IDENTIFYING INTERPRETABLE VISUAL FEATURES IN ARTIFICIAL AND BIOLOGICAL NEURAL SYSTEMS

David Klindt
Stanford University
klindt.david@gmail.com

Sophia Sanborn
UC Santa Barbara
sanborn@ucsb.edu

Francisco Acosta
UC Santa Barbara
facosta@ucsb.edu

Frédéric Poitevin
SLAC National Accelerator Laboratory
frederic.poitevin@stanford.edu

Nina Miolane
UC Santa Barbara
ninamiolane@ucsb.edu

ABSTRACT

Single neurons in neural networks are often "interpretable" in that they represent individual, intuitively meaningful features. However, many neurons exhibit *mixed selectivity*, i.e., they represent multiple unrelated features. A recent hypothesis proposes that features in deep networks may be represented in *superposition*, i.e., on non-orthogonal axes by multiple neurons, since the number of possible interpretable features in natural data is generally larger than the number of neurons in a given network. Accordingly, we should be able to find meaningful directions in activation space that are not aligned with individual neurons. Here, we propose (1) an automated method for quantifying visual interpretability that is validated against a large database of human psychophysics judgments of neuron interpretability, and (2) an approach for finding meaningful directions in network activation space. We leverage these methods to discover directions in convolutional neural networks that are more intuitively meaningful than individual neurons, as we confirm and investigate in a series of analyses. Moreover, we apply the same method to two recent datasets of visual neural responses in the brain and find that our conclusions largely transfer to real neural data, suggesting that superposition might be deployed by the brain. This also provides a link with disentanglement and raises fundamental questions about robust, efficient and factorized representations in both artificial and biological neural systems.

1 INTRODUCTION

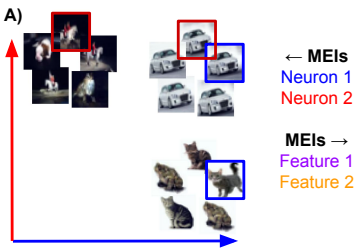

Figure 1: **Conceptual Overview. A**) A representation of two neurons' activations for different images. The highlights indicate maximally exciting images (MEIs) for each neuron. **B**) There exist directions in feature space that are more interpretable.

One of the oldest ideas in neuroscience is Cajal's *single neuron doctrine* (Finger, 2001) and its application to perception (Barlow, 1972), i.e., the hypothesis that individual sensory neurons encode individually meaningful *features*.[1] The idea dates back to the early 1950s, when researchers began to find evidence of neurons that reliably and selectively fire in response to particular stimuli, such as dots on a contrasting background (Barlow, 1953) and lines of particular orientation and width (Hubel & Wiesel, 1959). These findings gave rise to the *standard model* of the ventral visual stream as a process of hierarchical feature extraction and pooling (Hubel & Wiesel, 1968; Gross et al., 1972;

[1]In this work, we adopt a pragmatic definition of *feature* based on human discernability, measured through psychophysics experiments (see below). For an attempt at a more formal definition see Elhage et al. (2022).

Riesenhuber & Poggio, 1999; Quiroga et al., 2005). Neurons in the early stages extract simple features, such as oriented lines, while neurons at later stages combine simple features to construct more complex composite features. In the highest stages, complex features are combined to yield representations of entire objects encoded by single neurons—the shape of a hand, or the face of a friend. Notwithstanding a shift in focus towards population codes (Averbeck et al., 2006; Stanley, 2013; Hebb, 2005; Gao & Ganguli, 2015; Jacobs et al., 2009; Ebitz & Hayden, 2021), this model has remained a dominant paradigm in sensory neuroscience for the last seven decades and ultimately inspired (Hassabis et al., 2017; Zador et al., 2023) the development of convolutional neural networks (CNNs) (Fukushima, 1980; LeCun et al., 1989) (but see Gershman, 2023; Poggio et al., 2017).

Mechanistic interpretability research aims to uncover the coding properties of neurons within artificial neural networks. *Feature visualization* (Nguyen et al., 2019)—i.e. the single-unit electrophysiology of artificial neural networks—has revealed remarkable consistencies between neurons in image models and neurons in the visual cortex: neurons with center-surround receptive fields, color-contrast detectors, and oriented edge detectors that combine to form curve detectors in higher layers, for example (Olah et al., 2020; Willeke et al., 2023). However, the study of individual neurons, both *in vitro* and *in silico*, faces two major problems. First is the inherent subjectivity of "interpretability," which generally necessitates the hand-inspection of neuron response properties. Second is the ubiquitous existence of hard-to-interpret units with mixed selectivity (Fusi et al., 2016; Olah et al., 2020).[2] We address both problems in this work by (1) defining a quantitative, automated measure of interpretability for vision models that does not rely on human inspection, and (2) demonstrating a simple approach for finding meaningful directions in activity space.

A recent paper by Zimmermann et al. (2023) has taken a similar approach, using human perceptual judgments in large-scale psychophysics experiments to quantify the interpretability of neurons within deep image models (Zimmermann et al., 2021; Borowski et al., 2020). We automate this pipeline by replacing human judgments of perceptual similarity with a similarity metric grounded in deep image model representations (Zhang et al., 2018), and validate the approach against the large scale human data from Zimmermann et al. (2023). Thus, in line with recent work that uses *language models to interpret language models* (Bills et al., 2023), we use *image models to interpret image models*. We then leverage this automated index of interpretability to test whether non-axis aligned directions in the neural activation space of CNNs trained on real data may be more interpretable than the individual units—a test of the recently stated *superposition hypothesis* (Elhage et al., 2022).

Through a suite of experiments and analyses, we find evidence consistent with the hypothesis that neurons in both deep image models and the visual cortex encode features in superposition. That is, we find non-axis aligned directions in the neural state space that are more interpretable than individual neurons. In addition, across both biological and artificial systems, we uncover the intriguing phenomenon of what we call *feature synergy*—sparse combinations in activation space that yield more interpretable features than the constituent parts. Our work pushes in the direction of automated interpretability research for CNNs, in line with recent efforts for language models (Bills et al., 2023; Cunningham et al., 2023; Gurnee et al., 2023; Bricken et al., 2023). Simultaneously, it provides a new framework for analyzing neural coding properties in biological systems. Our results on neuroscience data add nuance to the concepts of *disentanglement* and *mixed selectivity* in the brain and suggest that insights gleaned from studying the coding properties of artificial neural networks may transfer to biological systems. These findings highlight potential synergy between mechanistic interpretability research and computational neuroscience, which may together reveal universal coding principles of neural representations.

2 METHODS

We propose an approach for quantifying the interpretability of neural network activations that is grounded in human judgement, yet is fully automated and scalable. In general, individual neurons — i.e., N directions corresponding to basis vectors of an activation space $\mathbb{R}^N$ — might not be interpretable. Yet, other directions in $\mathbb{R}^N$ might be: we refer to them as *features*. For example, in Fig. 1 B) the human observer can define three directions that are interpretable and correspond

[2]One might wonder why evolution or gradient descent would be so kind as to make any neurons interpretable. Anecdotally, researchers have explained this as a result of the use of pointwise nonlinearities in deep networks. We provide a more formal argument for this explanation in App. E.

approximately to horse-, car- and cat-like images. The superposition hypothesis stipulates that the activation space $\mathbb{R}^N$ of a neural network possesses several interpretable directions that are non-orthogonal (Elhage et al., 2022). Given a CNN, we aim to identify such directions and quantify their interpretability through the following three steps:

1. **Collect neural network activations for a given dataset**. Images are passed through the network up to the layer under analysis, for convolutional layers, we average activations across space (Zimmermann et al., 2023) to generate a dataset in activation space $\mathbb{R}^N$.
2. **Identify directions in activation space.** Directions may be provided by the neurons themselves (basis vectors) or by an algorithm (e.g., PCA, sparse coding, K-Means).
3. **Quantify the interpretability of each direction**. We compute an interpretability index (II) as the average pairwise similarity of the top M Maximally Exciting Images (MEIs, defined in the next subsection) for each direction. Through a suite of experiments, we argue that the II is a meaningful measure of interpretability.

2.1 Quantifying Interpretability in Neural networks

A neural network layer defines an activation space $f : \mathcal{X} \to \mathbb{R}^N$ with N the number of neurons of that layer. We consider directions in this space, for example, individual neurons are represented as directions: the basis vectors of $\mathbb{R}^N$, i.e., for neuron i, $f_i(x) = f(x)e_i$ with e_i the i^{th} canonical basis vector and similarly $f_u(x) = f(x)u$ for any direction $u \in \mathbb{R}^N$. In activation space, some directions may be *interpretable*, in the sense that they detect a single feature or concept within the image data. For example, an interpretable direction may detect features such as edges, corners, textures in early layers, or more abstract patterns in later layers such as dogs, cats, trucks. By contrast, other directions respond to several unrelated features or concepts. For instance, Fig. 1 (A) shows the first neuron firing in response to unrelated car- or cat-like images.

Maximally Exciting Images (MEIs) are defined as synthetic images that maximally activate a given direction in activation space (Erhan et al., 2009). Given a direction u, we propose an **Interpretability Index (II)** computed as the average pairwise similarity of its top $M = 5$ MEIs from a dataset of D images, i.e., $f_u(x_1) \geqslant ... \geqslant f_u(x_M) \geqslant ... \geqslant f_u(x_D)$:

$$\mathrm{II}(u) = \frac{1}{M} \sum_{j=1}^{M} \sum_{k=1}^{M} \mathrm{sim}\Big(x_j, x_k\Big). \tag{1}$$

In this work, we consider and compare several similarity metrics sim, detailed below.

