparsity can be introduced without significant performance loss.

3.2 Post-hoc explanation and concept extraction

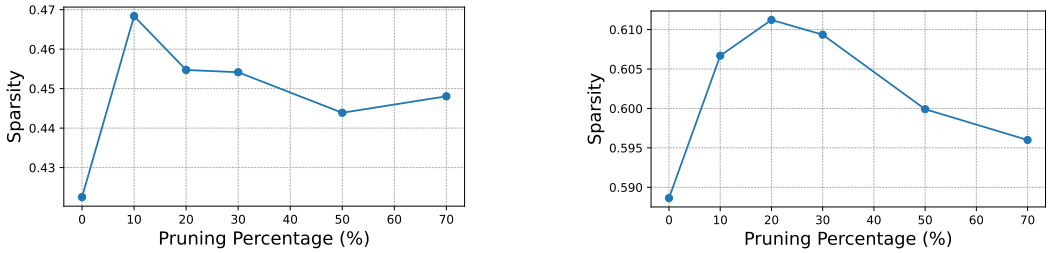
For pixel-level explanations, we generate saliency maps using Vanilla Gradients (VG) [17] and Integrated Gradients (IG) [19]. We quantify their quality using two metrics: **1) Sparsity** [3], measured via the Gini coefficient, which captures the concentration of attribution values in a small set of pixels. Higher values indicate sparse and comprehensible saliency maps. **2) Faithfulness**, measured using ROAD MoRF strategy [16]. Features are progressively removed in decreasing order of importance, and model accuracy is recorded at each step. Unlike Insertion/Deletion [15] or ROAR [13], ROAD avoids distribution shifts from synthetic perturbations and does not require retraining. We also measure faithfulness using the area over the perturbation curve (AOPC), where higher values correspond to more accurate identification of critical features. Further metric details are provided in Appendix B.

For high-level concept-level analysis, we adopt the CRAFT pipeline [5] to extract the top-ranked concepts for selected classes at different pruning stages. We use a patch size of 64 and concept extraction of 10 ranks (default in the official implementation), and follow the publicly available code[2]. We then qualitatively compare the discovered concepts before and after pruning, noting changes in semantic composition. Further details on CRAFT is provided in Appendix C.

4 Results

4.1 Saliency maps

Sparsity. Figure 1 shows saliency sparsity (Gini index) for Vanilla Gradients (VG) and Integrated Gradients (IG) as pruning increases. For both methods, sparsity rises with pruning, indicating more concentrated attributions in pruned models. VG peaks near $\sim$10% pruning and IG near $\sim$20%, after which improvements plateau. Across all levels, IG maps are consistently *absolutely* sparser than VG, reflecting IG's built-in tendency to concentrate attribution scores.



(a) Vanilla Gradient (VG)

(b) Integrated Gradients (IG)

Figure 1: Sparsity evaluation of saliency maps.

[2] https://github.com/deel-ai/Craft

Faithfulness. We assess faithfulness with ROAD–MoRF curves (Fig. 2) and summarize with AOPC scores (Fig. 3). A steeper accuracy drop under MoRF indicates that removed, high-ranked features were truly critical.

Vanilla Gradients (VG). Pruning produces *sharper* ROAD curves than the natural model across most levels, with clear gains already at 10–20% pruning. Additional pruning maintains or slightly reduces these gains. AOPC mirrors this trend: it peaks at 20% and remains above baseline thereafter.

Integrated Gradients (IG). Faithfulness is largely unchanged at $\leq$20% pruning: the curves closely track the unpruned model, and early drops can be slightly shallower. Improvements emerge beyond 30%, with the largest gains at heavy pruning (50–70%), where curves are steepest and AOPC reaches its maximum.

(a) Vanilla Gradient (VG)

(b) Integrated Gradients (IG)

Figure 2: Faithfulness evaluation. Sharper drop in model accuracy signifies faithful saliency maps.

Takeaway. Pruning's faithfulness gains are *method-dependent*: VG benefits at light-to-moderate sparsity (peaking at $\sim$20%), whereas IG requires moderate to higher pruning.

(a) Vanilla Gradient (VG)

(b) Integrated Gradients (IG)

Figure 3: Faithfulness evaluation using AOPC of ROAD-MoRF plots.

