Choosing the right basis for interpretability: Psychophysical comparison between neuron-based and dictionary-based representations

Anonymous authors
Paper under double-blind review

Abstract

Interpretability research often adopts a neuron-centric lens, treating individual neurons as the fundamental units of explanation. However, neuron-level explanations can be undermined by superposition, where single units respond to mixtures of unrelated patterns. Dictionary learning methods, such as sparse autoencoders and non-negative matrix factorization, offer a promising alternative by learning a new basis over layer activations. Despite this promise, direct human evaluations comparing neuron-based and dictionary-based representations remain limited.

We conducted three large-scale online psychophysics experiments (N=481) comparing explanations derived from neuron-based and dictionary-based representations in two convolutional neural networks (ResNet50, VGG16). We operationalize interpretability via visual coherence: a basis is more interpretable if humans can reliably recognize a common visual pattern in its maximally activating images and generalize that pattern to new images. Across experiments, dictionary-based representations were consistently more interpretable than neuron-based representations, with the advantage increasing in deeper layers.

Critically, because models differ in how neuron-aligned their representations are—with ResNet50 exhibiting greater superposition, neuron-based evaluations can mask cross-model differences, such that ResNet50's higher interpretability emerges only under dictionary-based comparisons.

These results provide psychophysical evidence that dictionary-based representations offer a stronger foundation for interpretability and caution against model comparisons based solely on neuron-level analyses.

1 Introduction

A central goal of explainable AI (XAI) in computer vision is to identify the visual elements that drive the decisions of deep neural networks (DNNs) (Selvaraju et al., 2017; Bau et al., 2017; Kim et al., 2018; Ghorbani et al., 2019; Cammarata et al., 2020b). Doing so requires recovering the visual patterns that systematically influence internal activations and ultimately shape model outputs.

This goal is shared with the study of biological vision, where decades of work have sought the "preferred stimulus" of individual neurons in the visual cortex (Hubel & Wiesel, 1959; Quiroga et al., 2005). Early neuroscience-inspired XAI similarly focused on neuron-level analyses (Zhou et al., 2016; Bau et al., 2017), alongside methods that synthesize maximally activating images for individual units (e.g., Erhan et al., 2009; Zeiler & Fergus, 2014; Olah et al., 2017).

At the same time, neuroscience has increasingly shifted from single-neuron selectivity to population codes (see Ebitz & Hayden (2021)), reflecting the view that neural representations are sparse and distributed rather than purely local (Haxby et al., 2001; Quiroga et al., 2008). A similar shift is emerging in XAI because neuron axes can be a poor explanatory basis under the "superposition" hypothesis (Elhage et al., 2022; Fel

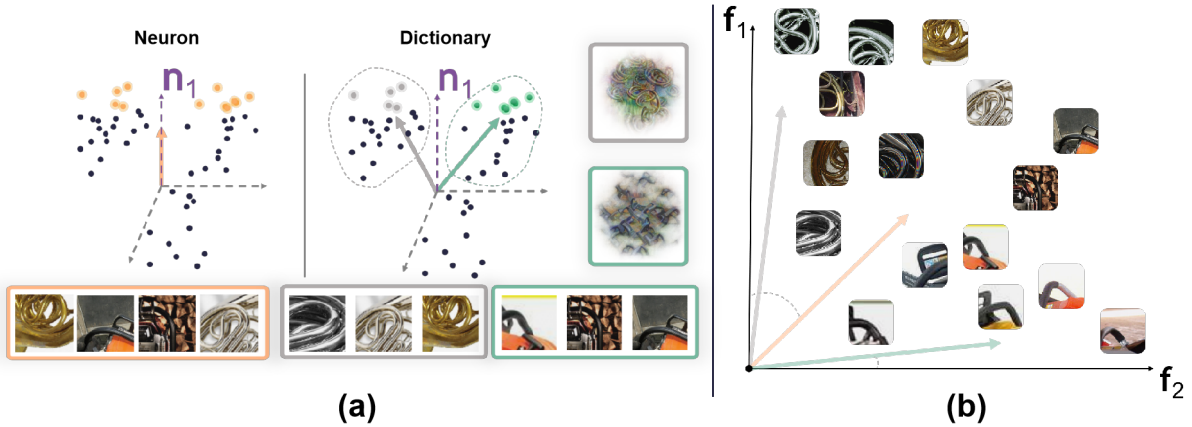


Figure 1: **(a)** • **Neuron (axis) basis** versus • **Dictionary basis**. Individual neurons may respond to multiple unrelated visual patterns (bottom), whereas dictionary learning aims to recover feature directions that isolate simpler patterns. **(b)** Dictionary learning yields a new basis over activations whose elements ideally correspond to single patterns. The interpretability hope is that these patterns align with the set of human-recognizable concepts $S = \{f_1, f_2, ..., f_n\}$.

et al., 2023a): when a model encodes more patterns than it has units, individual neurons may respond to mixtures of unrelated patterns—often referred to as polysemantic neurons.

To address this challenge, recent work applies dictionary learning methods (Fel et al., 2023b; Bricken et al., 2023; Gao et al., 2024; Costa et al., 2025; Fel et al., 2025; Zaigrajew et al.; Bussmann et al., 2025) to learn a dictionary basis over layer activations, ideally yielding directions that correspond to single, human-recognizable patterns. From an interpretability perspective, a learned dictionary basis is appealing because it aims to replace those polysemantic neuron axes with more monosemantic directions (Fig. 1).