2.2 Image Similarity Metrics

We consider image similarity metrics that capture notions of similarity at different levels of abstraction: **1) Low-Level: Color** The color similarity between two images is defined as the difference between the average color (averaged across space, independently, for each color channel) in each image; **2) Mid-Level: LPIPS** Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) is a perceptual metric used for assessing the perceptual differences between images. It relies on a CNN such as VGG or AlexNet that has been pre-trained on an image classification task. Given two images, LPIPS extracts their respective feature maps from multiple layers of the CNN. LPIPS then computes the distance between the corresponding feature maps. The distances are scaled by learned weights and then aggregated to yield a single scalar value representing the perceptual similarity between the two images; **3) High-Level: Labels** The label similarity between two images is a value equal to 0 if the two images have been assigned different labels during a reference classification task, and equal to 1 if the two images have been assigned the same label. In our experiments, we use the CIFAR-10 dataset and associated classification task.

2.3 From Human Psychophysics to In-Silico Psychophysics

How can we validate whether the proposed interpretability index from Eq. 1 is indeed a sensible measure of interpretability? The concept of *interpretability* is intimately tied to human judgment. A long history of theoretical inquiry has demonstrated the impossibility of identifying necessary

and sufficient conditions for many natural semantic categories (Stekeler-Weithofer, 2012). Due to this difficulty, we adopt a pragmatic view, converting the question of whether a representation is interpretable into an empirical measure of the human interpretability judgment (Wittgenstein, 1953).

HUMAN PSYCHOPHYSICS

Psychophysics is an experimental paradigm for quantifying the relationship between stimuli (e.g. images) and the perceptions they produce for human observers. Borowski et al. (2020) and Zimmermann et al. (2023) have demonstrated that large-scale psychophysics experiments can be leveraged for conducting quantitative interpretability research. In these works, researchers used the judgments of human participants to quantify the interpretability of neurons in trained artificial neural networks. In Zimmermann et al. (2023), participants are shown 9 minimally and 9 MEIs for a given neuron. The participant is then asked to select one of two query images x_1, x_2 that they believe also strongly activates that neuron (see App. A for an illustration). The *(human) psychophysics accuracy* obtained for that neuron is defined as the percentage of participants that are able to select the correct image.

IN-SILICO PSYCHOPHYSICS

Psychophysics experiments provide a way of crowd-sourcing and quantifying human intuition of interpretability at scale. However, such experimentsts are time consuming, noisy and costly ($12,000 for (Zimmermann et al., 2023)). Here, we propose a method for automating psychophysics experiments, with a model that faithfully approximates human judgments while requiring no human input. We replicate, *in-silico*, the experiments of Zimmermann et al. (2023), comparing different similarity metrics as proxies for human judgments. In our experiments, the model computes the maximum similarity, according to the image similarity metric, between each of the query images x_1, x_2 and the set of MEIs. The model then chooses as its response the image that is the closest to that set

$$\text{sim}(x, \text{MEI}(u)) = \max_{k=1,\dots,9} \text{sim}(x, x_k), \tag{2}$$

where $x_1, ..., x_9$ are the top 9 MEIs for a neuron or direction u, and sim the image similarity metric. The *psychophysics accuracy* for a given neuron or direction u is defined as the percentage at which the model selects the correct query image for that neuron, i.e.:

$$\text{Acc}(u) = \frac{\text{\# of correct selections for direction } u}{\text{\# of queries with direction } u}. \tag{3}$$

We check in practice that directions u with high interpretability index $\text{II}(u)$ are indeed more interpretable to a human observer. We note that we could have chosen the *in-silico* psychophysics accuracy $\text{Acc}(u)$ of direction u to quantify its interpretability: the more interpretable u is, the easier it is for participants to correctly select images associated with it (Zimmermann et al., 2023). However, we observe in practice that $\text{Acc}(u)$ is often at ceiling, and does not contain as much information as our proposed $\text{II}(u)$. Since it is also more expensive to compute, we use it only to validate the viability of the $\text{II}(u)$. Since the human psychophysics accuracy was computed only for directions corresponding to individual neurons, having in-silico psychophysics experiments is a key component of our approach. Note that we chose to work with MEIs rather than feature visualisations (Simonyan et al., 2013; Nguyen et al., 2019), because the latter has shown consistently lower interpretability in psychophysics studies (Borowski et al., 2020; Zimmermann et al., 2021; 2023).

3 RESULTS

3.1 COMPARISON: HUMAN VS. IN-SILICO PSYCHOPHYSICS

We first test the LPIPS similarity metric as a model of human perception of neuron interpretability. The experiments from Zimmermann et al. (2023) quantify the interpretability of neurons in models trained on ImageNet-1k through crowd-sourced human perceptual judgments. We reproduce this experiment *in-silico* by presenting the same image queries to a simple model based on the LPIPS metric. For each query, we evaluate whether the image selected by the model agrees with the image selected by human participants: it is a correct classification if they agree, incorrect otherwise. The

results of this binary classification task are shown in Figure 2 for the LPIPS image similarity metric (Zhang et al., 2018), on five of its AlexNet layers.

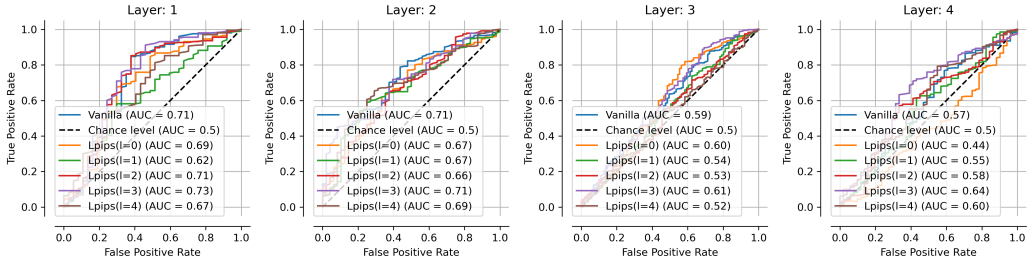

Figure 2: **Interpretability Metric vs. Human Behaviour.** Data from Zimmermann et al. (2023). Left to Right: Agreement between human and in-silico psychophysics on the predictability of the outputs of four 'layers' (see text) within a ResNet50. Human and model agree on feature predictability for the ResNet50's early layers. For these layers, the proposed interpretability metric is a valid representation of the human's perception of interpretability. AUC: Area Under the Curve.

The predictions of the LPIPS model match human judgments well for earlier layers of the ResNet50—*layers* 1 and 2[3]—with an AUC up to 0.71 (Figure 2, left two panels). While there is certainly room for improvement, we conclude that this metric, based on LPIPS-based pairwise image comparison, serves as a good first proxy of human perception of interpretability. Crucially, our metric has the added benefit of not having to recruit a cohort of human participants. Thus, we will use this metric to evaluate the interpretability of features across neural network layers in the next subsections. Since the interpretability metric is more accurate for early layers, we focus the remainder of our analyses on layer 1 of the same ResNet50 architecture trained on CIFAR-10.

3.2 Identifying Interpretable Directions in Feature Space

We next apply the II to analyze the interpretability of features in a ResNet50 pre-trained on the CIFAR-10 image dataset (Krizhevsky et al., 2009)[4]. We evaluate several methods for identifying interpretable directions in activation space: PCA, ICA, NMF, K-Means with cosine similarity, and the shallow sparse autoencoder used in Sharkey et al. (2022) (see Appendix B for sparse AE analyses). We evaluate several similarity metrics and compute the II for each, comparing the interpretability of individual neurons in a layer with the interpretability of identified directions from that layer.

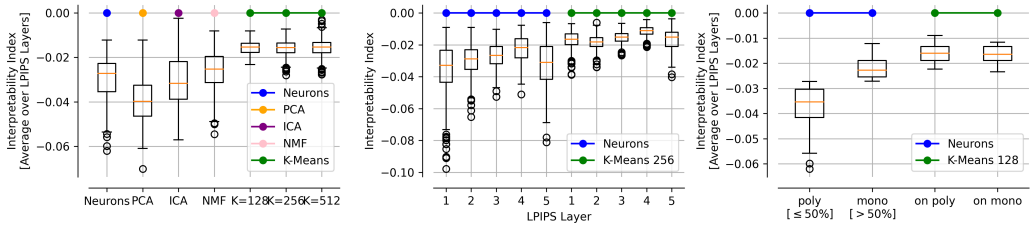

Figure 3: **Quantification of Interpretability.** Left: II score [a.u.] distribution for neurons ($N = 256$), PCA, ICA and NMF baselines, and K-Means as a function of $K \in \{128, 256, 512\}$. Middle: II score distribution for neurons ($N = 256$) and K-Means ($K = 256$) as a function of LPIPS layer. Right: II score distribution for uninterpretable neurons ($N = 128$), i.e. those with II score below the median, and interpretable neurons ($N = 128$), i.e., those with II score above the median; II score distribution for K-Means ($K = 128$) computed on each of these subsets separately.

For a quantitative comparison, we present comparative box plots for the distributions of II indices for neurons and directions in Figure 3 while we vary: the LPIPS layer defining the similarity metric

[3]This refers to the PyTorch module names, corresponding to layers 10 and 23 in the network.

[4]Hosted at https://github.com/huyvnphan/PyTorch_CIFAR10

used in the II (left), the number of K directions in activation space (middle), and distributions after splitting neurons into uninterpretable and interpretable groups (right). We observe that the K-means approach detects directions that are indeed more interpretable (higher II) than the individual neurons of the activation space — independently of the LPIPS layer considered for the II (Figure 3 left). Our method achieves higher II values (mean $= -0.0159$) than all baselines and the sparse autoencoder (best mean $= -0.0188$) (detailed comparison in App. B) [note that the II has arbitrary units]. Interestingly, the number of directions K does not impact their IIs in the regime tested (Figure 3 left). Thus, we focus our analyses on the $K = N = 256$ setting for a fair comparison.

The K-means approach can detect directions within subsets of uninterpretable neurons as well as within interpretable neurons, as we do not observe II differences in Figure 3 (right). Further, transforming interpretable neurons and uninterpretable neurons into their direction increases the II of both. We see a trend where the II increases with respect to the LPIPS layer used, which is a similar pattern as we saw in Figure 2.

For a qualitative comparison, Figure 4 shows the Maximally Exciting Images (MEIs) for 5 neurons (left) and 5 directions extracted from K-Means (right) selected in 5 different quantiles of II values (to avoid cherry picking in this qualitative comparison). The distributions of II indices is shifted towards the higher values for the directions detected by K-means, as shown by the II values associated with each quantile. This is confirmed by the visualization of MEIs, which appear more visually coherent to the human observer for the directions (right) compared to the neurons (left).