Qualitative maps. Visual inspection (provided in Appendix A) aligns with these patterns: 10–30% pruning sharpens object focus for both VG and IG; $\geq$50% pruning begins to reintroduce background noise and reduces the apparent sparsity gains.

4.2 Concept analysis under pruning

We complement pixel-level metrics with a concept-level probe using CRAFT [5] on the "parachute" class (Appendix D). The unpruned model's top concepts are highly distinctive and semantically pure, dominated by brightly colored canopy sections against clear skies, with secondary cues such as fabric color blocks. Light pruning (10–20%) largely preserves these object-centric patterns, concentrating importance on zoomed-out parachute shots and bold canopy stripes while reducing the variety of environmental cues. From 30% pruning onward, concept coherence declines: object-relevant features increasingly co-occur with unrelated textures, water scenes, or fabric folds, and background reliance becomes more pronounced. By 70% pruning, the top concept still contains parachute cues, but most remaining concepts are mixed or irrelevant, including abstract textures, text, and structural imagery, indicating a substantial loss of semantic clarity.

These results suggest that moderate pruning can reweight the model toward high-confidence object cues while retaining coherent concept structure, but heavy pruning forces disparate visual patterns

into fewer activation clusters, degrading interpretability despite occasional improvements in saliency sparsity.

5 Conclusion

This work examined how structured magnitude pruning influences the quality of post-hoc saliency explanations and concept-based interpretations, focusing on sparsity, faithfulness, and coherence. Across ImageNette experiments with ResNet-18, we observed that moderate pruning increases saliency sparsity and faithfulness, and compact concept activations. These results suggest pruning not only compresses models but can also improve post-hoc explanations.

Acknowledgment

This work was supported by Toyota InfoTech Labs through Unrestricted Research Funds.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *NeurIPS 2018*.

[2] Dipkamal Bhusal, Rosalyn Shin, Ajay Ashok Shewale, Monish Kumar Manikya Veerabhadran, Michael Clifford, Sara Rampazzi, and Nidhi Rastogi. Sok: Modeling explainability in security analytics for interpretability, trustworthiness, and usability. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–12, 2023.

[3] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. Concise explanations of neural networks using adversarial training, 2020. URL `https://arxiv.org/abs/1810.06583`.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[5] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023.

[6] Jonathan Frankle and David Bau. Dissecting pruned neural networks. *ICLR Workshop*, 2019.

[7] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[8] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[9] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*, 2018.

[12] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *NeurIPS*, 2019.

[13] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Jeremy Howard. imagenette. URL https://github.com/fastai/imagenette/.

[15] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 151. BMVA Press, 2018. URL http://bmvc2018.org/contents/papers/1064.pdf.

[16] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18770–18795. PMLR, 2022.

[17] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

[18] Ilya M Sobol. Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.*, 1(4):407–414, 1993.

[19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

[20] Sanish Suwal, Shaurya Garg, Dipkamal Bhusal, Michael Clifford, and Nidhi Rastogi. Smaller is better: Enhancing transparency in vehicle ai systems via pruning. *arXiv preprint arXiv:2509.20148*, 2025.

[21] Hanxiao Tan. Evaluating explanation robustness to model pruning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[22] David Weber, Florian Merkle, Pascal Schöttle, and Stephan Schlögl. Less is more: The influence of pruning on the explainability of cnns. *arXiv preprint arXiv:2302.08878*, 2023.

A Qualitative result

Figures 4, 5, and 6 present Vanilla Gradient (VG) and Integrated Gradients (IG) saliency maps for the normal and pruned models. Across all examples, VG tends to produce noisier and less focused explanations compared to IG, but pruning influences both methods.