Despite growing interest in dictionary-based representations, direct human evidence that they provide a better foundation for interpretability than neuron-based representations remains limited. This paper aims to fill this gap. Furthermore, model comparisons remain underexplored and have largely relied on neuron-based evaluations (Zimmermann et al., 2023; 2024); if models differ in their degree of superposition, conclusions drawn from neuron-based analyses may be unreliable.

We operationalize interpretability via *visual coherence*: a basis is more interpretable to the extent that humans can reliably identify the common visual pattern in its maximally activating images and generalize that pattern to new images[1]. We additionally require that the directions evaluated within a basis be demonstrably involved in the model's decisions.

We conducted three large-scale psychophysics experiments (N=481 participants, 16,835 responses) comparing the interpretability of neuron-based and dictionary-based representations in ResNet50 and VGG16. Our main contributions are:

- We identify a semantic confound in widely used human-evaluation protocols (Borowski et al., 2021; Zimmermann et al., 2023) and introduce a control that mitigates it.
- Across experiments, we find that dictionary-based representations are consistently more interpretable than neuron-based representations, with the advantage increasing in deeper layers.
- By quantifying axis-alignment[2], we find that ResNet50 exhibits greater superposition than VGG16, obscuring cross-model differences in neuron-based analyses that become apparent only in dictionary-based comparisons, revealing ResNet50 to be the more interpretable model.

[1]In the remainder of this paper, we use *interpretability* to refer specifically to this notion of human-recognizable visual coherence.

[2]Through the paper, axis-alignment refers to the alignment of the representation with neuron axes.

- We empirically demonstrate that the choice of representational basis materially affects model comparisons, cautioning against drawing conclusions from neuron-level analyses alone.

2 Related work

From neuron-based to dictionary-based representations. Early work in explainable AI (XAI) for computer vision developed attribution methods (e.g., Zeiler & Fergus, 2014; Sundararajan et al., 2017; Selvaraju et al., 2017) to explain individual predictions. These approaches primarily address the *where* question, that is, identifying which pixels or regions are most influential for a given output. However, they often fall short on the *what* question (Kim et al., 2018; Colin et al., 2022): the visual patterns or factors the model relies on to make its decision. This limitation motivated the development of feature visualization methods (Nguyen et al., 2016; Olah et al., 2017), which aim to characterize what neurons or layers represent. A consistent finding is that single neurons can respond to multiple, visually distinct patterns (Nguyen et al., 2016; Cammarata et al., 2020a; Bricken et al., 2023), echoing similar observations in natural language processing (Elhage et al., 2022).

These polysemantic responses suggest that neuron axes may not align with the underlying factors of variation in a model's representations. Under the superposition hypothesis (Elhage et al., 2022; Fel et al., 2023a), models can encode more patterns than they have neurons, causing individual units to act as mixtures of multiple patterns. In that regime, interpreting single neurons may be no more principled than interpreting arbitrary directions in the activation space, motivating the search for alternative bases that better isolate interpretable elements. This perspective has spurred increased interest in learning dictionary bases over latent activations via dictionary learning and related concept-extraction methods (Ghorbani et al., 2019; Fel et al., 2023b). In parallel, sparse autoencoders (SAEs) have emerged as a promising approach for discovering more monosemantic directions in deep networks (Bricken et al., 2023; Cunningham et al., 2023; Gao et al., 2024; Costa et al., 2025; Fel et al., 2025; Zaigrajew et al.; Bussmann et al., 2025). In this paper, we compare the suitability of the neuron basis versus a learned dictionary basis as a foundation for interpretability in computer vision models.

Human-evaluation of interpretability. Since a primary goal of XAI is to make model behavior understandable to people, interpretability ultimately requires human-centered evaluation. Psychophysics-style experiments provide a direct benchmark for assessing whether an explanation is comprehensible and usable to human observers, complementing automated or proxy metrics.

Borowski et al. (2021) were the first to quantify the interpretability of deep neural network representations using psychophysics experiments. Their protocol visualized unit selectivity by contrasting maximally and minimally activating natural stimuli. In particular, they evaluated neuron-based representations (single-unit activations) in an Inception V1 (Szegedy et al., 2015) trained on ImageNet (Deng et al., 2009). In each trial, participants viewed sets of maximally and minimally activating images for a given neuron and then selected which of two query images also strongly activated that neuron (see Fig. 2). They concluded that, from a human-centered perspective, natural exemplars are more effective than synthetically generated feature visualizations (Olah et al., 2017).

Zimmermann et al. (2021) proposed a variation of this task to investigate if humans gained causal insights from those visualizations. Participants predicted the effect of an intervention (*e.g.*, occluding an image region) on unit activation. They found that synthetic feature visualizations could improve performance, but provided limited advantage over natural exemplar images.

Interestingly, in a similar vein but in NLP, Bricken et al. (2023) compared the interpretability of 162 elements drawn from both neuron-based and dictionary-based representations. In their study, a single author rated each element using examples sampled across its activation range. They reported that dictionary-based elements were substantially more interpretable than individual neurons. Our work differs in domain (vision rather than language), experimental design, and scale (a single rater versus 16,835 behavioral responses from 481 participants).

Comparative analyses of model interpretability. Zimmermann et al. (2023) were the first to compare the interpretability of models using human evaluations. They extended the work of Borowski et al. (2021) to a broader range of computer vision architectures, including ResNet50 (He et al., 2016), and concluded that increasing model scale does not enhance interpretability. While we both build on the experimental protocol introduced by Borowski et al. (2021), their focus is on scaling these original insights across architectures. In contrast, we adapt the same protocol to address a different research question: which type of representation (neuron-based or dictionary-based) is more interpretable to humans. To the best of our knowledge, this is the first study that compares models using dictionary-based representations.