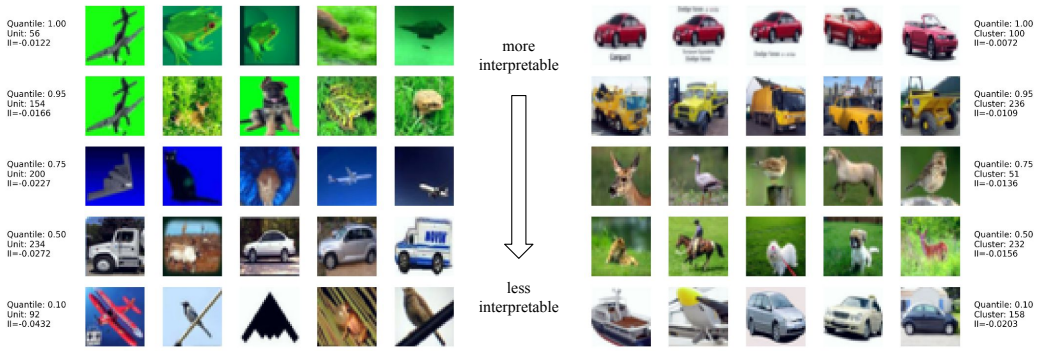

Figure 4: **MEIs of Neurons and Interpretable Directions.** These are the Maximally Exciting Images (MEIs) for neurons (left) and directions (right) as retrieved with K-Means. To represent the interpretability index (II) distribution, we show neurons and directions at different II quantiles.

3.3 Comparing Similarity Metrics for In-Silico Psychophysics

We now compare the interpretability of the directions measured using the three image similarity metrics described in Section 2. Each metric defines similarity at a different level of abstraction, from low-level to high-level: same *color* (Figure 5 left), same *perceptual structure* as defined by LPIPS (Figure 5 middle) or same *category* (Figure 5 right). For each metric, we perform the *in-silico* psychophysics task from Section 3.1, varying the difficulty of the psychophysics experiment. The difficulty of a task is controlled by choosing query images that cause less extreme activations—i.e. are farther away from the set of MEIs (Borowski et al., 2020). This allows us to probe a more general understanding of the interpretability of a neuron or direction instead of limiting our analyses to the most preferred stimuli (Vinken et al., 2023).

As expected, we see in Figure 5 that both the neurons and the directions have a decreased psychophysics accuracy as the task becomes more difficult. The directions detected by our approach are more predictable than the individual neurons across low, mid and high-level semantics and across task difficulties. The largest improvement over individual neurons is observed for the low-level semantics using colors, and the improvement decreases as we move towards higher level semantics. Additionally, as observed in Figure 3, the number of clusters K does not impact the accuracy.

Lastly, in recently published work, Bricken et al. (2023) test whether observed interpretability in activation space is a function of the model or the data—that is, they test whether untrained models possess non-axis aligned directions that are more interpretable than individual neurons and find

evidence that they do. We perform the same experiment, running our analysis on untrained versions of the models we analyze here, and find that there is indeed—even before training—a gap in interpretability between neuron axes and activation clusters (see Appendix G). This aligns with prior work on the expressive power of untrained CNNs (Frankle et al., 2020) and suggests paths for further investigation.

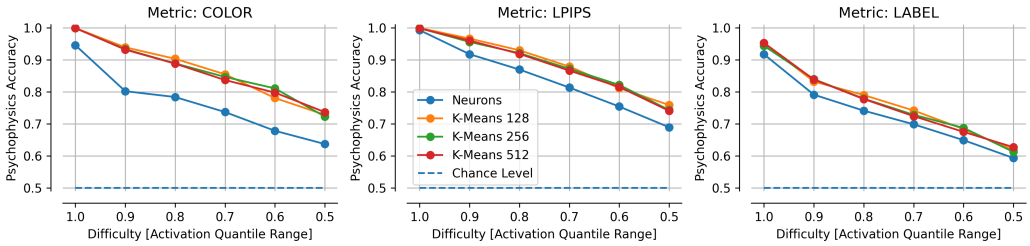

Figure 5: ***In Silico* Psychophysics Performance.** Accuracy across neurons and interpretable directions revealed by K-Means clusters ($K \in \{128, 256, 512\}$) for *in silico* psychophysics task for different levels of difficult, i.e., limiting query and reference image selection to the central range of activations (e.g., from the 0.45^{th} until the 0.55^{th} quantile, see (Zimmermann et al., 2023)). Predictions are made based on different metrics from low level semantics (colour match, left), over mid level semantics (LPIPS average over layers, center), to high level semantics (label match, right).

3.4 Pairwise Synergies Between Neurons

Efficient coding principles such as minimal wiring length (Laughlin & Sejnowski, 2003), as well as the circuit analysis approach of mechanistic interpretability (Conmy et al., 2023; Nanda et al., 2023) inspire us to look for minimal subcircuits that increase interpretability. Specifically, we investigate the synergies between pairs of neurons. For all pairs of neurons a, b in the same ResNet50 layer, we compute the II score for their added (z-scored) activity. The *synergy* is the difference between this II score and the maximum of their individual II scores to account for pairings with highly interpretable neurons:

$$\text{Synergy}(a, b) = \text{II}(a + b) - \max\left[\text{II}(a), \text{II}(b)\right]. \tag{4}$$

The synergy measures whether adding these neurons produces a direction in activation space that is more interpretable than taking each neuron individually. This is visualized in Figure 6 A) and B) which show two pairs of neurons a, b with the highest synergy: the MEIs resulting from their addition are more interpretable that their individual MEIs.

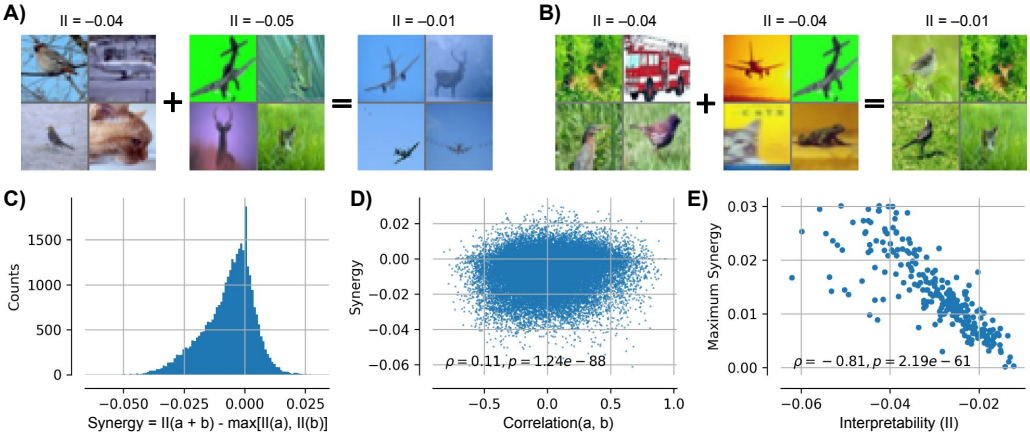

Figure 6: **Synergies. A), B)** Example pairs (two highest synergies) of neurons and the result when adding them (all visualized by their 4 MEIs). **C)** Histogram of synergies for every pair of neurons. **D)** A slight positive relationship between the correlation and synergy over all pairs of neurons (i.e., more correlated neuron pairs have higher synergies). **E)** A strong negative relationship between the II of a neuron and the maximum synergy it can achieve (i.e., pairings dilute interpretable neurons).

The histogram of Figure 6 C) shows a large fraction of negative values of the synergy, i.e., most pairings are, as expected, detrimental for interpretability. However, a good fraction of the added neurons $a+b$ become more interpretable. Figure 6 D) shows that correlated neurons tend to have higher synergy but correlation alone does not explain everything: two neurons can be uncorrelated, yet their addition can produce a very interpretable feature. This shows that our notion of interpretability is distinct from the familiar notion of decorrelation. Lastly, we find that more interpretable neurons (higher II) show lower maximal synergy (Figure 6 E)). This suggest that their representation is already interpretable and that any pairing would only dilute it.

3.5 Application to Biological Neural Data

Findings of *mixed selectivity*, i.e., hard to interpret neurons that code for multiple unrelated features have been reported before in neuroscience (Yoshida & Mori, 2007; Rigotti et al., 2013; Fusi et al., 2016). This suggests that the cortex may also encode meaningful features in superposition. Below, we perform the same analysis as above, but for cortical recordings from inferior temporal (IT) visual cortex in macaque monkeys—a cortical area involved in high level visual object recognition (Hung et al., 2005) with a specific preference for faces (Tsao et al., 2006).