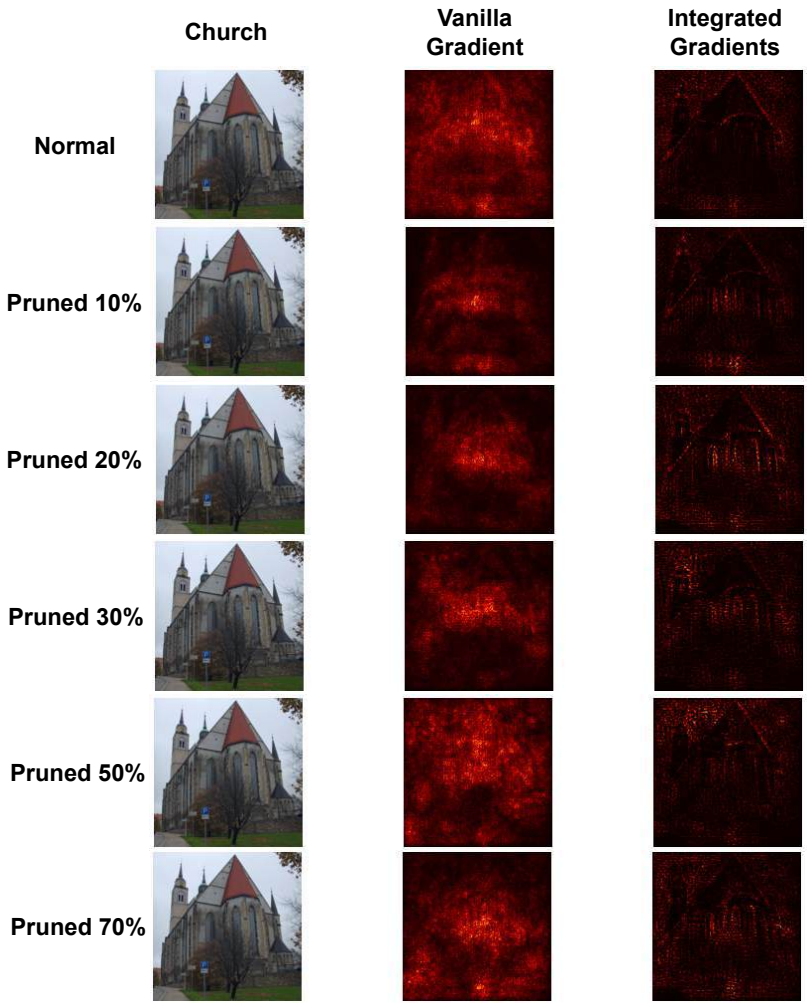


Figure 4: Church

In Church (Fig. 4), with VG, noise is prominent for the normal model and persists after pruning, though slight visual sharpening occurs up to 30% pruning. IG produces clearer focus on the frontal architecture in the pruned models; however, at 50–70% pruning, maps become noisier, aligning with the sparsity drop seen in quantitative results.

In Springer (Fig. 5), VG explanations remain diffuse and noisy across all pruning levels. IG maps, however, become sharper and more focused up to 30% pruning, before degrading in clarity at higher pruning levels. This mirrors the trend where pruning beyond a moderate threshold reduces sparsity and comprehensibility.

In Parachute (Fig. 6), VG maps start noisy and improve moderately with pruning up to 30%, after which sparsity gains are minimal. IG maps are clearer for low-to-moderate pruning (10–30%), but higher pruning again introduces noise, consistent with reduced sparsity.