To further scale the evaluation of model interpretability, Zimmermann et al. (2024) automated the human evaluation protocol introduced by Borowski et al. (2021). In each trial, human judgments were approximated by computing pairwise perceptual similarities between queries and explanations, which were then used by a binary classifier to predict the correct query. The resulting metric was found to correlate strongly with previous results (Zimmermann et al., 2023). This work shows that although most models contain a substantial number of interpretable units, differences across models primarily arise from the prevalence of highly uninterpretable units. Such units are often attributed to superposition, whose impact may vary across models. Consequently, comparing models solely based on neuron-based representations may disproportionately penalize certain architectures, potentially yielding misleading conclusions. Our work compares models using both neuron-based and dictionary-based representations and demonstrates that the choice of representational basis can substantially impact comparative outcomes.

3 Methodology

In this section, we first provide an overview of the technical methods used in our experiments (named Experiments I, II and III), followed by a description of the psychophysical experiments conducted to compare the interpretability of neuron-based versus dictionary-based representations.

3.1 Technical methods

Models Experiments I and II described below used a ResNet50 (He et al., 2016) and Experiment III used a VGG16 (Simonyan & Zisserman, 2014), both sourced from the Torchvision (Marcel & Rodriguez, 2010) library and pre-trained on ImageNet-1k (Deng et al., 2009). We focus on convolutional neural networks (CNNs) because they remain widely used in practical computer vision applications, making our findings broadly relevant, and they produce the positive activations required by our dictionary learning method.

Dictionary-based representations. To compute dictionary-based representations for the psychophysics experiments (see Section 3.2), we employed CRAFT (Fel et al., 2023b), a dictionary learning method based on Non-negative Matrix Factorization (NMF). Specifically, given a model $f : \mathcal{X} \to \mathcal{A}$ that maps from an input space $\mathcal{X} \subseteq \mathbb{R}^d$ to an activation space $\mathcal{A} \subseteq \mathbb{R}^p$ (*i.e.*, any layer of the network), we compute the activations $\mathbf{A} = f(\mathbf{X}) \in \mathbb{R}^{n \times p}$, where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$ represents a set of n input data points. Each row $\boldsymbol{a}_i \geq \mathbf{0}$ of $\mathbf{A}$ contains the non-negative activations for a given data point $\mathbf{x}_i$, due to the use of ReLU activations. NMF approximates $\mathbf{A}$ as:

$$(\mathbf{Z}^\star, \mathbf{D}^\star) = \underset{\mathbf{Z} \geq \mathbf{0},\ \mathbf{D} \geq \mathbf{0}}{\arg\min} \left\| \mathbf{A} - \mathbf{Z}\mathbf{D}^\top \right\|_F ,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Here, $\mathbf{Z} \in \mathbb{R}^{n \times k}$ are the **codes**, and $\mathbf{D} \in \mathbb{R}^{p \times k}$ forms the **dictionary** with k elements. Both $\mathbf{Z}$ and $\mathbf{D}$ are constrained to have non-negative entries and tend to be sparse due to the properties of NMF. The dictionary matrix $\mathbf{D}$ provides a new set of basis vectors aligned with the activation patterns of the neural network, while $\mathbf{Z}$ contains the coefficients representing the original activations $\mathbf{A}$ in terms of these basis vectors.

We use NMF rather than sparse autoencoders (SAEs) because NMF respects the geometry of post-ReLU activations: both the dictionary and codes are constrained to be non-negative, whereas SAEs can learn negative dictionary atoms corresponding to directions that cannot exist in the activation space (Fel et al., 2023a). NMF is also more stable, avoiding the hyperparameter sensitivity inherent in SAE training (Fel et al., 2025).

Element importance. Let $\mathbf{v}$ be a vector from an intermediate representation, either a dictionary element in $\mathbf{D}$ or a neuron axis in $\mathcal{A}$. Let $\mathbf{x}$ be an input that strongly activates $\mathbf{v}$, and let $f(\mathbf{x})$ denote the corresponding logit score. To assess the importance of $\mathbf{v}$, we measure how sensitive $f(\mathbf{x})$ is to perturbations of $\mathbf{v}$ using Gradient×Input (Shrikumar et al., 2017), which has been shown to provide a faithful importance measure in latent spaces (Fel et al., 2023a):

$$\mathrm{GI}(\mathbf{x}, \mathbf{v}) \;=\; \mathbf{v} \odot \frac{\partial f(\mathbf{x})}{\partial \mathbf{v}}\,. \tag{1}$$

The more important $\mathbf{v}$ is for the model's decision at $\mathbf{x}$, the larger $\mathrm{GI}(\mathbf{x}, \mathbf{v})$.

Axis-alignment of dictionary elements. If a model contains polysemantic neurons, then a learned dictionary is expected to recover directions that are not well-approximated by any single neuron axis. To quantify how closely each dictionary element aligns with the neuron (coordinate) basis—or equivalently, how distributed it is across neurons—we use the sparsity measure of Hoyer (2004).