3.5.1 Face Cell Responses

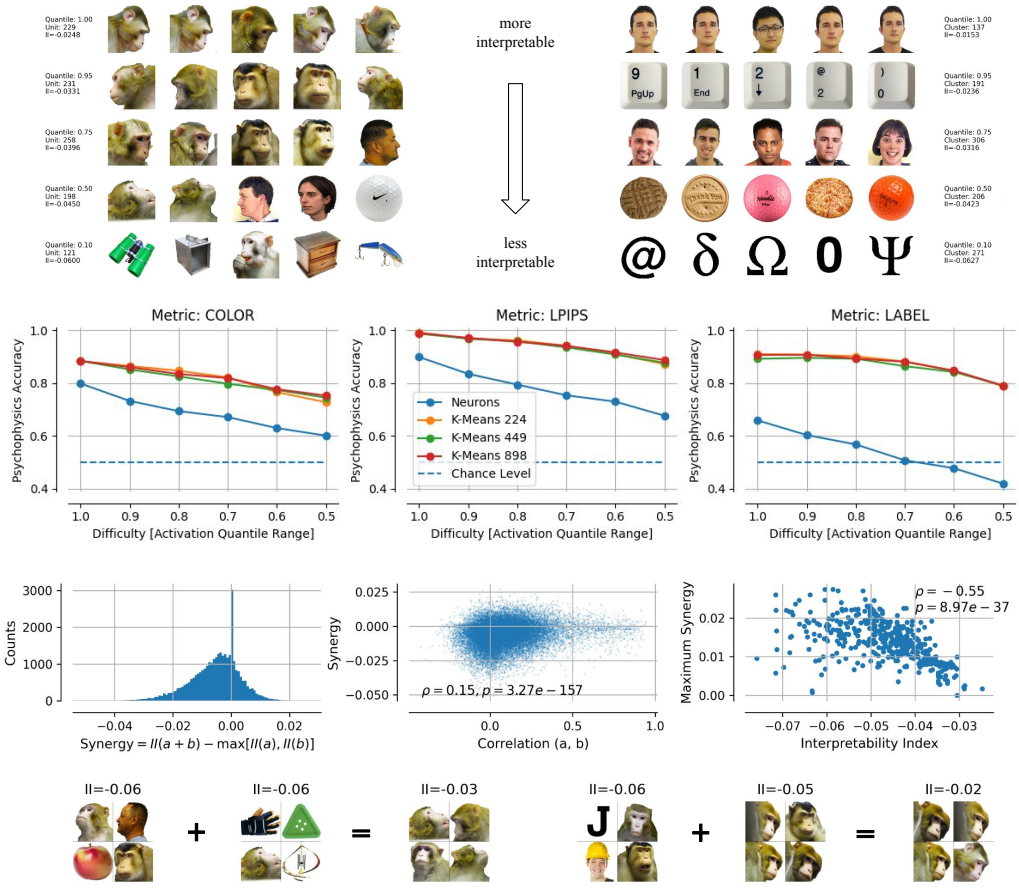

Figure 7: **Face Cell Responses in IT Cortex. 1st row)** Maximally Exciting Images (MEIs) for neurons (left) and directions (right) as retrieved with K-Means (see Fig. 4). **2nd row)** Accuracy across neurons and interpretable directions revealed by K-Means clusters for the *in silico* psychophysics task for different levels of difficulty (see Fig. 5). **3rd row)** (left to right, see Fig. 6) Histogram of synergies for each pair of neurons; Relationship between the correlation and synergy over all pairs of neurons; Relationship between the II of a neuron and the maximum synergy it can achieve. **4th row)** Example pairs (two highest synergies) of neurons and the result when adding them (see Fig. 6).

We first examine a dataset from Vinken et al. (2023), consisting of the responses of 449 neurons (sites) to 1379 images (447 faces 932 non-face objects). We perform the same analysis pipeline as for the CNN above, i.e., we clustered the activations using K-Means and studied the learned *features* by their MEIs and our *in silico* psychophysics. The results are shown in Fig. 7. The top row is the same as Fig. 4, showing quantiles of II sorted neurons/features. This is a rather striking demonstration of the previous claim that IT cortex represents a 'domain-general object space' (Vinken et al., 2023), i.e., we find highly interpretable activation clusters that code, e.g., for faces, keyboard keys, round objects or characters. In the center row of Fig. 7, we perform the same *in silico* psychophysics experiment as above (labels are now provided by 3 distinct image conditions, see (Van der Maaten & Hinton, 2008)). Intriguingly, we find the same if not a larger effect of increased interpretability when moving from individual neurons to the features we find with K-Means. Lastly, in the bottom row of Fig. 7, we perform the same synergy experiment as in Fig. 6 and also find the same qualitative pattern, including a skewed distribution over synergies, a positive link with pairwise neural correlations, and a negative link with the II score.

These results are an interesting extension of the conclusions from the original Vinken et al. (2023) paper, in which the authors concluded that MEIs give an incomplete picture and that face cells should rather be understood as representing a domain-general object space. We fully agree with the former conclusion—when limited to individual neurons. The additional insight that we obtain here is that the object space, represented by multiple IT neurons, is spanned by groups of features (in *superposition*) whose MEIs are meaningful in the sense that they correspond to interpretable coding directions (Fig. 7 (1^{st} row)) and whose activations are interpretable across a wide range of quantiles (Fig. 7 (2^{nd} row)).

3.5.2 Disentangling Interpretable Features in IT Cortex

Next, we apply this analysis to the dataset from Higgins et al. (2021), which consists of 159 neurons in anterior middle (AM) macaque face area that were presented with 2100 human and monkey face images. We apply the same analysis pipeline as before, i.e., we cluster the activations using K-Means and study the learned *features* through their MEIs and *in silico* psychophysics. The results are shown in Fig. 8.Note that the images used in the experiment were greyscaled. Thus, we are limited to considering only brightness rather than color for the low-level metric. For the mid-level LPIPS metric, we feed the greyscale value into all three colour channels. For this dataset, it is not possible to consider the label-based method, as there is no category information provided for the images in this dataset.

Again, the MEIs Fig. 8 (top) and *in silico* psychophysics Fig. 8 (center) tell a consistent story. That is, we find a significant increase in interpretability (psychophysics performance across all levels of difficulty) when moving from individual neurons to the K-Means features. Lastly, the synergy experiment Fig. 8 (bottom) also shows the same result pattern with a skewed synergy distribution, a positive link between pairwise neural correlations and synergies, and a negative link between interpretability and maximal synergy.

In the original paper by Higgins et al. (2021), the key insight was that *individual* neurons encode disentangled, interpretable features of the data. By contrast, we find that directions in activation space that mix multiple neurons are more interpretable than individual units. To gain more insight into this discrepancy, we perform a similar analysis of disentanglement as the original paper. In the original paper, they trained 400 β-variational autoencoder (β-VAE) models (Higgins et al., 2016) with different seeds and hyper-parameters that, empirically, find interpretable factorizations of the data (although see Hyvärinen & Pajunen, 1999; Locatello et al., 2019). They used an unsupervised metric of *disentanglement* (Duan et al., 2019) to check if more disentangled models have a better one-to-one correspondence with IT neurons, and found a positive relationship.[5] We find the same positive relationship in Fig. 9 (top) for both neurons (left) and K-Means features (right).

In the middle of Fig. 9, we report the distribution over different disentanglement metrics for neurons and features. Surprisingly, we find that the *features* tend to achieve higher scores (across the 400 model instances), suggesting that they are more disentangled than individual neurons. This

[5]Note that the same desideratum of having a sparse readout that links model features with neural responses leads to the identification of functional cell types in neural system identification (Klindt et al., 2017; Ustyuzhaninov et al., 2019).

finding is particularly interesting because *disentanglement* and *interpretability* are logically separable concepts—the former can be mathematically formalized as *source recovery* in (non-)linear ICA (Hyvarinen & Morioka, 2016; 2017; Klindt et al., 2020; Hyvärinen et al., 2023), while the latter is a complex function of human semantics. However, based on these results, we hypothesize that our measures of interpretablity are in fact related to classical notions of disentanglement or *source recovery*. Further supporting this idea, in the bottom of Fig. 9, we find a strong relationship between a supervised measure of disentanglement (i.e., the maximal absolute correlation between a neuron/feature and the model units, for the model with the highest UDR score) and our interpretability score (in terms of psychophysics logits).[6]

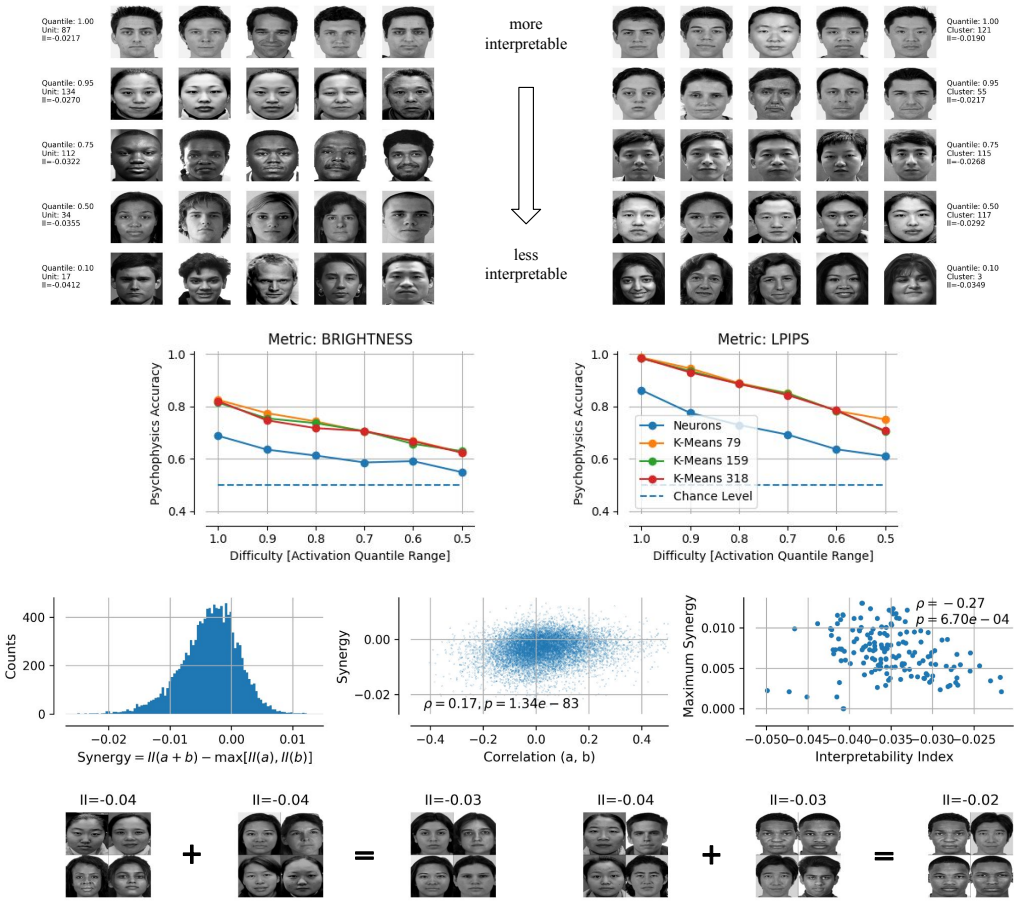

Figure 8: **Interpretability in IT Cortex. 1st row)** MEIs for neurons (left) and directions (right) as retrieved with K-Means (see Fig. 4). **2nd row)** Accuracy across neurons and interpretable directions revealed by K-Means clusters for *in silico* psychophysics task for different levels of difficult (see Fig. 5). **3rd row)** (left to right, see Fig. 6) Histogram of synergies for every pair of neurons; Relationship between the correlation and synergy over all pairs of neurons; Relationship between the II of a neuron and its maximal synergy. **4th row)** Example pairs (highest synergies) of neurons and the result when adding them (see Fig. 6).

