Activation Matching for Explanation Generation and Circuit Discovery

Editors: List of editors' names

Abstract

In this paper we introduce an activation-matching–based approach to generate minimal, faithful explanations for the decision-making of a pretrained classifier on any given image and reveal the underlying compact internal circuits that suffice for its decisions. Given an input image x and a frozen model f, we train a lightweight Autoencoder to output a binary mask m such that the explanation $e = m \odot x$ preserves both the model's prediction and the intermediate activations of x. Our objective combines: (i) multi-layer activation matching with KL Divergence to align distributions and cross-entropy to retain the top-1 label for both the iamge and the explanation; (ii) mask priors—L1 area for minimality, a binarization penalty for crisp 0/1 masks, and total variation for compactness; and (iii) abductive constraints for faithfulness and necessity. Beyond producing per-image explanations, we also introduce a circuit readout procedure wherein using the explanation's forward pass, we identify active channels and construct a channel-level graph, scoring inter-layer edges by ingress weight magnitude times source activation and feature-to-class links by classifier weight magnitude times feature activation. This reveals sparse data-dependent sub-circuits and or internal pathways providing a practical bridge between explainability in the input space and mechanistic circuit analysis.

Keywords: Inversion, Explanability, Interpretability, Ciruits

1. Introduction

Explanations are increasingly recognized as essential for understanding and trusting the decision-making of modern machine learning models. Deep neural networks, despite their remarkable predictive performance, often arrive at their outputs through complex, high-dimensional computations that are not directly human-interpretable. These models typically learn a vast repertoire of decision rules, any of which may be activated for a given input. As a result, simply observing the final prediction provides little insight into why the decision was made or which aspects of the input were most responsible.

Minimality has therefore emerged as a favored criterion for explanations. By isolating the smallest possible set of input features that suffices for a given prediction, one obtains an explanation that is both human-readable and faithful to the model's internal computation. Minimal explanations highlight a compact subset of pixels in the case of images, or features in general, that directly support the output. Such explanations serve not only as cognitive aids for human understanding but also as a practical diagnostic tool: they can expose spurious correlations, highlight shortcut learning, and reveal when the model relies on inappropriate evidence. This is critical in safety-sensitive applications such as medical diagnostics, autonomous driving, and security, where knowing the precise basis for a decision can determine whether the system is trustworthy.

In this work, we propose an *activation-matching* approach that, given an image and a frozen pretrained classifier, learns a lightweight autoencoder to produce a binary mask selecting a minimal set of pixels whose masked input preserves the model's behavior. We further use the explanation's activations to derive a concise, channel-level view of the model's internal computation, revealing sparse, data-dependent subcircuits sufficient for the decision. Together, these components bridge input-level explanations with mechanistic insight; providing detailed understanding of the working on the machine learning model.

2. Prior Work

Inversion attempts to reconstruct inputs that elicit desired outputs or internal activations of a neural network. Unlike explanations, which are tied to a specific input and model decision, inversion focuses on synthesizing representative patterns that expose what a model has learned. Early studies on multilayer perceptrons applied gradient-based inversion to visualize decision rules, but these often yielded noisy or adversarial-like images Kindermann and Linden (1990); Jensen et al. (1999); Saad and Wunsch (2007). Evolutionary search and constrained optimization were explored as alternatives Wong (2017). Later work introduced prior-based regularization, including smoothness constraints and pretrained generative models, to improve realism and interpretability of reconstructions Mahendran and Vedaldi (2014); Yosinski et al. (2015); Mordvintsev et al. (2015); Nguyen et al. (2016, 2017). Recent advances include learning surrogate loss landscapes to stabilize inversion Liu et al. (2022), and generative methods that conditionally reconstruct inputs likely to produce a given output Suhail and Sethi (2024). Alternative formulations recast inversion into logical reasoning frameworks, encoding classifiers into CNF constraints for deterministic reconstruction Suhail (2024).