Let $D \in \mathbb{R}^{p\times k}$ denote the learned dictionary for a layer with p neurons, and let $\mathbf{d}_\ell = D_{:\ell} \in \mathbb{R}^p$ be the ℓ-th dictionary element expressed in the neuron basis. We define the average axis-alignment score as

$$\mathcal{H}(D) \;=\; \frac{1}{k}\sum_{\ell=1}^{k} \frac{\sqrt{p} - \|\mathbf{d}_\ell\|_1/\|\mathbf{d}_\ell\|_2}{\sqrt{p}-1}\,. \tag{2}$$

Here $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1 and ℓ_2 norms. For nonnegative vectors, this index lies in $[0, 1]$: it is 0 for maximally dense vectors (mass evenly spread across coordinates) and approaches 1 for one-hot vectors (mass concentrated on a single neuron axis). Thus, higher values indicate that dictionary elements are closer to individual neuron axes, *i.e.,* more coordinate-sparse or axis-aligned, whereas lower values indicate elements that combine many neuron axes.

3.2 Psychophysics experiments

3.2.1 Experimental protocol

Interpretability is ultimately a human-centered concept. To examine how the choice of representational basis shapes human judgments, we conducted three large-scale online psychophysics experiments comparing units expressed either in the neuron basis or in a learned dictionary basis. Throughout, we use *visual pattern* (or *visual feature* in the perceptual sense) to refer to recurring, human-recognizable regularities in images, and we reserve *dictionary element* for a learned direction (dictionary element) in the activation space. Across all experiments, we evaluate how easily observers can identify the *consistent visual pattern* associated with a unit (neuron axis or dictionary element) from its maximally activating stimuli.

Because standard interpretability evaluations do not easily scale to large numbers of units (Colin et al., 2022), we adapted the psychophysics protocol of Borowski et al. (2021) to measure a unit's *visual coherence*, or equivalently, its lack of ambiguity, as a proxy for interpretability.

Each participant was assigned to one of two between-subject conditions: the neuron-basis condition or the dictionary-basis condition. After completing a practice session of 9 trials, participants performed 40 trials of the same task, with each trial corresponding to a different unit (a specific neuron axis or a dictionary element, depending on the condition).

In each trial, participants were shown two panels of 9 reference images, one on each side of the screen, separated by two *query* images in the center (Fig. 2). The right panel displayed images selected to strongly activate the target unit (maximally activating stimuli). The left panel displayed images intended to contrast with the right panel (the exact selection procedure depends on the experiment; see Section 3.2.3). Participants were asked to select the query image that they believed matched the reference images on the right, *i.e.*, the query image that shared the same consistent visual pattern as the maximally activating reference set.

Intuitively, when the maximally activating images are visually coherent, participants should more reliably select the correct query. Visual coherence suggests the unit's activation is driven by a consistent, recognizable

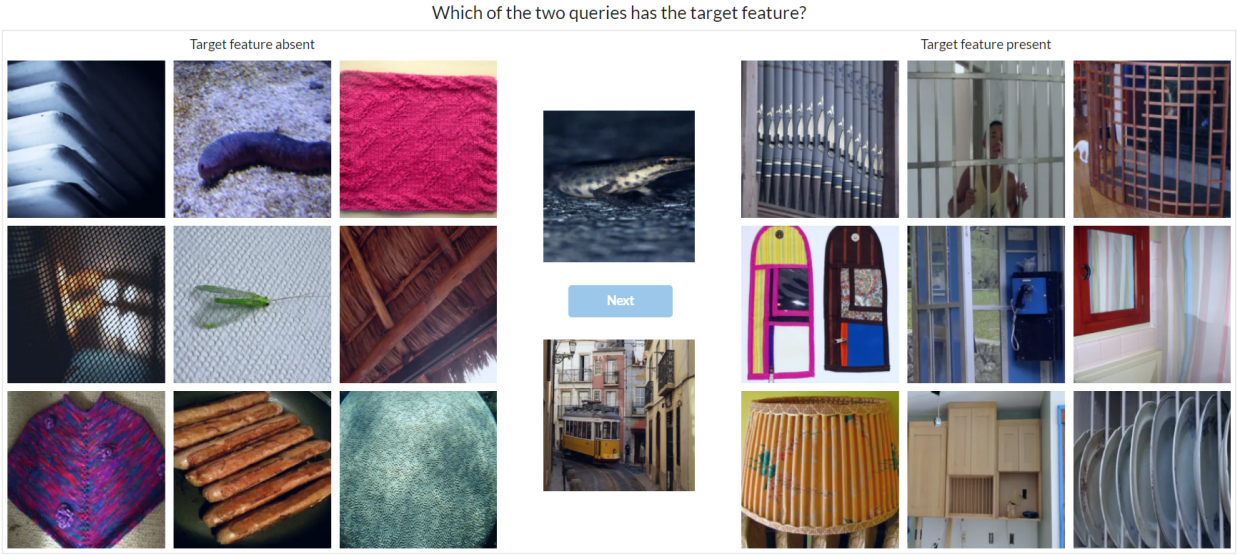


Figure 2: **Illustration of a trial.** Example of a trial in our study corresponding to Experiment I, dictionary-based condition for a unit located in *layer*2.0.*bn*2. Two panels of 9 reference images are located on the left and right-hand side of the display, separated by 2 query images in the center. Participants were asked to select the query image they believed shared the same visual pattern as the reference images displayed on the right panel, corresponding to maximally activating stimuli. The less ambiguous this shared pattern is, the more visually coherent the set of images and the more likely participants are to select the correct query. In this case, the correct query is the bottom image depicting a yellow tram.

pattern. Conversely, if the maximally activating images are heterogeneous (consistent with polysemanticity), the task becomes more ambiguous and accuracy should drop. We therefore summarize performance as the proportion of correct responses per unit, which serves as our primary measure of visual coherence and hence interpretability.