[6]We could not use the same disentanglement metrics above since those are across models, while here, we report scores across neurons/features.

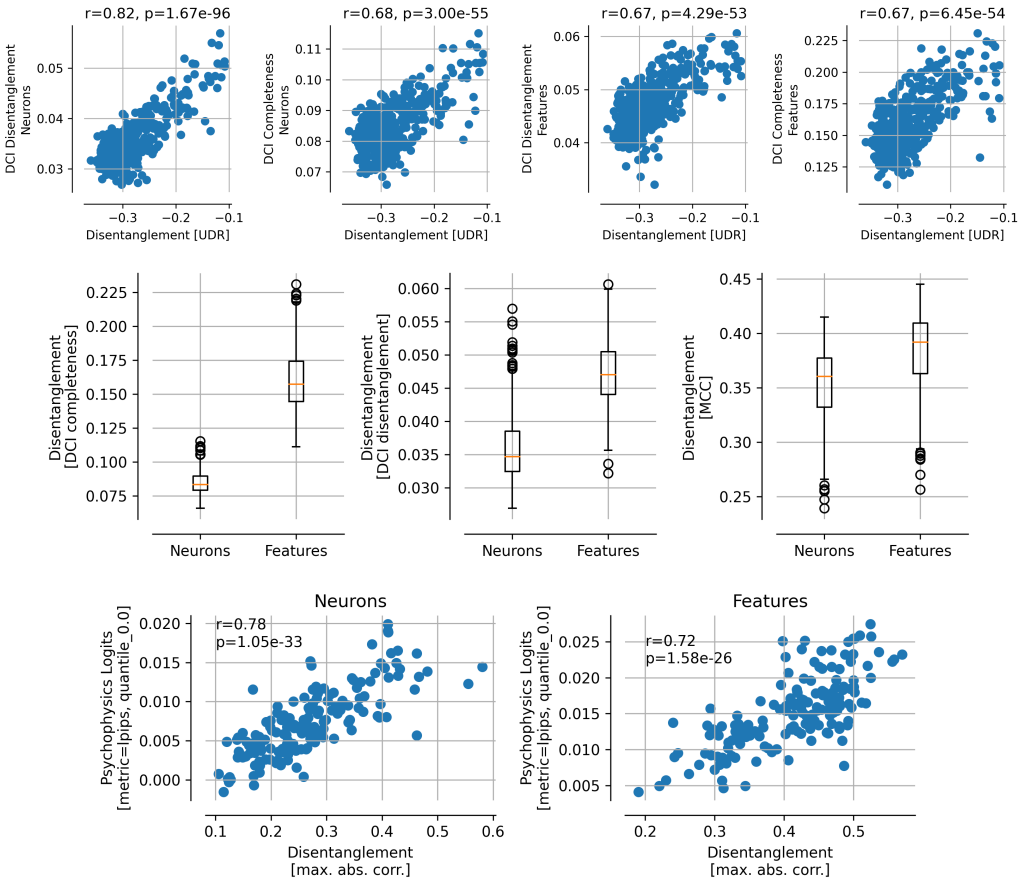

Figure 9: **Disentanglement and Interpretability in IT Cortex. Top)** Relation between *unsupervised* disentanglement (Duan et al., 2019) of 400 models (β-VAE) and *supervised* (models vs. neurons / features) disentanglement of neurons and features (Higgins et al., 2021). **Center)** Different measures of disentanglement (*DCI* (Eastwood & Williams, 2018) and *MCC* (Hyvarinen & Morioka, 2016; 2017)) for neurons and features. *DCI Disentanglement* corresponds to the metric named *alignment* in the original paper (Eastwood & Williams, 2018; Higgins et al., 2021). **Bottom)** Relationship between interpretability (psychophysics logits for LPIPS metric and full quantile range, i.e., $x = 1.0$ in Fig. 8, 2$^{\text{nd}}$ row, left) and disentanglement (see text).

The results of these analyses are interesting for two reasons. First, they provide an alternative interpretation of the original data: Neurons may sometimes align with disentangled factors of the data, however, activity clusters that involve multiple neurons tend to be even better aligned with disentangled models. Second, we find that our measures of interpretability (here *in silico* psychophysics accuracy) is strongly related to the more mathematically-grounded concept of disentanglement (Hyvärinen et al., 2023).

3.5.3 Universality and Representational Drift

A key finding in interpretability research is the phenomenon of *convergent learning* Li et al. (2015)—the observation that diverse neural network architectures

universality hypothesis Olah et al. (2020) The third neuroscience dataset that we analyze is [ref, Fran, details?]...

This dataset is interesting because it allows us to study representations across brain areas, across time and across subjects.

We find Fig. 10 (top), again, that for all brain areas (subject one) the clusters achieve a higher psychophyscis accurcay, suggesting that they are more interpretable than the individual voxels [tbd]. Moreover, we can see a tendency for lower areas (e.g., V1v) to have a larger gain for the low level

color metric, whereas higher areas (e.g., FBA-2, FFA-2) see a larger gain for the high level label metric. In Fig. 10 (bottom), we observe that clusters are systematically more transferable across brain areas (left) and across subjects (right). This corresponds to the findings of Vinken et al. (2023) about the *universality* of interpretable features. That is, more interpretable features are more likely to transfer across different representations.

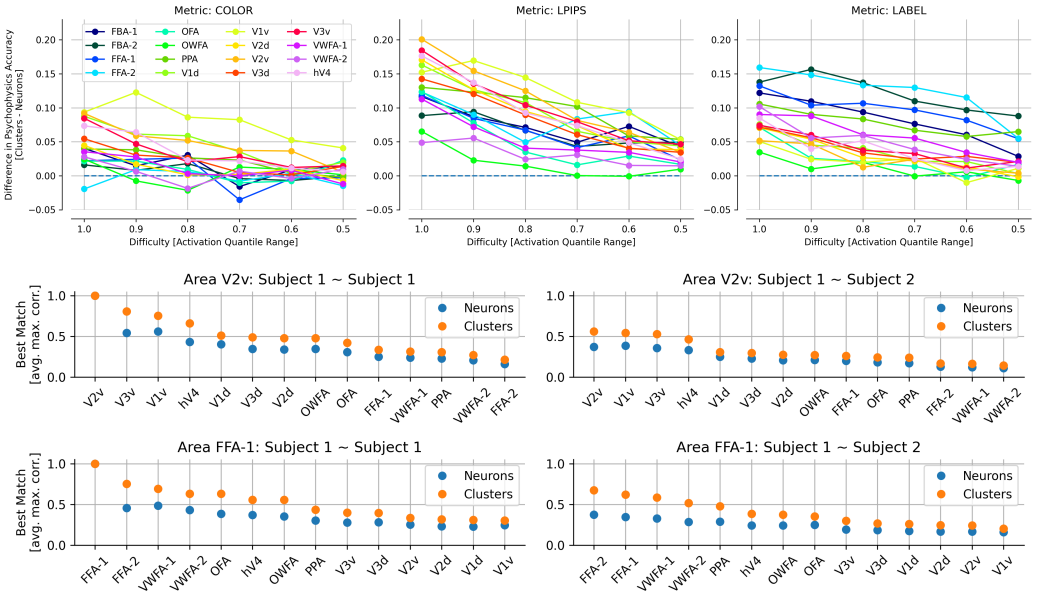

Figure 10: **Universality and Representational Drift across Brains. Top)** Differences in psychophysics accuracy between clusters and neurons for different difficulties (horizontal axis), metrics (columns) and brain areas (colours). A value above zero indicates that the clusters achieve a higher psychophysics accurcay, i.e., are more interpretable. **Bottom, left)** Best matching units (maximal correlation between activations) for two exemplary areas (V2v, FFA-1) in comparison with all other areas for both neurons and clusters, within the same subject—testing representational similarity across brain areas. **Bottom, right)** Same as bottom left but across subjects—testing representational universality across brains (Vinken et al., 2023).

4 Discussion

In this work, we have proposed a quantitative metric of *interpretability* and a method for finding interpretable features in activation space. We hope that further research will find better metrics and better feature identification methods. Nevertheless, we believe that our initial combination of metric and feature recovery method used here demonstrates the viability of our framework for automating interpretability research for vision models and visual cortex. In particular, we emphasize the value of validating quantitative metrics of interpretability against large-scale human psychophysics experiments of interpretability (Zimmermann et al., 2023). This allows us to scale human intuition to large-scale, complex neural network models—thus automating what is ordinarily done in mechanistic interpretability research by hand (Leavitt & Morcos, 2020). We hope that this approach will ultimately lead to a better understanding of neural coding principles and cast light into the black box of deep network representations.