While inversion aims to characterize model behavior in aggregate, explanation generation focuses on providing faithful rationales for a specific prediction. Explainable AI has therefore emerged as a major research area Ali et al. (2023); Hsieh et al. (2024); Gilpin et al. (2018), motivated by the need to enhance trust, transparency, and accountability in high-stakes applications. Post-hoc attribution methods remain dominant: LIME produces local surrogate models Hamilton et al. (2022), Grad-CAM highlights salient image regions via gradient-weighted activations Selvaraju et al. (2019), and more recent work emphasizes concept-based explanations that map predictions to semantically meaningful parts Lee et al. (2025). The quality of explanations is itself a key open challenge, with surveys stressing the need for rigorous metrics combining fidelity, stability, and human-centered evaluation Zhou et al. (2021). Explanations are also being integrated into interactive systems, allowing users to steer, debug, or refine models through explanation-guided feedback Teso et al. (2022). Beyond heuristic methods, abductive reasoning approaches compute subset- or cardinality-minimal explanations with formal guarantees Ignatiev et al. (2018).

Mechanistic interpretability seeks to discover the *circuits* within a model—sparse subgraphs of neurons and connections that implement specific algorithms. Minimal explanations highlight the smallest sufficient evidence for a model's decisions providing mechanistic understanding of its internals. Early circuit analyses relied heavily on manual inspection, but recent work has introduced scalable discovery methods. Conmy et al. (2023) proposed ACDC, an automated framework that rediscovered known transformer circuits through activation patching. Rajaram et al. (2024) extended these ideas to vision models, extracting circuits responsible for concept recognition and showing that targeted edits can alter predictions and improve robustness. Nainani et al. (2024) investigated how circuits generalize across varied inputs, finding that networks often reuse core components while adapting connectivity—a form of representational superposition.

3. Methodology

We aim to generate minimal, faithful explanations for a frozen classifier f and use them to expose compact internal circuits. We use a lightweight autoencoder to generate a binary mask m, trained with a composite loss consisting of activation-matching, fidelity, sparsity, binarization, smoothness, and robustness terms, each weighted appropriately.

Activation matching and output fidelity. Given an input image x and a frozen classifier f, our goal is to find a binary mask m such that the masked input $e = m \odot x$ preserves the model's behavior. Both x and e are passed through f, and we enforce that their internal representations remain aligned. Specifically, we minimize a multi-layer activation distance $\mathcal{L}_{\text{act}} = \sum_{\ell} \alpha_{\ell} d(\phi_{\ell}(x), \phi_{\ell}(e))$, where ϕ_{ℓ} denotes features at layer ℓ. In addition, we encourage output fidelity using KL divergence between the softmax distributions of $f(x)$ and $f(e)$, together with cross-entropy to preserve the top-1 label.

Mask priors for minimality. To ensure explanations are compact and interpretable, we impose priors on the mask. An area loss $\mathcal{L}_{\text{area}} = \|m\|_1$ encourages sparsity, a binarization penalty $\mathcal{L}_{\text{bin}} = \|m - m^2\|_1$ drives values toward 0/1, and a total variation term $\mathcal{L}_{\text{tv}}$ reduces speckle by promoting smooth, contiguous regions.

Abductive constraint. Alonsode minimality we also enforce a robustness constraint: *random perturbations outside the explanation should not change the prediction.* Concretely, given a perturbed background r, we form $\tilde{e} = m \odot x + (1 - m) \odot r$ and apply a cross-entropy loss to ensure that $f(\tilde{e})$ preserves the same label as $f(x)$.

Circuit discovery. Beyond input-level explanations, we analyze how evidence flows through the network. Using activations from e, we select the most energetic channels at each layer as nodes and assign edge weights between successive layers by ingress weight magnitude times source activation. Connections from the penultimate feature vector to class logits are similarly scored by $|\text{fc weight}| \times$ feature activation. This yields a sparse, channel-level graph that captures the dominant subcircuits sufficient for the model's decision.

4. Results

While our approach is general, we use it to explain the decision-making of a pretrained ResNet-18 classifier on ImageNet images. We define a simple U-Net–based autoencoder that generates a binary mask. Both the original image and the explanation are passed through the frozen ResNet, and we tap the post-ReLU activations at five layers along with the final logits. These activations are matched using mean squared error, while the outputs are aligned via KL divergence and cross-entropy. To enforce minimality, we heavily weight the area loss combined with the robustness constraint to generate crisp explanations.