3.2.2 Unit and stimuli selection

Unit selection. Following Zimmermann et al. (2023), we aimed to obtain a representative sample of units across different layers of the neural networks of interest. As a starting point, we selected the 80 units reported in Zimmermann et al. (2023) for a ResNet50 (He et al., 2016) model. Because CRAFT requires non-negative activations (Fel et al., 2023b), we mapped neurons from convolutional layers to their counterparts in the subsequent batch normalization layer (*e.g.*, *layer*1.1.*conv*1 neuron 53 -> *layer*1.1.*bn*1 neuron 53), which are followed by ReLU and thus yield non-negative post-ReLU activations. For VGG16 (Simonyan & Zisserman, 2014), we randomly selected 80 neurons across the ReLU layers. In total, we evaluated 80 units per model, distributed across 43 layers for ResNet50 and 11 layers for VGG16.

Stimuli selection for the neuron-based condition. For each of the selected neurons, we identified a representative set of 2,900 images from the validation set of ImageNet ILSVRC 2012 (Russakovsky et al., 2015): the 2,500 most strongly activating images and the 400 least strongly activating images. Following Borowski et al. (2021), we illustrated each neuron's selectivity through both maximally activating stimuli (images likely to contain the relevant visual pattern; see Fig.2, right panel) and minimally activating stimuli (images unlikely to contain it; see Fig.2, left panel). For the maximally activating panel, we selected a random sample of 9 images from the top 150; for the minimally activating panel, we uniformly sampled 9 images from the bottom 20. We created 10 different trials per neuron following this procedure to ensure image independence across results.

Stimuli selection for the dictionary-based condition. Our hypothesis is that applying dictionary learning to neuron-based representations allows us to recover directions corresponding to patterns in superposition, which are expected to be more interpretable. To test this, we derived a complementary dictionary-based representation for each neuron in the neuron-based condition. Specifically, for each neuron in the neuron-based condition, we selected its top N=300 maximally activating images, *i.e.*, those that most strongly activated it. We then applied CRAFT to learn a dictionary of k=10 elements over the activations of that neuron's layer (we select N and k following the recommendations in Fel et al. (2023b)). From these 10 dictionary elements, we selected the one that was most frequently maximally activated across the 300 images (see Appendix 5 for more details). Finally, we ranked the 2,900 images according to their activation along the chosen dictionary element to obtain the set of stimuli that illustrate the corresponding visual pattern.

3.2.3 Experiments

In this section, we describe each of the three psychophysics experiments conducted in this work. Each experiment consists of 1,600 trials (80 units $\times$ 2 conditions $\times$ 10 trials per unit).

Experiment I. This experiment is an adaptation of the methodology proposed by Borowski et al. (2021), conducted on a ResNet-50, with two conditions: neuron-based versus dictionary-based representations. An illustration of a trial from Experiment I can be found in Fig. 2, and the experimental protocol is described in Section 3.2.1.

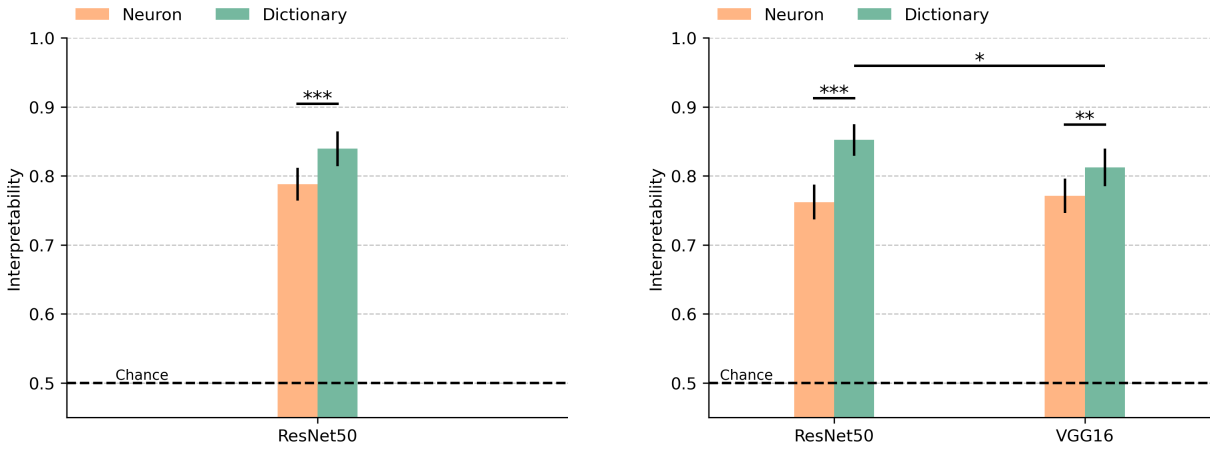
Experiment II. The main objective of Experiment II was to control for a potential semantic confound uncovered in Experiment I. If the reference images that contain or do not contain the pattern of interest belong to distinct semantic categories, then it may be possible to solve the task through simple semantic grouping. Fig. A1 illustrates this phenomenon: in the depicted example, it is easier to solve the trial by inferring that the pattern of interest is *not* about a monkey than to identify the actual visual pattern present in the reference images. In such cases, correctly solving the trial tells us little about the interpretability of the unit. To mitigate this semantic confound, we proceeded as follows. Given a set of reference images containing the pattern of interest, we extracted their semantic labels from ImageNet and searched the 400 minimally activating images for a set of 9 images matching these semantic labels. When ImageNet labels were insufficient, we expanded the search by moving upward through the WordNet (Fellbaum, 2010) hierarchy. After up to 4 iterations, all but one unit were successfully controlled. This unit was excluded from both the neuron-based and dictionary-based conditions.