Shifting focus from individual neurons to populations has been an important development in neuroscience (Averbeck et al., 2006; Stanley, 2013; Hebb, 2005; Gao & Ganguli, 2015; Jacobs et al., 2009; Ebitz & Hayden, 2021). In fact, *mixed selectivity* is widely observed in neuroscience, (Yoshida & Mori, 2007; Rigotti et al., 2013) and there are coding advantages believed to be conferred by such a representation (Fusi et al., 2016)—for instance, in the case of *representational drift*, a phenomenon observed in cortex in which neurons change their tuning over time while maintaining a stable representation as a population (Rule et al., 2019; Driscoll et al., 2022; Masset et al., 2022). We also tested a recent neural coding hypothesis that combines sparse coding with disentanglement in the framework of the *sparse manifold transform* (Chen et al., 2018). In App. C we find support for the notion

that interpretable features are more sparsely localized on the data manifold. Moreover, this theoretical framework could help further elucidate the link between interpretability (of discrete clusters) and disentanglement that we found in neural data (Higgins et al., 2021). Lastly, such a code may be more robust to input perturbations (Morcos et al., 2018), as suggested by our sensitivity analysis (App. D) (but see Barak et al., 2013; Johnston et al., 2020; Fusi et al., 2016). In App. F, we show that network activations follow the same spectral power law as cortical representations (Stringer et al., 2019). That is, they are low-dimensional enough to maintain differentiability (i.e. they are robust to input perturbations), while being high-dimensional enough to capture the data structure. This suggests a *universal* coding strategy employed by biological and artifical neural networks alike. We believe that future analyses grounded in a quantified metric of interpretability may illuminate the computational function of these convergent neural coding strategies.

ACKNOWLEDGEMENTS

We would like to thank Roland Zimmermann and Wieland Brendel for discussions, experiments with metrics and for sharing their psychophysics data. Moreover, thanks to Vinken et al. (2023) and Higgins et al. (2021) for publicly sharing their data and to Le Chang for helping with the extraction. We would also like to thank Katrin Franke and Andreas Tolias for discussions and feedback on the manuscript. This work was supported by the U.S. Department of Energy, under DOE Contract No. DE-AC02-76SF00515, the SLAC National Accelerator Laboratory LDRD program, and the National Science Foundation under Grant 2313150. Finally thanks to the complete Geometric Intelligence Lab at UCSB for providing feedback and support for this work.

REFERENCES

Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366, 2006.

Omri Barak, Mattia Rigotti, and Stefano Fusi. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *Journal of Neuroscience*, 33(9):3844–3856, 2013.

Horace B Barlow. Summation and inhibition in the frog's retina. *The Journal of physiology*, 119(1): 69, 1953.

Horace B Barlow. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1(4):371–394, 1972.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2023.

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. *arXiv preprint arXiv:2010.12606*, 2020.

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.

Yubei Chen, Dylan Paiton, and Bruno Olshausen. The sparse manifold transform. *Advances in neural information processing systems*, 31, 2018.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.

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.

Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.

Laura N Driscoll, Lea Duncker, and Christopher D Harvey. Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, 76:102609, 2022.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:1905.12614*, 2019.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

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, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 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.

Stanley Finger. *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, 2001.

Jonathan Frankle, David J Schwab, and Ari S Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152*, 2020.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, 37:66–74, 2016.

Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

Samuel J Gershman. What have we learned about artificial intelligence from studying the brain?, 2023.

Charles G Gross, CE de Rocha-Miranda, and DB Bender. Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of neurophysiology*, 35(1):96–111, 1972.

Chong Guo, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James Dicarlo. Adversarially trained neural representations are already as robust as biological neural representations. In *International Conference on Machine Learning*, pp. 8072–8081. PMLR, 2022.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456, 2021.

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.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Chou P Hung, Gabriel Kreiman, Tomaso Poggio, and James J DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *arXiv preprint arXiv:2302.02672*, 2023.

Adam L Jacobs, Gene Fridman, Robert M Douglas, Nazia M Alam, Peter E Latham, Glen T Prusky, and Sheila Nirenberg. Ruling out and ruling in neural codes. *Proceedings of the National Academy of Sciences*, 106(14):5936–5941, 2009.

W Jeffrey Johnston, Stephanie E Palmer, and David J Freedman. Nonlinear mixed selectivity supports reliable neural computation. *PLoS computational biology*, 16(2):e1007544, 2020.

DA Klindt, AS Ecker, T Euler, and M Bethge. Neural system identification for large 579 populations separating "what" and "where.". *Advances in Neural Information Processing 580 Systems*, 2017.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009.

Simon B Laughlin and Terrence J Sejnowski. Communication in neuronal networks. *Science*, 301 (5641):1870–1874, 2003.

Matthew L Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Paul Masset, Shanshan Qin, and Jacob A Zavatone-Veth. Drifting neuronal representations: Bug or feature? *Biological cybernetics*, 116(3):253–266, 2022.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 55–76, 2019.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 5(4):e00024–002, 2020.

Dylan M Paiton, Charles G Frye, Sheng Y Lundquist, Joel D Bowen, Ryan Zarcone, and Bruno A Olshausen. Selectivity and robustness of sparse coding networks. *Journal of vision*, 20(12): 10–10, 2020.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 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.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497 (7451):585–590, 2013.

Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current opinion in neurobiology*, 58:141–147, 2019.

Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders. `https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition#Method_1__The_presence_of_dead_neurons`, 2022. [Online; accessed 26-Sept-2023].

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Garrett B Stanley. Reading and writing the neural code. *Nature neuroscience*, 16(3):259–263, 2013.

Pirmin Stekeler-Weithofer. *Grundprobleme der Logik: Elemente einer Kritik der formalen Vernunft*. Walter de Gruyter, 2012.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

Ivan Ustyuzhaninov, Santiago A Cadena, Emmanouil Froudarakis, Paul G Fahey, Edgar Y Walker, Erick Cobos, Jacob Reimer, Fabian H Sinz, Andreas S Tolias, Matthias Bethge, et al. Rotation-invariant clustering of neuronal responses in primary visual cortex. In *International Conference on Learning Representations*, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Kasper Vinken, Jacob S Prince, Talia Konkle, and Margaret S Livingstone. The neural code for "face cells" is not face-specific. *Science Advances*, 9(35):eadg1736, 2023.

Konstantin F Willeke, Kelli Restivo, Katrin Franke, Arne F Nix, Santiago A Cadena, Tori Shinn, Cate Nealley, Gabby Rodriguez, Saumil Patel, Alexander S Ecker, et al. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. *bioRxiv*, pp. 2023–05, 2023.

Ludwig Wittgenstein. Philosophische untersuchungen. *Frankfurt: Suhrkamp*, 1953.

Ikue Yoshida and Kensaku Mori. Odorant category profile selectivity of olfactory cortex neurons. *Journal of Neuroscience*, 27(34):9105–9114, 2007.

Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

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. *arXiv preprint arXiv:2307.05471*, 2023.

APPENDIX

A TASK EXPLANATION

Here is an illustration of the task used by Zimmermann et al. (2023).

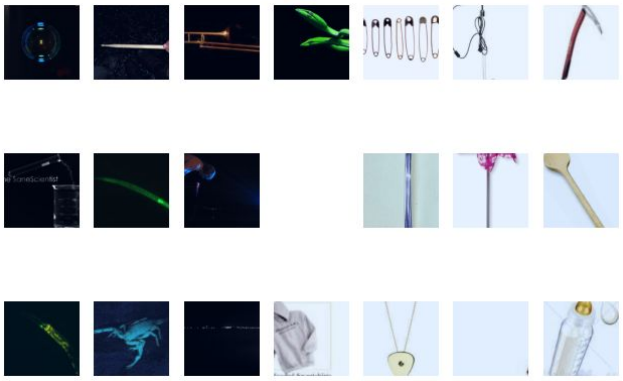

Figure 11: **Psychophysics Task.** Left images indicate positive reference ($MEI(y)$), right images indicate negative reference ($MEI(-y)$). The center images are the querries, the participant, here has to select the top image.

B COMPARISON WITH A SPARSE AUTOENCODER

We compare the K-Means approach for identifying interpretable directions in activation space to the shallow sparse autoencoder used in Sharkey et al. (2022). The model is a single-layer autoencoder trained with an L1 penalty on its hidden layer activation. We are training it for different numbers of hidden dimension and different values of sparsity regularisation. All training is for 200 epochs on the complete training set activations ($N = 50,000$) with the Adam optimizer and a learning rate of 10^{-3}. We verified manually that these settings lead to convergence for all hyperparameter settings.

We see in Fig. 12 that the directions identified by the sparse autoencoder are more interpretable according to our metric than the original neuron basis. However, we find that the K-Means approach performs better than the sparse autoencoder. In Sharkey et al. (2022), the authors assessed the relationship between source recovery and sparsity, using synthetic data containing known features. Here, we perform the same analysis, but with the II as a proxy for ground truth feature recovery. The functional relationship that we obtain between sparsity and II is remarkably similar Fig. 12 a&b, which suggests that the II may provide a good proxy for ground truth feature recovery. This also aligns with our conclusions from Fig. 9.

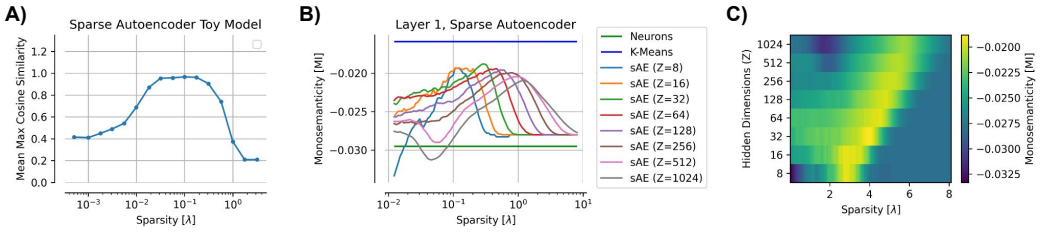

Figure 12: **Model Comparison with Sparse Autoencoder. A)** Sparse autoencoder results reproduced from (Sharkey et al., 2022). **B), C)** We train the same sparse autoencoder model and measure the II index for different number of hidden dimensions (Z) and L_1 sparsity penalties (λ).