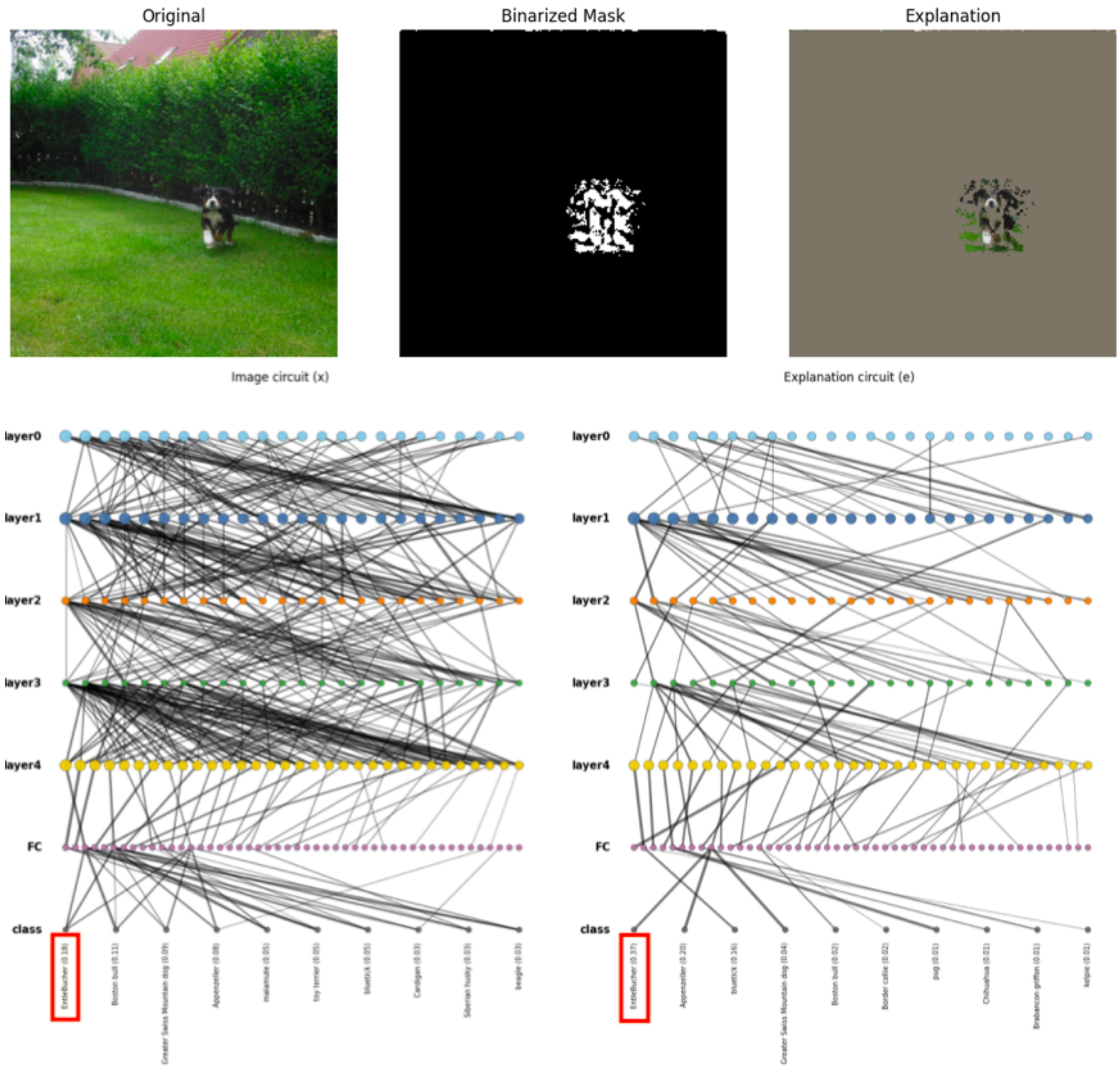

Figure 1: Top: Original Image, 0/1 Mask, and Explanation. Bottom: channel-level circuits derived from the original image and the explanation.

Figure 1 illustrates an example for the ImageNet class *EntleBucher*. The first row shows the original image, the binary mask, and the resulting explanation. The second row compares the circuit graphs obtained from the original image and from the explanation when passed through the ResNet. We observe that the explanation is highly minimal(only about 5% of active pixels), ignoring background regions of varying colors and textures, and focusing mostly on the object pixels. The explanation circuit highlights only the dominant pathways necessary for the decision. Interestingly, the top-1 confidence of the explanation is higher than that of the original image, as irrelevant background pixels have been turned off.