Experiment III. To compare the interpretability across models, Experiment III applied the protocol from Experiment II to a VGG16.

3.2.4 Participants

A total of 481 participants were recruited for Experiments I, II, and III through the Prolific [3] online platform. All participants were native English speakers who reported no visual impairments and completed the study on a laptop or desktop computer (not mobile devices). They provided informed consent electronically and were compensated $2.75 for their time, corresponding to $15 USD per hour (approximately 10–13 minutes). The protocol was approved by the Institutional Review Board (IRB) of an institution affiliated with the authors. Based on the power analysis of Zimmermann et al. (2023), a minimum of 60 participants per condition (*i.e.*, 120 participants per experiment) was needed to obtain statistically robust results ($p < 0.05$). Participants were required to: (1) succeed in at least 5 of the 9 practice trials, (2) correctly answer at least 4 of the 5 catch trials (attentiveness tests) randomly inserted throughout the experiment, and (3) complete the experiment within 3 standard deviations of the mean completion time for that experiment.

[3] `www.prolific.com`



(a) Experiment I.

(b) Experiment II and III.

Figure 3: **Results for (a) Experiment I, and (b) Experiments II (ResNet50) and III (VGG16).** Given a unit and a set of images illustrating it, we assess how visually coherent participants find this set of images, or equivalently, how unambiguous the underlying pattern is. Specifically, we measure the proportion of trials in which participants correctly identified the query image belonging to this set of images and refer to this value as *Interpretability*. Across all experiments, participants found it significantly easier to identify the consistent pattern in the dictionary-based condition than in the neuron-based condition. Additionally, ResNet50 is more interpretable than VGG16 in the dictionary-based condition but not in the neuron-based condition.

4 Results

4.1 Dictionary elements are at least as decision-relevant as individual neurons

Interpretability aims to enable humans to understand the decision-making processes of machine learning models. Accordingly, it is essential to first verify that the elements under investigation genuinely influence model decisions. We therefore assess the relative importance of dictionary elements relative to neuron axes, confirming that the former are at least as informative as the latter for driving model outputs.

We quantify the importance of each element used in our psychophysics experiments by measuring the sensitivity of model decisions to perturbations of that element across its 300 most activating images (see Eq. 1). For each matched pair, we compute the difference in importance between the dictionary element and its corresponding neuron. We find no significant difference in ResNet50 ($M = 0.0002, SD = 0.0018$), $t(77) = 0.93$, $p = 0.36$, while dictionary elements are slightly more important than neurons in VGG16 ($M = 0.002, SD = 0.005$), $t(76) = 3.34$, $p = 0.001$ (One-sample t-test against 0).

These results show that dictionary elements have similar influence on model outputs as their corresponding neurons, attesting to the relevance of the selected elements for interpretability and supporting the fairness of our comparisons by matching elements of comparable importance.

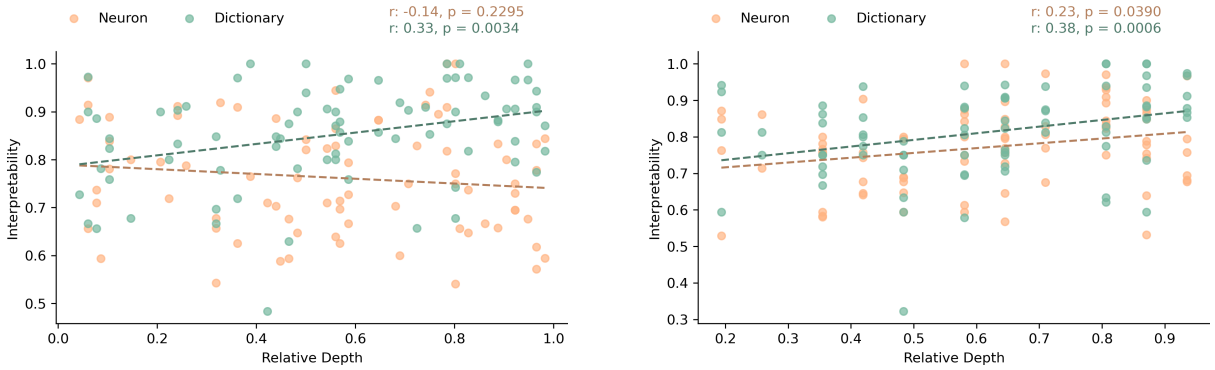
4.2 Interpretable features should be understandable by humans

In this section, we summarize the results obtained from analyzing the responses from 133, 130, and 129 participants who successfully completed Experiments I, II, and III.

Replication of previous findings. The experimental protocol employed in Experiment I for the neuron-based condition is the same as that described by Zimmermann et al. (2023). Thus, we first assess the extent to which our results replicate previous findings. For ResNet50, Zimmermann et al. (2023) report an average

task performance of $83.0\% \pm 2.0$ [4]. In our experiment, we obtain an average performance of $78.8\% \pm 1.5$. Given the similarity of these results, and considering that the specific selected units are not exactly the same (as described in Section 3), we conclude that Experiment I reproduces previously reported findings for neuron-based representations. This result also provides external validation of our experimental protocol.