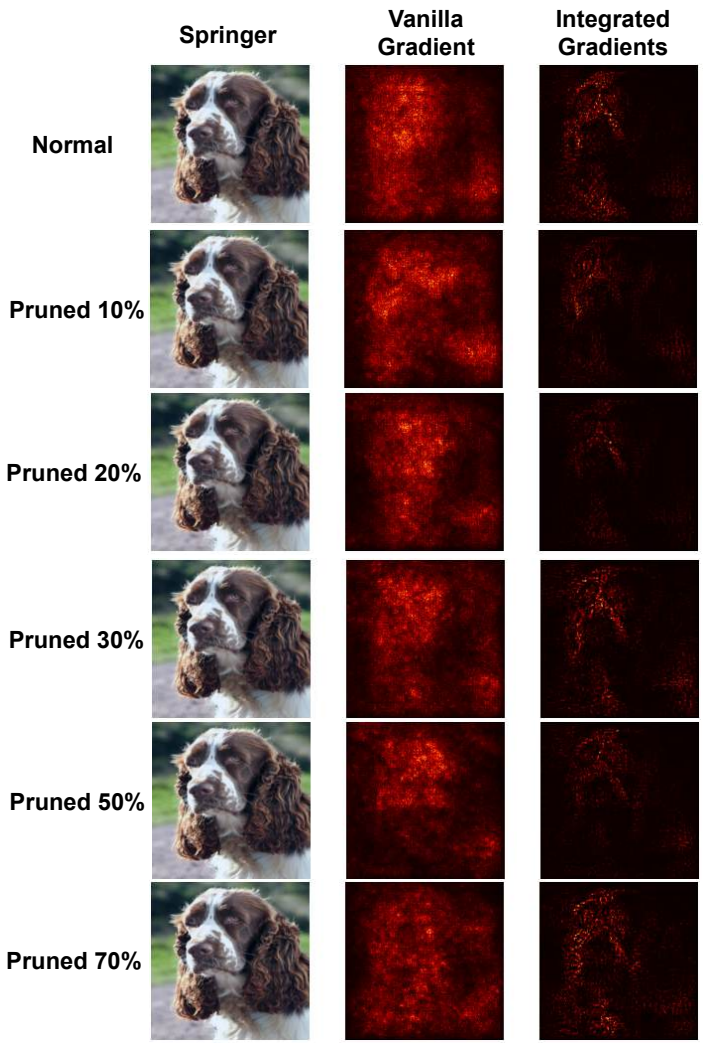


Figure 5: Springer

B Evaluation metrics

B.1 Sparsity

To measure the concentration of attribution in the saliency maps, we use the Gini Index [3]. The Gini Index is a measure of statistical dispersion that quantifies inequality. For a saliency map S, the Gini Index is calculated on the absolute values of its attribution scores. A score close to 1 indicates a highly sparse map where attribution is concentrated on a few input features (pixels), while a score close to 0 indicates a diffuse map where attribution is spread out evenly.

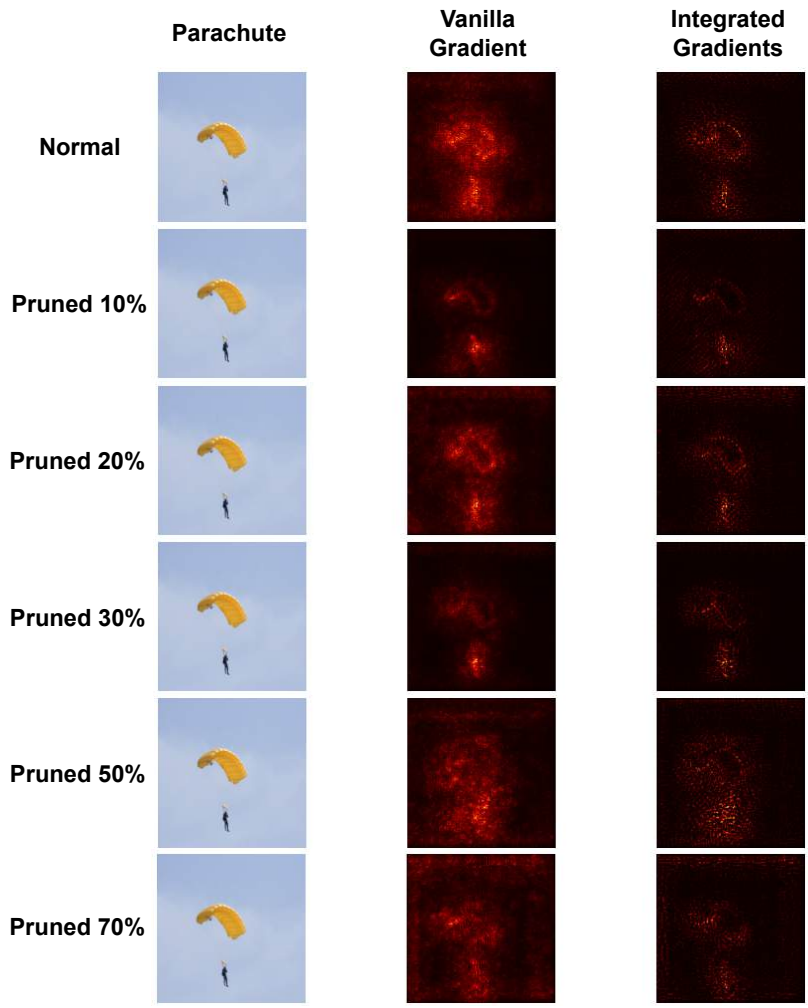


Figure 6: Parachute

$$G(\phi(\mathbf{x})) = 1 - 2\sum_{k=1}^{d} \frac{\phi(\mathbf{x})_{(k)}}{||\phi(\mathbf{x})||_1} \frac{d - k + 0.5}{d} \tag{1}$$

Here, $||\phi(\mathbf{x})||_1$ is the L_1-norm of $\phi(\mathbf{x})$, and $\phi(\mathbf{x})_{(k)}$ denotes the k-th smallest element in the sorted vector. The Gini index ranges from 0 to 1.