C Interpretable Neurons as Sparse Activations on the Data Manifold

One way of understanding interpretability is in terms of the distribution of a given neuron's activations over the image manifold. This concept and the following analysis are directly inspired by the *sparse manifold transform* (Chen et al., 2018). We take the top $M = 5$ MEIs for both neurons ($N = 256$ and K-Means ($K = 256$) features and compute all pairwise image similarities using LPIPS. We then embed this distance matrix into a 2D space for visualization purposes using t-SNE (Van der Maaten & Hinton, 2008) (perplexity$= 10$) (Fig. 13). Each point in the visualization corresponds to a different image, and is colored according to a different scheme in each subplot. In Fig. 13:A, the color of each point indicates the average color of the image. In Fig. 13:B, color indicates the image label. In Figs. 13:E and 13:F, color indicates the activation of a single neuron (E) or K-Means feature (F) over the dataset. Activations for both neurons and K-Means features are computed as follows:

$$f_i(x) = e^{-\frac{d'}{\tau}}, \quad d' = \frac{d - \text{avg}(d)}{\text{s.d.}(d)} d = \|x - \mu_i\|_2^2 \tag{5}$$

Where μ_i is the location of the cluster centroids for K-Means features and a one hot vector for neurons. Taking the z-score before the exponential ensures a fair comparison. Finally, the temperature $\tau = 2$ is introduced for visualization purposes.[7]

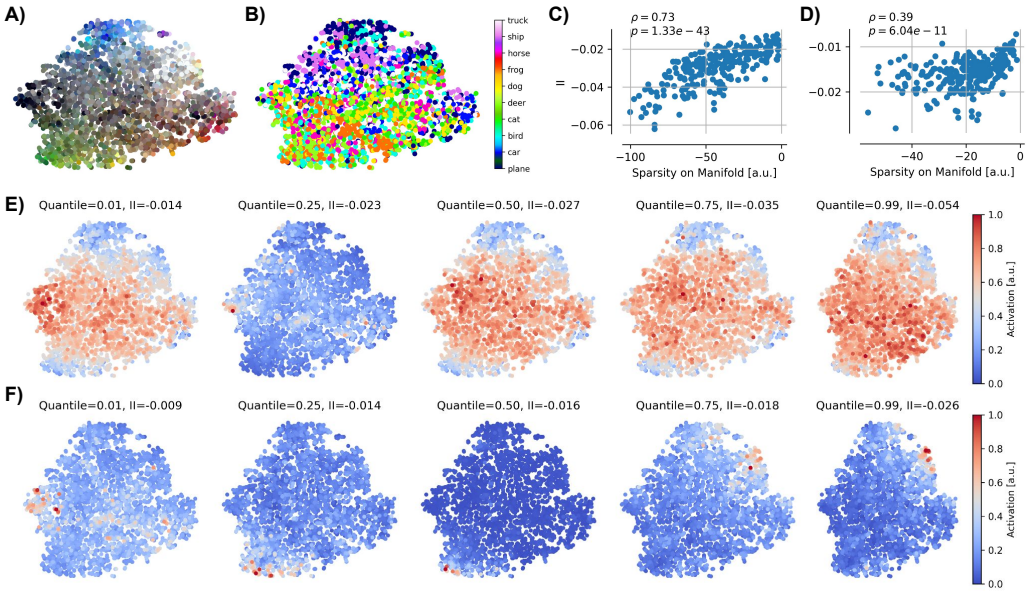

Figure 13: **Sparse Manifold Activations.** The natural image manifold (subset of MEIs) embedded in 2D with tSNE and coloured by average image colour **A)**, image label **B)**, activation of neurons **E)** and activation of features **F).** There is a correlation between sparsity on the manifold (average distance of most activating points) and the II for both neurons **C)** and features **D)**.

We see that colour (Fig. 13:A) is a major factor in determining the layout of the manifold, and although labels tend to cluster locally, image category plays a lesser role (Fig. 13:B). Interestingly, we see that the features (Fig. 13:F) are much sparser on the manifold than the neural activations (Fig. 13:E). This suggests that more interpretable units are more sparsely active across the natural image manifold. Thus, we provide evidence (Fig. 13:C&D) for a long-standing hypothesis in the neural coding literature (Chen et al., 2018).

[7]Lower τ would make Fig. 13:F look even sparser, and higher τ would make Fig. 13:E look even more uniform; using the same τ ensures a fair comparison.

D Robustness of Interpretable Features

D.1 Sensitivity Analysis

We investigate the sensitivity of the interpretable directions (directions), i.e., of the K-Means centroids. Specifically, we perturb each direction and quantify whether the perturbed direction is still interpretable. Our perturbation process is explained below. We interpolate from one K-Means centroid μ_a to another μ_b (and beyond to test extrapolation) and compute the II for these different directions in latent space. The intermediate directions are:

$$v(\alpha) = \alpha\mu_a + (1-\alpha)\mu_b, \quad \text{for } \alpha \in \mathbb{R}. \tag{6}$$

We normalize each intermediate direction $v(\alpha)$ to maintain the same norm 1. For each direction v along the interpolation path we compute the II index from the images that lie closest to that point in latent space:

$$f_v(x) = -\|y(x) - v\|, \tag{7}$$

where x is an image (an input), y is the function representing the first layer of the CNN, $y(x)$ is the feature associated to image x in latent space, and $\|\|$ is the Euclidean norm in latent space. The results are shown in Figure 14 B). We observe that μ_a and μ_b, corresponding to interpolation factors $\alpha = 0$ and $\alpha = 1$ have higher II and that the II strongly drops for interpolating directions $v(\alpha)$. This supports the idea that the direction extracted via K-Means are interpretable. Moreover, based on signal detection theory (Dayan & Abbott, 2005), we hypothesize that more meaningful directions are so in virtue of being highly *selective* and less *sensitive* to input perturbations, i.e., to image perturbations Paiton et al. (2020). Thus, for each intermediate direction v, we also compute the norm of the input gradient:

$$\|\nabla_{f_v}|_x\| = \|\nabla[-\|y(x) - v\|]\| = \|\nabla[-\|y(x) - \alpha\mu_a - (1-\alpha)\mu_b\|]\|. \tag{8}$$

For a fixed intermediate direction v, this value quantifies how much the response of a feature $y(x)$ representing image x changes given perturbations on image x.

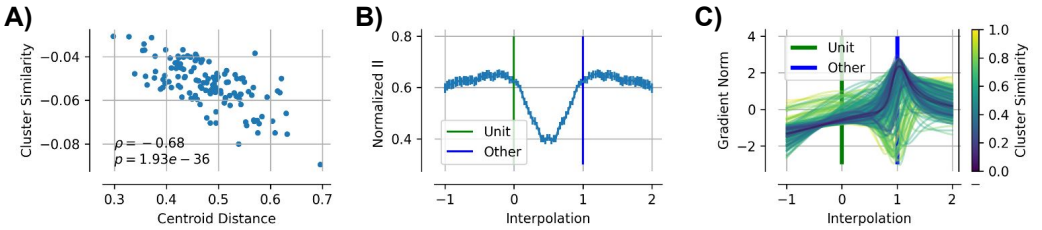

Figure 14: **Sensitivity Analysis. A)** Clusters that are further away from each other are have lower semantic similarity (measured as cross-II). **B)** Computing the II for interpolations (from one centroid to another) between and beyond all pairs of maximally separated clusters (determined with Hungarian algorithm). **C)** Sensitivity (i.e., average norm of the input gradient) for different interpolation points, coloured by the cluster similarity of start and end point in interpolation.

Figure 14 shows the average norm of the input gradient. We observe that the gradient's norm is generally much lower at a neuron's MEI ($\alpha = 0$) vs. the MEI of a different unit ($\alpha = 0$), unless they are very similar. We also find a weak but significant negative correlation (Spearman $\rho = -0.18$ $p < 3.110^{-5}$ between the interpretability and the minimal gradient norm, as shown in Figure 14 D). Together, these suggests that units which are more interpretable are also less sensitive to input perturbations at their preferred inputs. Consequently, a hypothesis derived from these analyses: neurons in CNNs that are more interpretable are also more robust to adversarial or noise perturbations.

D.2 Noise Robustness

To test the hypothesis that more interpretable neurons are more noise robust, we add Gaussian noise to the inputs (standard deviation $\sigma \in [0, 0.1]$) and measure the sensitivity, i.e., the maximum absolute change in response compared to the clean image as proposed by (Guo et al., 2022). We then compare

those scores to the II metric from the paper, as well as the logits from the *in silico* psychophysics task. We find a weak but significant relationship in both cases. This supports the hypothesis that more interpretable neurons are more robust to white noise input perturbations.

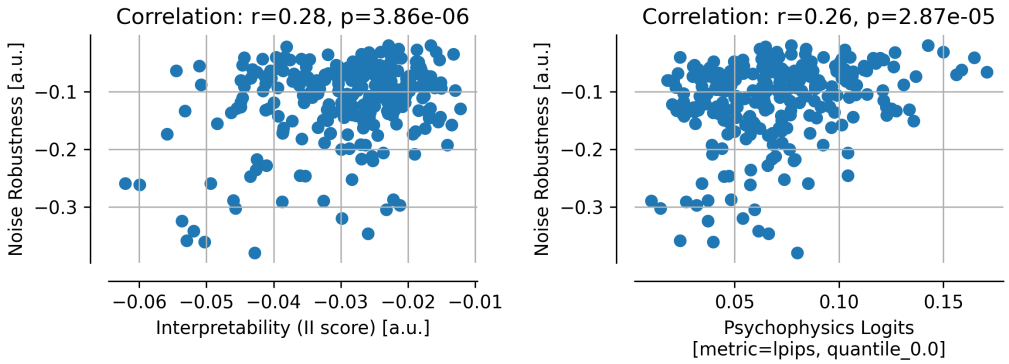

Figure 15: **Noise Robustness of Neurons.** Left, noise robustness (Guo et al., 2022) of neurons, plotted as a function of their II scores (see paper); linear correlation indicated above. Right, same but as a function of psychophysics logits (for largest quantile and LPIPS metric, see main text).

E Monosemanticity and the Privileged Basis Hypothesis

Why should we see individual neurons learning meaningful representations at all? Recent research in the mechanistic interpretability literature has suggested that there exists *privileged bases* in neural networks, corresponding to neurons, emerging from nonlinearities such as ReLU that operate per neuron. The intuition behind the privileged basis needs further explanation.