Human performance is superior in the dictionary-based condition. Based on our main hypothesis that dictionary-based representations constitute a better basis for interpretability than neuron-based representations, we predicted that participants would perform better in the dictionary-based condition. Results across all three experiments are illustrated in Fig. 3. In Experiment I, the average performance across participants in the dictionary-based condition was • $83.5\% \pm 1.4$ [5] when compared to • $78.8\% \pm 1.5$ in the neuron-based condition. A Mann-Whitney U test revealed that participants performed significantly better in the dictionary-based condition $z = 3.12$, $p < .001$. This result was corroborated in both Experiments II (see examples of trials in Figs. A2–A4) and III, with a mean participant performance of • $85.1\% \pm 1.4$ *vs.* • $76.2\% \pm 1.6$, $z = 4.83$, $p < .001$ and a mean performance of • $81.5\% \pm 1.5$ *vs.* • $77.1\% \pm 1.6$, $z = 2.46$, $p = 0.01$, respectively. In subsequent analyses, we focus on the results from Experiments II and III.



(a) ResNet50. (b) VGG16.

Figure 4: **Correlation between human performance and the relative depth of the layer where the feature is extracted from for (a) ResNet50 and (b) VGG16.** For both models, we only find a significant correlation between interpretability and depth in the dictionary-based condition, after correcting for multiple comparisons (Bonferroni correction).

The deeper the layer, the more prominent the benefits of dictionary-based representations. Interestingly, our results suggest that the benefits of dictionary-based representations increase with relative layer depth in both models. Specifically, we observe a significant positive correlation between *interpretability* and *depth* for the dictionary-based condition only: $r = 0.35$, $p = 0.001$ for ResNet50 and $r = 0.38$, $p < 0.001$ for VGG16. In contrast, the correlations for the neuron-based condition are weaker and not statistically significant after Bonferroni correction: $r = 0.09$, $p = 0.4$ for ResNet50 and $r = 0.24$, $p = 0.04$ for VGG16. These patterns are illustrated in Fig. 4a and 4b, corresponding to Experiments II and III, respectively.

4.3 Comparison of model interpretability

While evaluating the interpretability of a specific machine learning model is important for validating its transparency for downstream use cases, it is equally important to understand what makes certain models more interpretable than others. Doing so requires fair model comparisons. In this section, we report three main findings obtained by comparing ResNet50 and VGG16.

ResNet50 exhibits greater superposition than VGG16. Superposition arises when a model attempts to represent more patterns than it has neurons, resulting in patterns that are encoded by populations of

[4] Values inferred from Fig. 3 in Zimmermann et al. (2023)
[5] The values reported correspond to a 95% confidence interval.

neurons rather than by single units. As the number of encoded patterns increases, individual dictionary elements are expected to be less aligned with any single neuron axis. As summarized in Fig. 5, using Eq. 2, we measure the average axis-alignment of dictionary elements extracted with CRAFT for both models. We find that ResNet50 exhibits lower average axis-alignment than VGG16 (0.40 *vs.* 0.45, $z = 6.01, p < .001$, Mann–Whitney U test), indicating that its dictionary elements are more distributed across neurons. This finding may suggest a need for ResNet50 to represent a larger number of sparse features, a pattern consistent with greater superposition. Consequently, evaluating its neurons in isolation may lead to an underestimation of its interpretability.

ResNet50 is more interpretable than VGG16. Within the Rashomon set of similarly accurate models—*e.g.*, ResNet50 achieves 76% accuracy compared to 72% for VGG16 on ImageNet, some models are expected to be more interpretable than others. A common assumption in mechanistic interpretability is that sparser representations should be more interpretable.

Based on the theoretical intuition that neural networks encode patterns in superposition and that interpretability improves when dictionary elements are sparse, we hypothesized that ResNet50 would exhibit greater interpretability than VGG16 in the dictionary-based condition. Our results corroborate this intuition: ResNet50 achieves significantly higher interpretability than VGG16 ($85.1\% \pm 1.4$ *vs.* $81.5\% \pm 1.5$, $z = 2.14, p = 0.016$) in the dictionary-based condition. These findings contradict prior work (Zimmermann et al., 2024), which reported higher interpretability for VGG16 with 89.27% *vs.* 87.40% for ResNet50, underscoring the critical role of the representational basis when evaluating model interpretability.