Sparsity helps evaluate how concentrated the attribution scores are, with higher sparsity leading to more comprehensible and human-friendly explanations.

B.1.1 Faithfulness

To evaluate how accurately a saliency map reflects the model's decision-making process, we use the Remove and Retrain (ROAD) benchmark [16]. ROAD measures the change in model performance after removing the input features identified as most important by the saliency map. Specifically, we remove the top $k\%$ of pixels (ranked by their attribution scores) and replace them with a neutral

value (e.g., average of neighbors). We then measure the model's prediction probability for the correct class on the modified image. A larger drop in probability signifies a more faithful explanation, as it indicates that the removed features were indeed critical to the model's original prediction.

A sharper accuracy drop as features are removed indicates a better explanation, as the most relevant features have a greater impact on model predictions.

We quantify the ROAD plot by computing the area over perturbation curve (AOPC) score.

C CRAFT

To probe model decisions beyond pixel-level saliency, we adopt CRAFT (Concept Recursive Activation Factorization) [5], a method for discovering and quantifying concepts directly from model activations.

The first step is to select a set of images predicted as a target class y, i.e., $C = \{x_i : f(x_i) = y\}$. This ensures that the analysis reflects the model's internal representation of a class rather than human labels. From each image, localized crops are extracted via a simple crop-and-resize operator, avoiding artifacts introduced by segmentation and inpainting as in ACE [8]. These crops form an auxiliary dataset, which is passed through the network to obtain activations.

To extract concepts, CRAFT applies **Non-negative Matrix Factorization (NMF)** to the activations, decomposing them into a set of *Concept Activation Vectors (CAVs)* W (the "concept bank") and corresponding coefficients U. This factorization expresses each activation as a non-negative linear combination of concepts, yielding semantically interpretable clusters such as canopy textures, sky patches, or suspension lines in our parachute example.

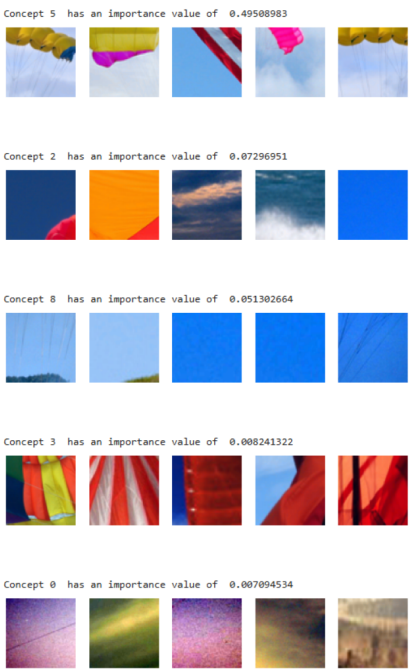
To evaluate the importance of concepts, CRAFT uses Sobol indices [18]. By perturbing concept coefficients and measuring the variance in model outputs, the method quantifies how much each concept (alone or in interaction with others) influences the prediction. High-importance concepts correspond to activation patterns whose removal or alteration substantially changes the model's output, while low-importance concepts reflect background or unused features.

We use the official implementation available in `https://github.com/deel-ai/Craft`.

