Causal Interpretation of Sparse Autoencoder Features in Vision

Sangyu Han Yearim Kim Nojun Kwak
Seoul National University, Seoul, Korea
{acoexist96, yerim1656, nojunk}@snu.ac.kr

Abstract

Understanding what sparse auto-encoder (SAE) features in vision transformers truly represent is usually done by inspecting the patches where a feature's activation is highest. However, self-attention mixes information across the entire image, so an activated patch often co-occurs with—but does not cause—the feature's firing. Consequently, interpretations based solely on top-activation patches can be misleading. We therefore propose ***Causal Feature Explanation (CaFE)****, which levarages Effective Receptive Field (ERF). We consider each activation of an SAE feature to be a target and apply input-attribution methods to identify the image patches that causally drive that activation. Across CLIP-ViT features, ERF maps frequently diverge from naive activation maps, revealing hidden context dependencies (e.g., a "roaring face" feature that requires the co-occurrence of eyes and nose, rather than merely an open mouth).. Patch insertion tests confirm that our CaFE more effectively recovers or suppresses feature activations than activation-ranked patches. Our results show that CaFE yields more faithful and semantically precise explanations of vision-SAE features, highlighting the risk of misinterpretation when relying solely on activation location.*

Figure 1. Interpretation of CLIP (ViT-L/14) sae features at layer 22. While most SAE feature activations are localized aligned with its meaning (**Top**), we found some of SAE feature activations scattered across the image, halting the explanation of SAE features (**Bottom**).

1. Introduction

Interpretable machine learning seeks to map deep model representations to human-understandable concepts. In vision, sparse autoencoders (SAEs) have emerged as a compelling approach to distill concise basis features from high-dimensional transformer latents by imposing sparsity constraints [4]. These SAE features ideally capture distinct visual patterns and have been widely used to reveal semantic structure in complex models.

To assign semantic labels to these SAE features, prior works have proposed several methodologies. One line of work retrieves the top-activated samples and derives feature labels from those images [7, 11]. Another approach further pinpoints high-activation tokens within those samples for fine-grained annotation [9]. More recently, PatchSAE combines both sample- and patch-level analyses, though it still primarily relies on top-activated images for interpretation [5]. All these approaches presuppose that features are spatially localized and semantically coherent.

However, as shown in Fig. 1, we find that there exists some *non-localized* sparse autoencoder (SAE) features : their highest-activation patches scatter across the image. For these non-localized features, neither reviewing top-activated images nor marking their activated patches yields a coherent interpretation, since the visible activation peaks represent correlational, not necessarily causal, evidence.

To interpret non-localized SAE features faithfully, we propose examining each activated token's *effective receptive field* (ERF) [3]: the set of input patches that causally

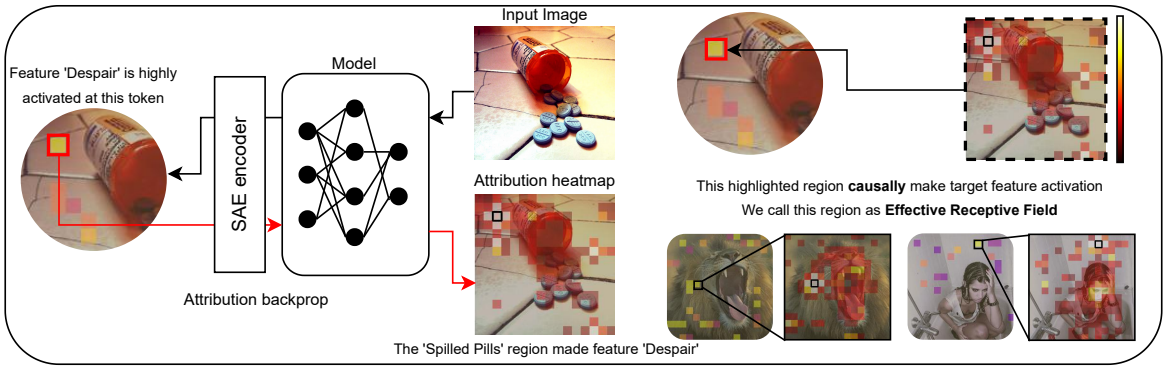

Figure 2. Method overview. The feature 'Despair' highly activated at the patch at background. Using attribution method, we can find which part of image causally contribute to the feature. We call this effective region of interest as **Effective Receptive Field.**

drive the token's activation.