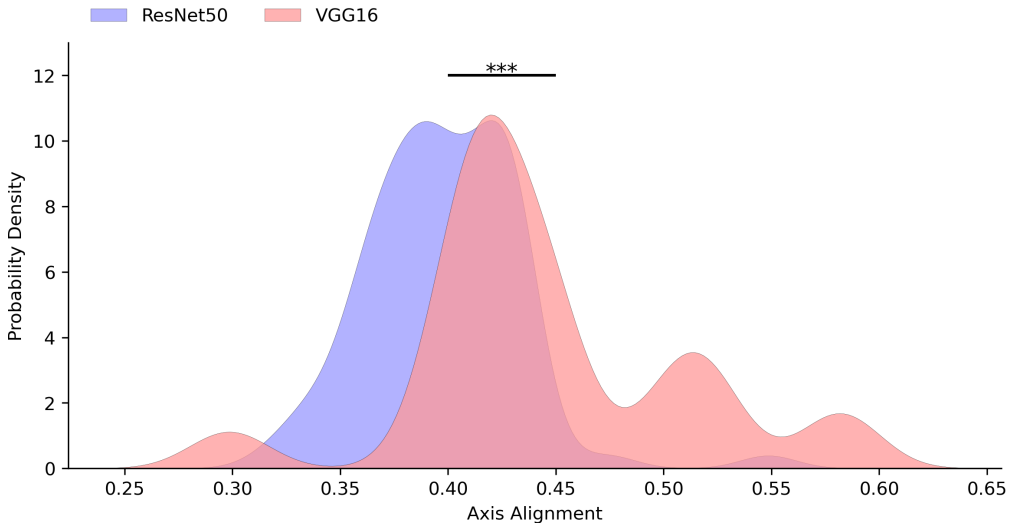


Figure 5: **Axis-alignment as a proxy measure of superposition.** We measure the axis-alignment of dictionary elements recovered by CRAFT for both models using the sparsity measure of Hoyer (2004) (see Eq. 2). Higher values indicate stronger alignment with individual neuron axes, *i.e.*, the recovered dictionary elements can be well approximated by single neurons. We find that ResNet50 has significantly lower axis-alignment than VGG16: 0.40 *vs.* 0.45, $z = 6.01, p < .001$ (Mann Whitney U test). This suggests that ResNet50 is more impacted by superposition.

The choice of basis matters when comparing model interpretability. Fig.3b depicts the average interpretability of ResNet50 and VGG16 for each representational basis (neuron-based vs. dictionary-based). While ResNet50 achieves higher interpretability than VGG16 in the dictionary-based condition, this advantage disappears when interpretability is assessed using neuron-based representations ($z = 0.56, p = 0.71$). These results suggest that a model's apparent interpretability critically depends on the choice of representational basis. Importantly, they challenge prior claims about model interpretability (Zimmermann et al., 2023), and highlight the need to consider dictionary-based representations when evaluating and comparing models.

5 Conclusion

In this work, we have investigated the suitability of neuron-based and dictionary-based representations as a basis for interpretability in two convolutional neural networks, ResNet50 and VGG16. Across three large-scale psychophysics experiments, we find converging evidence that the choice of representational basis significantly influences human interpretability: dictionary-based representations are consistently more interpretable than neuron-based representations, with the advantage increasing in deeper layers. Furthermore, we find models rely no less on dictionary elements than on neuron axes to make their decisions.

We also observe that the extent to which models are affected by superposition influences their apparent interpretability. ResNet50 exhibits more superposition than VGG16, resulting in comparable interpretability when assessed using neuron-based representations but superior interpretability when evaluated using dictionary-based representations. This finding is particularly notable given that ResNet50 is the model with higher classification accuracy.

These findings demonstrate that the choice of representational basis matters when comparing models, suggesting a need to reinterpret prior results. Overall, our results highlight that dictionary-based representations not only constitute a superior basis for interpretability but also that comparing models using neuron-based representations alone can lead to misleading conclusions.

Limitations and future work. Our study is not without limitations. First, the methodology proposed by Borowski et al. (2021) and followed by Zimmermann et al. (2023) utilizes the entire ImageNet validation set with the goal of studying a broad range of stimuli and thereby increasing the likelihood of identifying stimuli that are representative of a neuron's selectivity. While this motivation is sound, the trade-off is that neurons selective to class-specific patterns (*e.g.*, fish scales) will be maximally activated by stimuli from the corresponding classes (*e.g.*, fish) and minimally activated by stimuli from other classes (*e.g.*, dogs). In such cases, the task can be solved trivially using semantics. The design of Experiment II reflects an initial attempt to mitigate this confound. However, manual exploration of the trials by the authors suggests only partial success in addressing this challenge. We leave further refinement of the experimental protocol to future research.

Second, this work focused on investigating whether analyzing individual neurons reflects a mixture of pattern detectors in superposition. This approach enabled a fair one-to-one comparison of elements recovered from both neuron-based and dictionary-based representations, providing evidence that dictionary learning methods can improve our understanding of model representations. However, future research should examine the interpretability of dictionary elements recovered by applying dictionary learning to full layer activations.

Finally, our experiments were limited to two models. While two models were enough to illustrate that the choice of representational basis influences interpretability comparisons, future work should expand the range of architectures to understand why certain models are inherently more interpretable than others.

References

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Judy Borowski, Roland S. Zimmermann, Judith Schepers, Robert Geirhos, Thomas SA. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain cnn activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020a.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020b. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2832–2845, 2022.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

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*, pp. 248–255. Ieee, 2009.

R. Becket Ebitz and Benjamin Y. Hayden. The population doctrine in cognitive neuroscience. *Neuron*, 109 (19):3055–3068, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Mathieu Chalvidal, Thomas Serre, et al. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16543–16572. PMLR, 2025.

Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pp. 231–243. Springer, 2010.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pp. 9273–9282, 2019.

James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293 (5539):2425–2430, 2001.

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 (CVPR)*, 2016.

Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. Proceedings of the International Conference on Machine Learning (ICML), 2018.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, 2010.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *Visualization for Deep Learning workshop, Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017.

R. Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

R. Quian Quiroga, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Sparse but not 'grandmother-cell'coding in the medial temporal lobe. *Trends in cognitive sciences*, 12(3):87–91, 2008.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting clip with hierarchical sparse autoencoders. In *Forty-second International Conference on Machine Learning*.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of cnn activations? *Advances in Neural Information Processing Systems*, 34:11730–11744, 2021.

Roland S. Zimmermann, Thomas Klein, and Wieland Brendel. Scale alone does not improve mechanistic interpretability in vision models. *Advances in Neural Information Processing Systems*, 36, 2023.

Roland S. Zimmermann, David Klindt, and Wieland Brendel. Measuring per-unit interpretability at scale without humans. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 48448–48483. Curran Associates, Inc., 2024. doi: 10.52202/079017-1535. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/56ed2bd15b66f709cd81cb1aaa0496b9-Paper-Conference.pdf`.