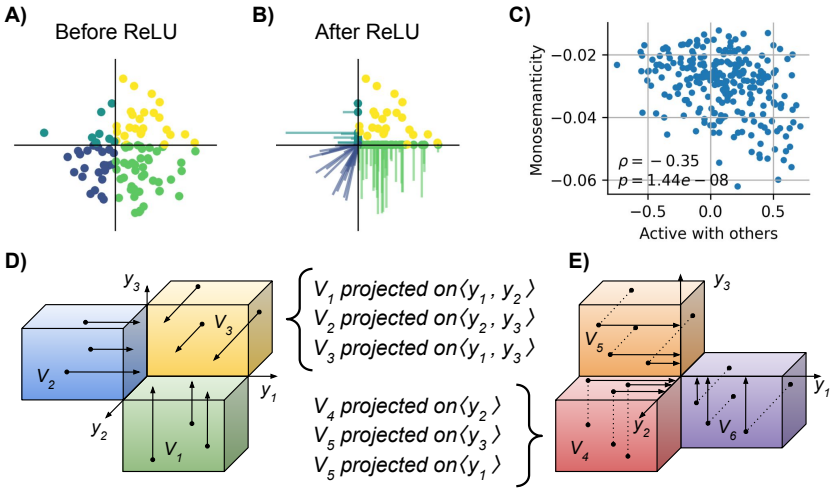

Figure 16: **Privileged Basis Hypothesis.** Activations before **A)** and **B)** after ReLU nonlinearity. In this scenario, the first neuron is more active as it has to represent the large number of (green) points in the bottom right quadrant, while the second neuron only needs to represent few (blue) points in the top left quadrant. **C)** Significant negative correlation between neurons that are more active together with others (measured as the correlation between the neuron's activity and the population, excluding the neuron, average) and the interpretability index (II). **D** In a three dimensional space, the positive only quadrant remains untouched by ReLUs, the negative only quadrant gets mapped to 0, three quadrants (V_1, V_2, V_3, **D**) are projected onto two dimensional subspaces, and three quadrants (V_4, V_5, V_6, **D**) are projected onto one dimensional subspaces. Thus, e.g. neuron y_1 has to represent (together with y_2) all of V_1, and it also has to represent (completely on its own) all of V_6.

For K neurons, there are 2^K quadrants as shown in Figure 16 A) for $K = 2$. We consider what happens for a feature vector in each of these quadrant after the application of ReLU. Of those, 1 (all positive) stays untouched, as shown in yellow in Figure 16 B. Another 1 quadrant (all negative) becomes zero: shown in purple. Next, K quadrants get represented by 1 neuron (i.e. K dimensions are collapsed to 1) shown in blue and green; $K - 1$ get represented by 2 neurons (i.e. K dimensions are collapsed to 1), etc. In other words, each neuron participates in encoding points in $K!$ quadrants of compressed dimensionality. Now, we ask: which of these neurons should be more interpretable?

Consider data points unequally distributed into the different quadrants, with one quadrant (bottom right in Figure 16 A)) containing more points than another (top left in Figure 16 A)). Neurons in charge of representing the features from a quadrant with many point is more active. A resulting hypothesis is that neurons which are more active when others are inactive, i.e., which are active alone, are more interpretable — and form the priviledged basis.

Figure 16 C) uses our our interpretability index (II) to confirm this hypothesis by showing a negative correlation between a co-activation measure ("Active with Others") and II. The co-activation measure is defined as the correlation between each neuron's response and the average population response.

F Dimensionality of the Neural Activation Manifold

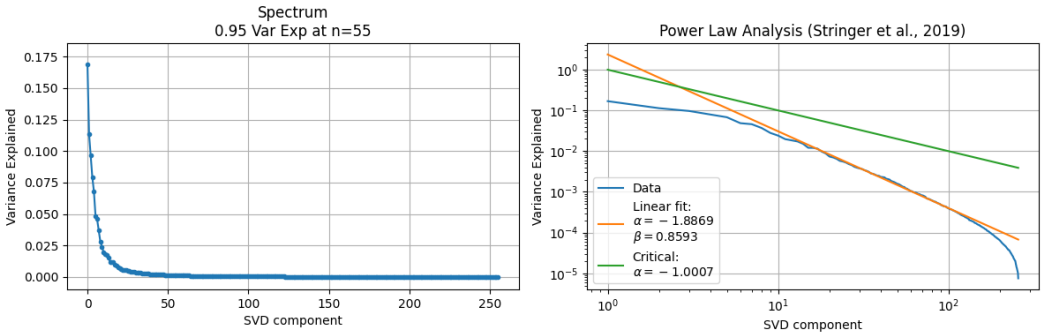

Figure 17: **High Dimensional Smooth Activation Manifold.** Same analysis as in Stringer et al. (2019), showing that activations in CNN feature space are high dimensional within the constraints of remaining differentiable. A spectrum that decays slower than the critical value (green line in right plot, Stringer et al. (2019)), would be non-differentiable and, therefore, highly non-robust. Remarkably, this spectral behaviour is the same as observed across many cortical areas.

G Additional Comparisons to Psychophysics Data

Here we look at some of the other exemplary models studied by Zimmermann et al. (2023).

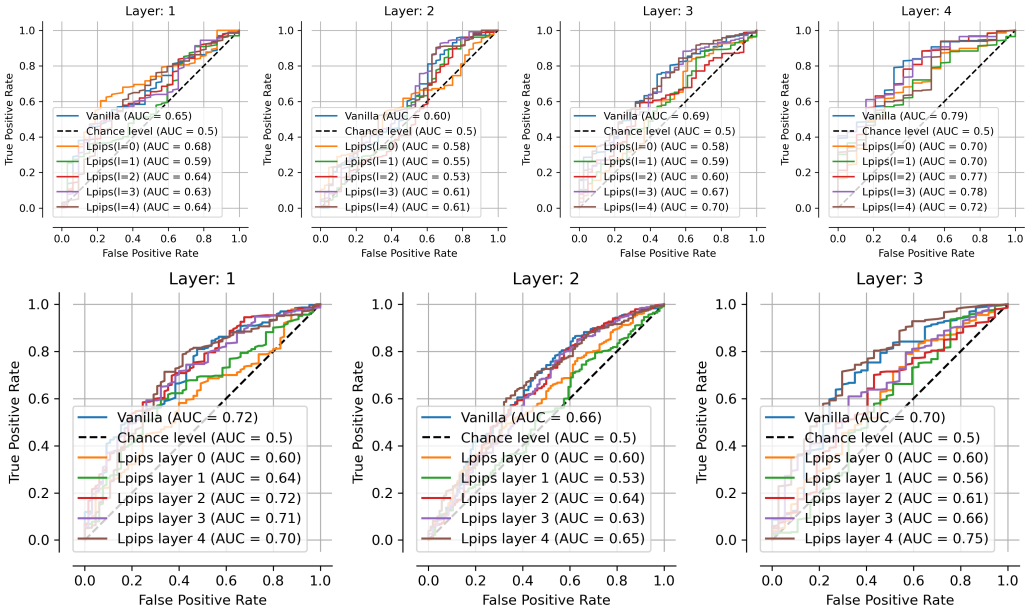

Figure 18: **Psychophysics Match for Other Models.** Data from Zimmermann et al. (2023). Top, Clip ResNet50; Bottom, GoogLeNet. Left to Right: Agreement between human and in-silico psychophysics on the predictability of the outputs of different layers in the network. Human and model agree on what makes a feature predictability for the early layers. For these layers, the proposed interpretability metric is a valid representation of the human's perception of interpretability. AUC: Area Under the Curve.

H DISENTANGLING MODEL AND DATA

Like Bricken et al. (Bricken et al., 2023), we run our analyses again on an untrained model to assess how much of the interpretability results we obtained are due to training (model) versus derived from properties of the data. Interestingly, we find that, indeed, the gap in interpretability between neurons and activity clusters is already present at initialization. This confirms prior experiments that showed that CNNs are, even untrained, already very useful representations of image data (Frankle et al., 2020). However, we note a clear training effect: the semantic level of interpretability shifts over the course of training. That is, while the untrained model neurons and features are easily predicted by low level image properties, such as color, the trained model is better predicted by high level semantics, such as label information.

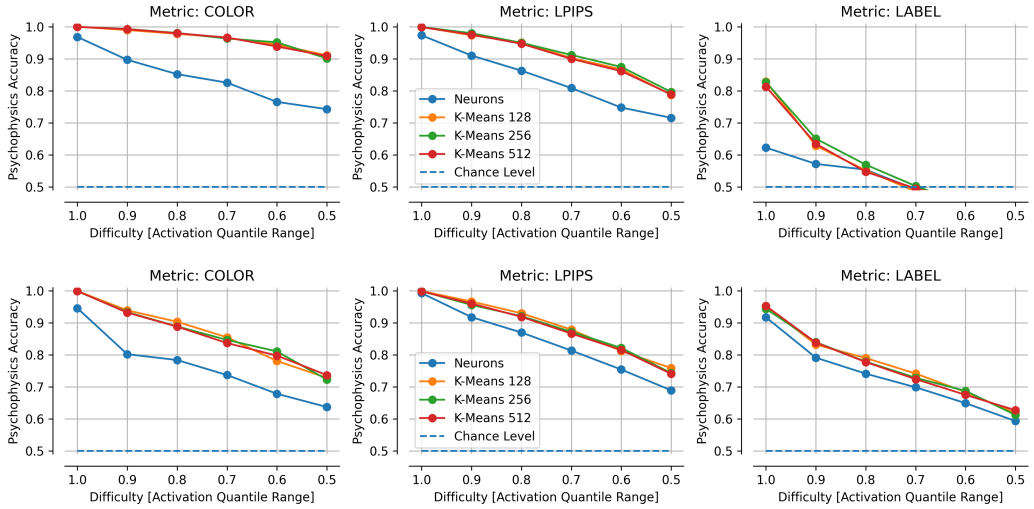

Figure 19: **Untrained Psychophysics.** Top, untrained model; Bottom, trained model, same as Fig.5. The interpretability gap is already apparent at initialisation, however, there is a clear shift in semantics from low level (color) at initialisation, to high level (label) after training.