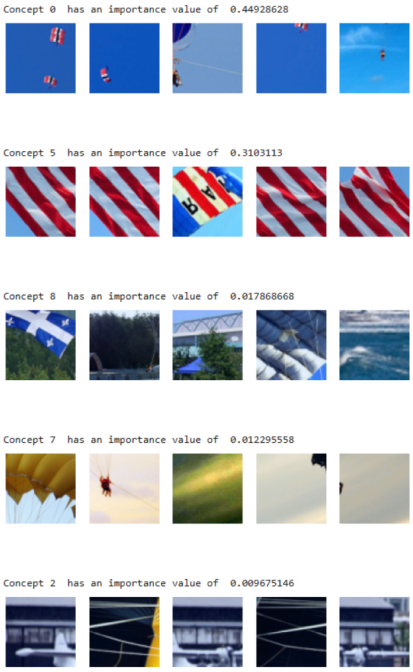
D Concept extraction

Figures 7a–9b illustrate the top CRAFT-extracted concepts for the parachute class across the baseline and progressively pruned models.

For the normal (non-pruned) model, the most influential concept (Concept 5, importance ≈ 0.495) captures brightly colored parachute canopies (yellow, magenta, red) against a clear blue sky—highly distinctive for the target class. Secondary concepts include fabric color blocks and environmental elements , sky segments with suspension lines, and multicolored close-ups with lower relevance.



(a) Normal model



(b) 10% pruned model

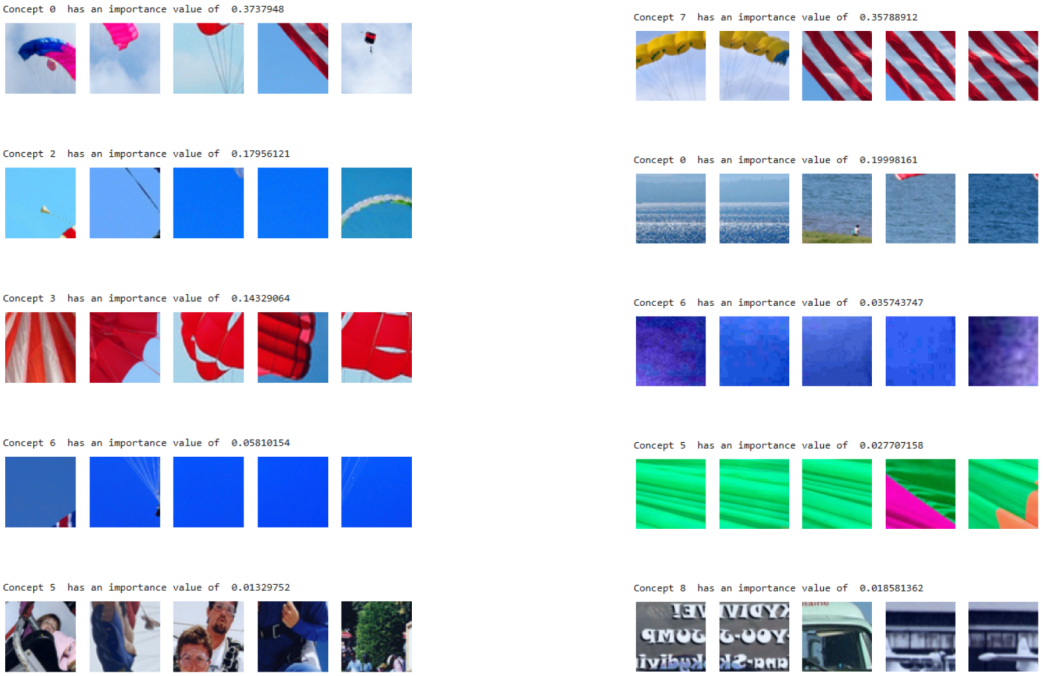
Figure 7: Concept extraction

In the 10% pruned model, the dominant concept (Concept 0, ≈ 0.449) shifts toward small parachutes in a clear sky—similar similar the normal model's top concept but more zoomed out and object-centered. The second concept (Concept 5, ≈ 0.310) emphasizes red-and-white canopy stripes. Lower-ranked concepts include mixed environmental scenes, partial canopy edges with background blur, and structural cord-like patterns.

For the 20% pruned model, the top concept (Concept 0, ≈ 0.374) focuses on close-up canopies in varied colors (blue, pink, red) with sky backgrounds, while Concept 2 (≈ 0.180) captures large blue-sky patches with minimal object detail, indicating greater reliance on background. Concept 3 (≈ 0.143) retains strong red canopy patterns. Concept 6 and Concept 7 include secondary cues like sky textures and even human figures and gear, suggesting increased drift toward non-object-specific features.