References

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101805. URL https://www.sciencedirect.com/science/article/pii/S1566253523001148.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2304.14997.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018. URL https://api.semanticscholar.org/CorpusID:59600034.

Nicholas Hamilton, Adam Webb, Matt Wilder, Ben Hendrickson, Matt Blanck, Erin Nelson, Wiley Roemer, and Timothy C. Havens. Enhancing visualization and explainability of computer vision models with local interpretable model-agnostic explanations (lime). In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 604–611, 2022. doi: 10.1109/SSCI51031.2022.10022096.

Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Pohsun Feng, Yizhu Wen, Xinyuan Song, Tianyang Wang, Ming Liu, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, Junhao Song, Hong-Ming Tseng, Yichao Zhang, Lawrence K. Q. Yan, Qian Niu, Silin Chen, Yunze Wang, and Chia Xin Liang. A comprehensive guide to explainable ai: From classical models to llms, 2024. URL https://arxiv.org/abs/2412.00800.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models, 2018. URL https://arxiv.org/abs/1811.10656.

C.A. Jensen, R.D. Reed, R.J. Marks, M.A. El-Sharkawi, Jae-Byung Jung, R.T. Miyamoto, G.M. Anderson, and C.J. Eggen. Inversion of feedforward neural networks: algorithms and applications. *Proceedings of the IEEE*, 87(9):1536–1549, 1999. doi: 10.1109/5.784232.

J Kindermann and A Linden. Inversion of neural networks by gradient descent. *Parallel Computing*, 14(3):277–286, 1990. ISSN 0167-8191. doi: https://doi.org/10.1016/0167-8191(90)90081-J. URL https://www.sciencedirect.com/science/article/pii/016781919090081J.

Jae Hee Lee, Georgii Mikriukov, Gesina Schwalbe, Stefan Wermter, and Diedrich Wolter. Concept-based explanations in computer vision: Where are we and where could we go? In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi, editors, *Computer Vision – ECCV 2024 Workshops*, pages 266–287, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-92648-8.

Ruoshi Liu, Chengzhi Mao, Purva Tendulkar, Hao Wang, and Carl Vondrick. Landscape learning for neural network inversion, 2022. URL https://arxiv.org/abs/2206.09027.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014. URL https://arxiv.org/abs/1412.0035.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Jatin Nainani, Sankaran Vaidyanathan, AJ Yeung, Kartik Gupta, and David Jensen. Adaptive circuit behavior and generalization in mechanistic interpretability, 2024. URL https://arxiv.org/abs/2411.16105.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016. URL https://arxiv.org/abs/1605.09304.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space, 2017. URL https://arxiv.org/abs/1612.00005.

Achyuta Rajaram, Neil Chowdhury, Antonio Torralba, Jacob Andreas, and Sarah Schwettmann. Automatic discovery of visual circuits, 2024. URL https://arxiv.org/abs/2404.14349.

Emad W. Saad and Donald C. Wunsch. Neural network explanation using inversion. *Neural Networks*, 20(1):78–93, 2007. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2006.07.005. URL https://www.sciencedirect.com/science/article/pii/S0893608006001730.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Pirzada Suhail. Network inversion of binarised neural nets. In *The Second Tiny Papers Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=zKcB0vb7qd.

Pirzada Suhail and Amit Sethi. Network inversion of convolutional neural nets. In *Muslims in ML Workshop co-located with NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=f9sUu7U1Cp.

Stefano Teso, Öznur Alkan, Wolfang Stammer, and Elizabeth Daly. Leveraging explanations in interactive machine learning: An overview, 2022. URL https://arxiv.org/abs/2207.14526.

Eric Wong. Neural network inversion beyond gradient descent. In *WOML NIPS*, 2017. URL https://api.semanticscholar.org/CorpusID:208231247.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015. URL https://arxiv.org/abs/1506.06579.

Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021. ISSN 2079-9292. doi: 10.3390/electronics10050593. URL https://www.mdpi.com/2079-9292/10/5/593.

