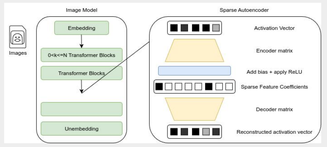
Causal Interpretation of Sparse Autoencoder Features in Vision

Sangyu Han, Yearim Kim, Nojun Kwak

Seoul National University, Seoul, Korea
{acoexist96, yerim1656, nojunk}@snu.ac.kr

TL;DR: We explain SAE features causally by tracing which input patches actually give rise to each latent.

Preliminaries: SAE

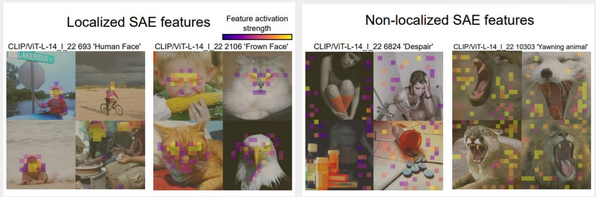


- Purpose: Extract a set of interpretable features from hidden representations (e.g., ViT patch embedding $\mathbf{h} \in \mathbb{R}^n$) within an image model.
- Encoder: Transforms $\mathbf{h}$ into a higher-dimensional sparse latent vector $\mathbf{z} = \mathrm{ReLU}(W_e(\mathbf{h} - \mathbf{b}_d))$.
- Decoder: Reconstructs the original representation $\hat{\mathbf{h}} = W_d\mathbf{z} - \mathbf{b}_h$.
- Loss Function: Minimizes $\mathcal{L} = \left\|\mathbf{h} - \hat{\mathbf{h}}\right\|_2^2 + \lambda\|\mathbf{z}\|_1$.

Motivation

Sparse autoencoders (SAEs) yield interpretable latents, and interpreting these features is done by taking the most-activated patch as "what the feature looks at".

However, we found **non-localized SAE features** - the features which we can identify the meaning of the SAE features by the image samples, but their activated region is not related to the causation.



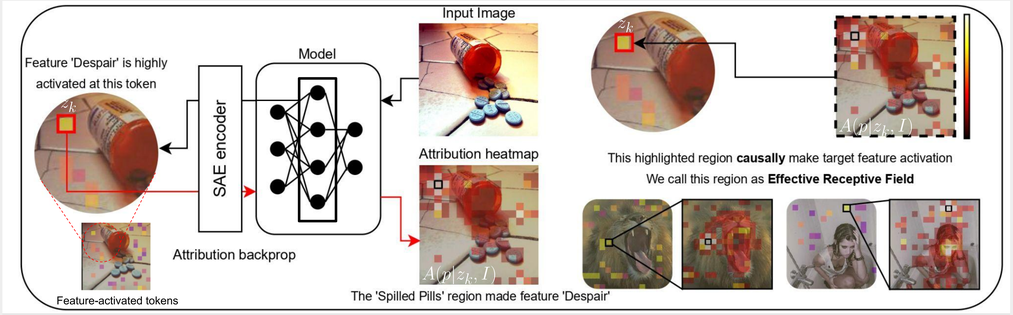
For example, as the roaring mouth region would have made latent, but the activated feature region is not related to the mouth region.

Thus, "most-activated patches" do not truly make SAE feature.

This is also obvious - "most-activated patches approach" presumably assumes that *the token which SAE feature activated is only correlated with one image patch.*
So we ask a stricter question: which input patches cause the feature z_k to arise?

- We propose Causal Feature Explanation (CaFE), a causal attribution view of SAE features using instance-specific effective receptive fields(iERF)
- We define iERF maps targeted at a latent z_k and select cause patches from them
- We evaluate causal faithfulness via insertion tests and report quantitative and qualitative gaines over activation-based views.

Method: Causal Feature Explanation (CaFE)



Given an input image I and SAE latent $z_k \in \mathbb{R}^n$, we compute attribution to z_k (e.g. AttnLRP, Integrated Gradients, …) through the model + SAE, yielding an attribution heatmap $A(p|z_k, I)$ over input patches p. We define this attribution heatmap as instance-specific Effective Receptive Field (iERF).
We then take iERF as the cause of feature.

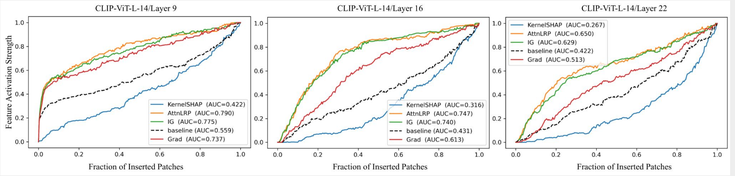
- z_k is a n-th feature id of SAE at position k. we omitted feature id n for brevity.

Quantitative Evidence

CaFE can find faithful causal evidence

To quantitatively validate that our CaFE accurately identifies the true causal regions, we perform insertion/deletion tests.
As naive feature-activated patch visualization and our CaFE ranks patches with their importance, we measure how that 'ranking' is accurate to reconstruct the feature at interest.
In insertion test, we start with a blank image and gradually insert patches from the original image in order of their importance ranking, measuring how quickly the feature activation is recovered. For deletion test, we delete gradually from the original image. however, we omitted the result; as removing the patches at selected location immediately zero out the feature, that all method have same deletion score.

The result show that CaFE outperforms the naive activation map in reconstructing feature activation. Among CaFE method, AttnLRP iERF is the most faithful compared with out methods.

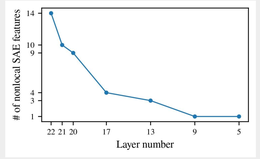


Manual analysis

Nonlocal features are more frequent in higher layers.

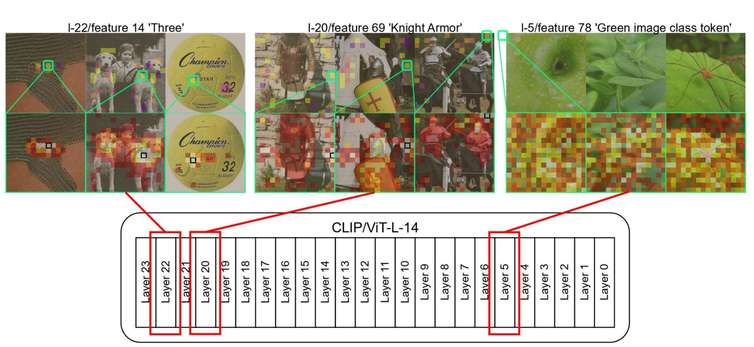
We manually reviewed the first 100 random SAE features and counted how many nonlocal features were across layers.

- Higher layer has more non-local features.
- Lower layer only had non-local features in class-token related features.



Qualitative Results

ERF maps of non-local features uncover hidden context dependencies.
Even when the feature is spatially displaced from the region that encodes its meaning, its iERF still correctly pinpoints the area that triggered the activation.
We collected top 3 nonlocal-feature-activating samples and compared them with naive feature activation maps.
While feature activation map failed to locate the cause of feature, our CaFE approach can reveal the region that caused feature to activate.



CaFE can interpret class token feature too!
As these class token feature only activate at class token, the "most-activated patches approach" cannot locate the cause of the feature thus has difficulty to interpret with top-activating samples.
However, CaFE approach can successfully locate the cause of the class token features, enabling interpretation of "summary token"