In the 30% pruned model, the leading concept (Concept 7, ≈ 0.358) mixes yellow canopy shots with red-and-white patterns, still object-relevant but less semantically pure. Concept 0 (≈ 0.200) shifts to water and shoreline views, marking heavier environmental dependence. Other concepts include abstract blue/purple textures, unrelated green fabric folds, and irrelevant structural or text-based patterns. This stage shows increased background noise infiltration.

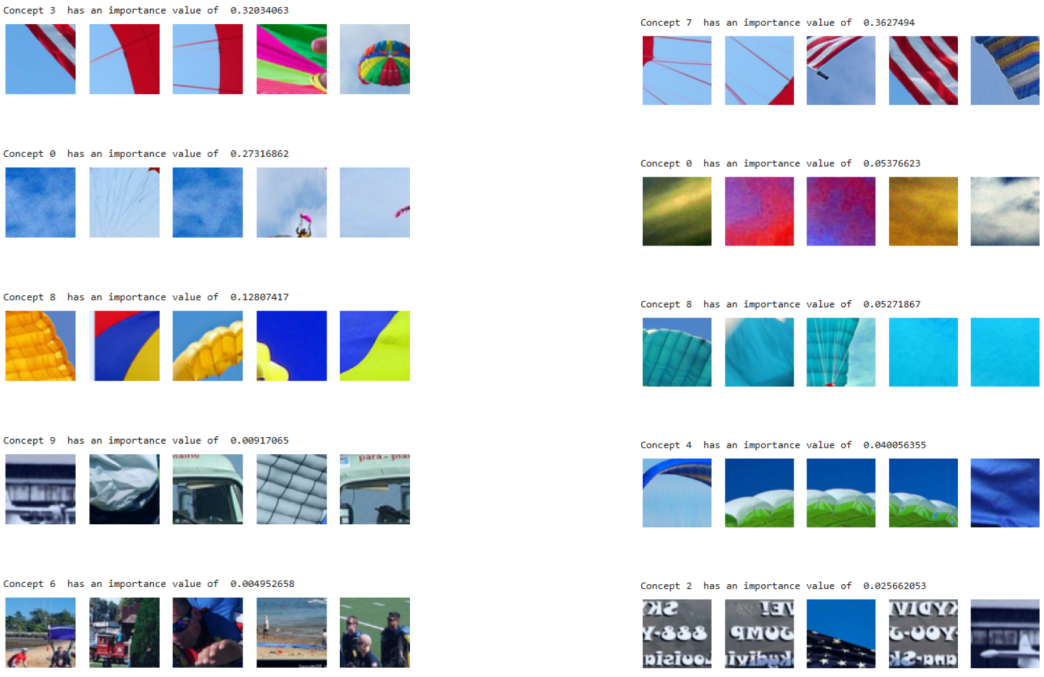
By 50% pruning, the model's top concepts retain some object cues but lose exclusivity. Concept 3 (≈ 0.320) includes red canopy sections and multicolor parachute views, while Concept 0 (≈ 0.273) largely captures sky backgrounds with partial parachutes. Concept 8 (≈ 0.128) is still object-related, but lower concepts contain unrelated scenes with vehicles, people, and beach environments, reflecting growing concept drift.



(a) 20% pruned model

(b) 30% pruned model

Figure 8: Concept extraction



(a) 50% pruned model

(b) 70% pruned model

Figure 9: Concept extraction

The 70% pruned model relies heavily on a single dominant concept (Concept 7, ≈ 0.363) that contains suspension lines, red/white stripes, and yellow-blue canopy shots. Subsequent concepts drop sharply in importance and often mix in non-object patterns: colored textures, blue canopy folds and sky areas, mixed parachute/fabric views, and unrelated text or structural imagery. At this stage, semantic coherence is low outside the top concept.

Overall, these results indicate that low-to-moderate pruning ($\leq$20%) preserves core parachute concepts while slightly reweighting toward object-centered or background-focused cues. Beyond 30%, the top concepts increasingly mix heterogeneous patterns, and unrelated or noisy concepts rise in importance. By 70% pruning, concept purity degrades substantially, suggesting that aggressive sparsification forces the model to compress diverse discriminative features into fewer, less distinct activation patterns, impairing interpretability.