In this paper, we introduce **Causal Feature Explanation (CaFE)**, which integrates: (1) SAE feature extraction on transformer latents, and (2) patch-level input attribution (e.g., Integrated Gradients or AttnLRP). For each feature, we compute attribution scores over all patches and identify those with highest causal contribution, forming the feature's ERF. By revealing the causal drivers of feature activations, CaFE enables more trustworthy and accurate interpretations of vision models, avoiding misleading correlations and deepening our understanding of complex architectures.

2. Related works

Individual units within convolutional neural networks (CNN) and transformer architectures are often *polysemantic*, entangling several concepts and thereby hampering interpretation. Sparse autoencoders (SAEs) address this limitation by learning an overcomplete, l_1-regularized basis in which each latent dimension activates for a single concept. Originally applied to language models, SAEs have seen adapted to vision transformers, where they capture object parts and textures from patch embeddings [5]. Most existing explanation methods simply visualize top-activating images or tokens [9, 11]; these works are effective only when features are spatially localized and may fail when confronted with the non-localized features we observe.

3. Method

3.1. Preliminaries: SAE Features

Given a hidden representation $\mathbf{h} \in \mathbb{R}^n$ from a backbone (e.g. patch embeddings of a ViT), we learn a SAE

$$\mathbf{z} = \mathrm{ReLU}(W_e(\mathbf{h} - \mathbf{b}_d)), \qquad \hat{\mathbf{h}} = W_d\mathbf{z} - \mathbf{b}_h,$$

with the number of features $m \gg n$, where $W_e \in \mathbb{R}^{m\times d}$ is the SAE encoder weight matrix, $W_d \in \mathbb{R}^{d\times m}$ is the decoder weight matrix, and $\mathbf{b}_d, \mathbf{b}_h$ are learned bias vectors. Training minimizes

$$\mathcal{L} = \left\|\mathbf{h} - \hat{\mathbf{h}}\right\|_2^2 + \lambda\|\mathbf{z}\|_1,$$

so that each latent dimension z_k activates only for patches exhibiting a specific visual concept. The k-th row of W_e therefore serves as an interpretable feature detector, often corresponding to an object part, texture, or scene element.

3.2. Causal Feature Explanation (CaFE)

From activations to causes. SAE features are usually inspected by ranking the patches whose activations $z_k(I) \in \mathbb{R}$ are highest, where I denote the input image. While this works well for *localized* features, Fig. 1 shows that *non-localized* features appear in disparate regions, offering only correlational hints. This limitation motivates us to ask not merely *where* z_k fires, but *which image evidence truly drives the activation*.

Effective Receptive Field (ERF). Let $A(p \mid z_k, I)$ denote the attribution of input patch p to the scalar output $z_k(I)$. The *effective receptive field* of feature k on I is the score map

$$\mathrm{ERF}_k(I) = \big\{\big(p, A(p \mid z_k, I)\big) : p \in I\big\}, \qquad (1)$$

whose intensity measures how much p caused to form z_k. High-score regions are the real evidence behind the activation, independent of where the network reports the maximum. Fig. 2 illustrates that although the token in the background floor attains the highest activation for the "*Despair*" feature, the ERF pinpoints the spilled-pill region as the causal driver.

As shown in Fig. 2, the attribution A is obtained by backpropagating relevance scores from the target SAE neuron through the SAE encoder and subsequently through the vision transformer. Inside transformer layers, we employ **Attention-LRP (AttnLRP)** [1], an adaptation of

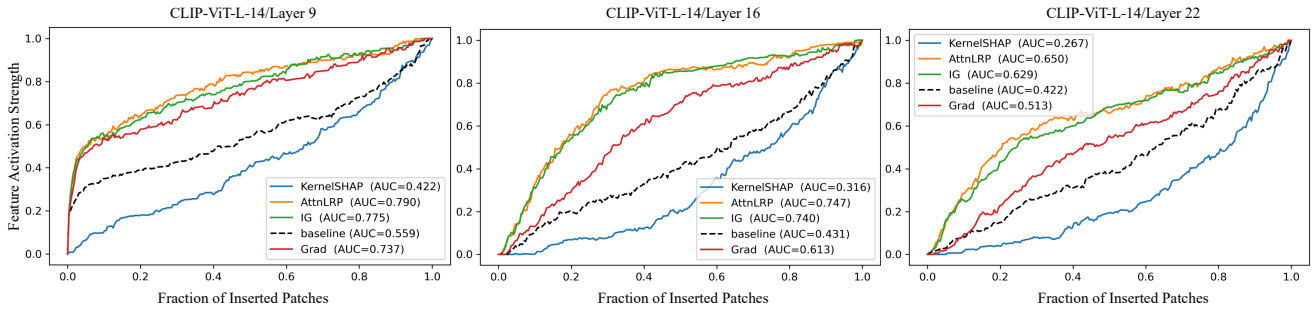

Figure 3. Causality validation. We compare causality of our CaFE method with baseline (naive activation-based patch ranking). We also compare various attribution methods including KernelSHAP, Attention-LRP, Integrated Gradients, and Gradients. It shows that our CaFE method with AttnLRP surpassed other attribution methods.