Appendix A. More Results and Analysis

In Figure 2 how varying the relative weighting of area and smoothness terms affects the explanations. In the first case, heavily weighting the area and total variation losses yields a very compact mask that captures only a small discriminative region. In the second example, the explanation reveals shortcut learning, as the model highlights both the dog and the leash, consistent with biases in training data. In the third case, relaxing the minimality constraints results in broader coverage of the dog and partial inclusion of the background. Finally, further relaxation expands the mask to cover the entire object, suggesting that our framework could be adapted for instance-level segmentation within a class.

As shown in Figure 3, when strong minimality constraints are applied, the explanation for a single otter reduces to a remarkably small region—roughly 2% of pixels—focusing primarily on the facial features and fur texture. Despite this extreme sparsity, the classifier's label is preserved with high confidence. In contrast, when applied to an image with multiple otters, the method produces separate explanations that selectively attend to each animal, demonstrating how the approach can adapt to multi-instance settings and highlight distinct decision-supporting evidence for each occurrence.

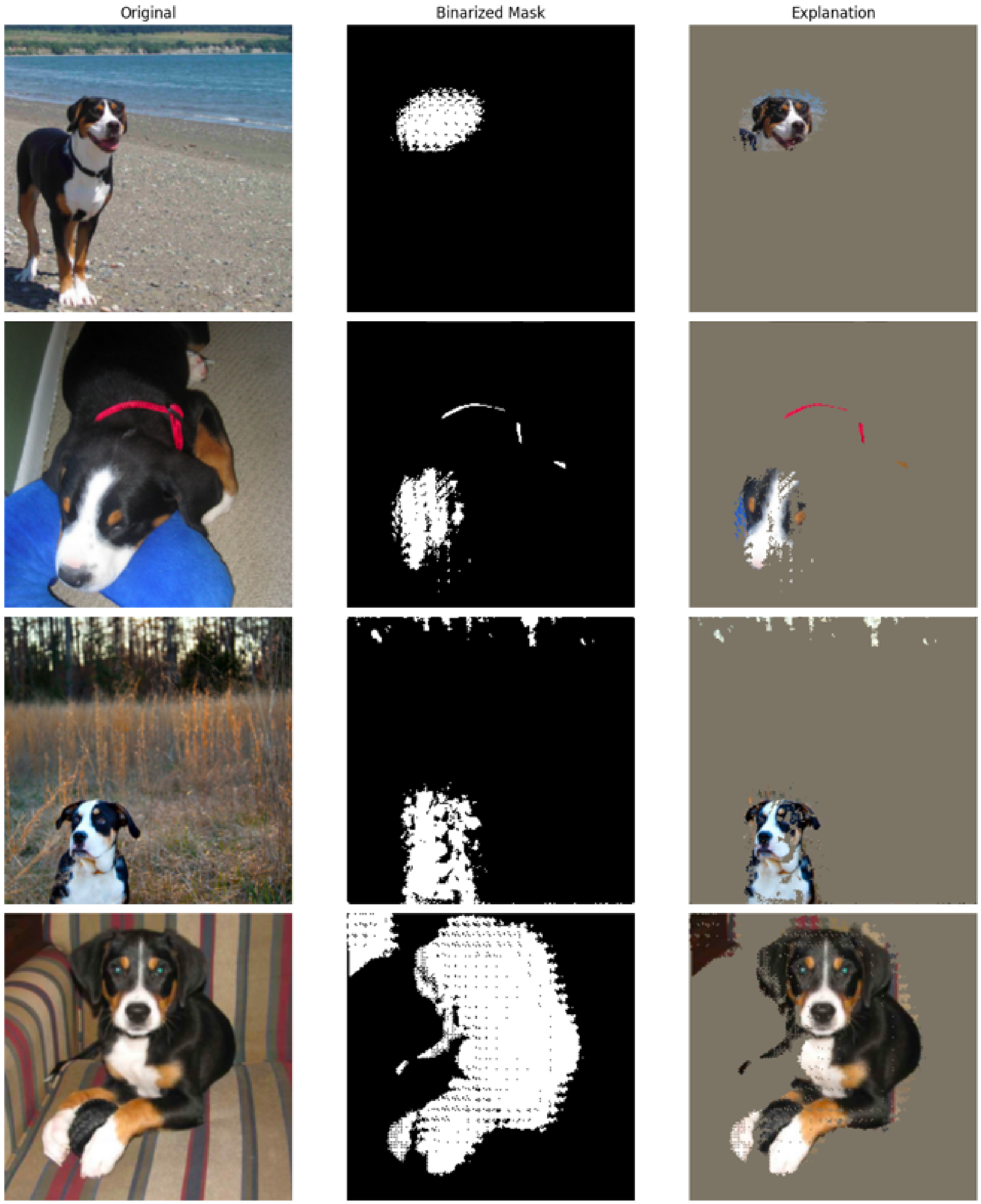

Figure 2: Effect of varying loss weights on generated explanations. Each triplet shows the original image, the generated mask, and the resulting explanation. (1) With heavily weighted area and total variation losses, the explanation becomes extremely small and localized. (2) Example of shortcut learning: the model highlights not only the dog but also the leash, reflecting dataset biases where dogs frequently appear with leashes. (3) With relaxed constraints, a larger portion of the dog and some background regions are included. (4) Further relaxation of the area loss highlights the entire dog, demonstrating how the approach can be extended toward instance-level segmentation.

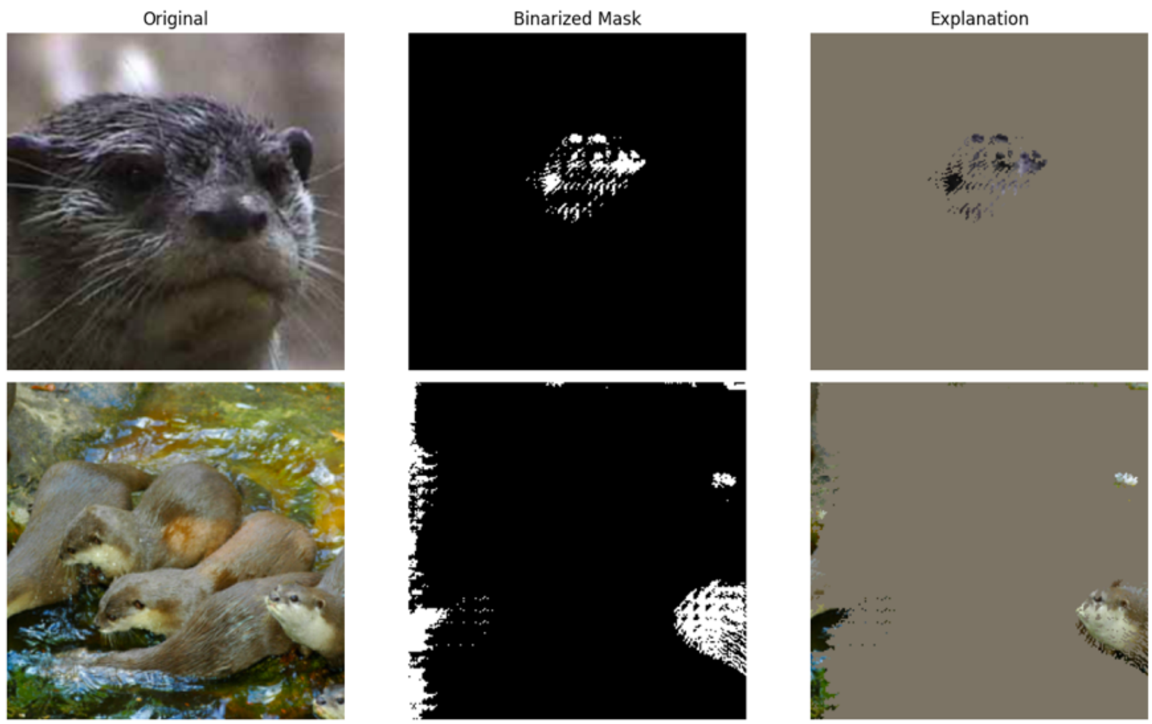

Figure 3: Explanations for sample Images of *Otter*. (Top) With heavily weighted area and total variation losses, the mask retains only about 2% of the image pixels, yet these are sufficient to classify the otter. The highlighted regions focus on the eyes, nose, and characteristic texture. (Bottom) For an image containing multiple otters, the framework produces distinct explanations, each selectively highlighting a different otter in the frame, illustrating the capacity of the method to isolate multiple instances.