Layer-wise Relevance Propagation that distributes relevance properly considering attention edges. For comparison, we also report *Integrated Gradients* (IG) [10], KernelSHAP [6], and Gradients as baselines. These methods produce the same ERF interface, allowing plug-and-play replacements.

By replacing naive activation maps with ERF attribution, our **Causal Feature Explanation (CaFE)** pipeline delivers faithful explanations for both localized and non-localized SAE features, laying the foundation for the analyses in the following sections.

4. Experiments

We conduct a series of experiments to (i) assess the fidelity with which CaFE identifies the *causal* image evidence behind SAE activations and (ii) characterize when and where non-local SAE features arise in a vision transformer.

Experimental Setup. All experiments are conducted with the CLIP–ViT-L/14 encoder [2]. For each transformer layer, we train a Matryoshka SAE [11] on $\mathbf{5 \times 10^8}$ image patches extracted from the ImageNet-1K training set. Following the prior works, the reconstruction and sparsity hyper-parameters are held fixed across layers.

4.1. Quantitative Causality Evaluation

To quantitatively validate that our Causal Feature Explanation (CaFE) framework accurately identifies the true causal regions, we perform insertion tests, an established evaluation protocol in the explainability evaluations [1, 3, 8]. In an insertion test, we start with a blank image and gradually insert patches from the original image in order of their importance, measuring how quickly the feature activation is recovered. For completeness, we also considered deletion tests; however, across all methods, removing the patches at the method's own selected locations immediately drives the feature activation z_k to zero, so we omit deletion results. We then apply these insertion tests to compare the efficacy of attribution-based importance maps with that of the naive activation-based patch ranking. If our method with ERF truly pinpoints the true causal patches, then inserting solely those patches into a blank canvas should yield a higher z_k activation than inserting the patches selected purely on the basis of their activation magnitudes. We perform these tests over a set of images and features, computing metrics like the area under the insertion curve (AUC) as a summary of explanation efficacy. We compared several CaFE methods with different attribution with baseline approach. Fig. 3 confirms that ERF-guided insertion highly outperformed the causal recovery rate of the activation baseline across the layers. Between CaFE methods, AttnLRP surpassed other attribution methods - which is known as the most faithful attribution method for the transformer architecture. Importantly, even for traditional "local" features ERF still gives a modest boost, indicating that attribution sharpens patch selection beyond raw activation.

4.2. Qualitative Analysis of Non-local Features

ERF maps of non-local features frequently diverge from the naive activation map, uncovering hidden context dependencies. **Bottom right** of Fig. 2 shows the examples of non-local SAE features with their corresponding ERF. The "Despair" feature fires only when *spilled pills* co-occur with a *frowning face*, even though its maximal activation patch is far from the pills. Such phenomena remain imperceptible when one relies solely on activation-based inspection.

Moreover, for each layer, we manually reviewed the **first 100** SAE features and flagged those whose top-activation patches were spatially inconsistent with their corresponding ERF. The results depicted in Fig. 5 show a clear trend. Firstly, non-local features are *scarce* in early layers ($<$layer 9). All such cases are class-token features whose activations reside exclusively at the CLS position. Secondly, their frequency rises sharply in higher layers, peaking at layer 22 where $\approx$14% of features are non-local. These fea-

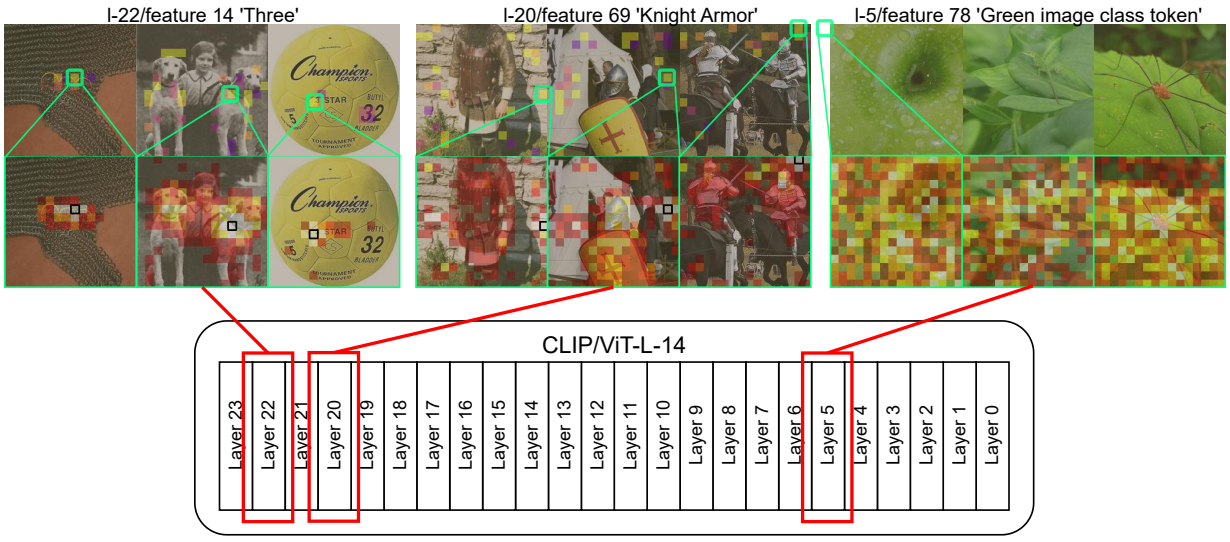

Figure 4. Qualitative examples of non-local SAE features and their ERFs at the points of highest activation across different layers. Even when the feature is spatially displaced from the region that encodes its meaning, its ERF still correctly pinpoints the area that triggered the activation. In lower layers (right-most column), the non-local SAE feature appears only when its semantic meaning is linked to class tokens.

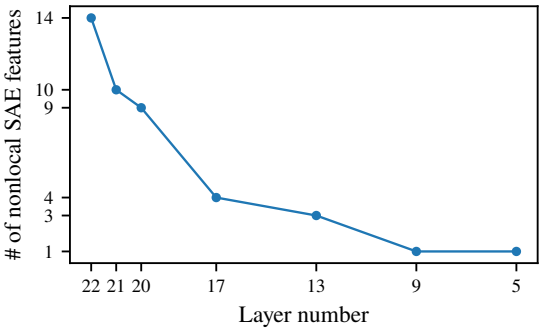

Figure 5. The number of non-local SAE features across layers. The non-local SAE features become rarer as layer number decreases. We manually inspected the number of non-local features out of the first 100 features.

tures encode highly abstract, often compositional concepts (e.g. “knight in armour”, “three”).

This distribution supports the intuition that self-attention progressively mixes global context, making later-layer activations increasingly difficult to interpret without ERF.

4.3. Discussion

Limitations and open questions. Computing the ERF for each feature requires both forward and backward passes, which may be costly for large SAEs. Furthermore, manual annotation of non-local features is inherently subjective; devising scalable, automated criteria remains future work. Finally, while we focus on vision transformers, analogous patterns of context mixing arise in large-scale language models; extending our CaFE with ERF-based attribution to text modalities is also a compelling avenue for further exploration.

5. Conclusion

We presented the Causal Feature Explanation (CaFE) framework for interpreting vision model features by leveraging effective receptive field attribution, and empirically demonstrated its superiority to the conventional practice of relying on top activations. Our study reveals that many sparse autoencoder features in vision transformers cannot be adequately understood by looking only at where they activate; one must also consider why they activate. By applying input attribution to each feature, we obtain causal explanations that often differ from correlational activation maps – shedding light on context dependencies and complex concepts encoded by the model. The main contribution of our work is to show that causal, ERF-based explanations provide a more faithful and semantically precise interpretation of visual features than activation-based methods. This is a crucial diagnostic insight for vision model interpretability: it prevents us from mislabeling features or overlooking the true factors influencing model representations.

6. Acknowledgement

This work was supported by NRF grant (2021R1A2C3006659) and IITP grants (RS-2022-II220953, RS-2021-II211343), all funded by MSIT of the Korean Government.

References

[1] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. *arXiv preprint arXiv:2402.05602*, 2024. 2, 3

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 3

[3] Sangyu Han, Yearim Kim, and Nojun Kwak. Respect the model: Fine-grained and robust explanation with sharing ratio decomposition. *arXiv preprint arXiv:2402.03348*, 2024. 1, 3

[4] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[5] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 3

[7] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models, 2025. 1

[8] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 3

[9] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models, 2025. 1, 2

[10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 3

[11] Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. 1, 2, 3