DORA: Exploring Outlier Representations in Deep Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Although Deep Neural Networks (DNNs) are incredibly effective in learning complex abstractions, they are susceptible to unintentionally learning spurious artifacts from the training data. To ensure model transparency, it is crucial to examine the relationships between learned representations, as unintended concepts often manifest themselves to be anomalous to the desired task. In this work, we introduce DORA (Data-agnOstic Representation Analysis): the first *data-agnostic* framework for the analysis of the representation space of DNNs. Our framework employs the proposed *Extreme-Activation* (EA) distance measure between representations that utilizes self-explaining capabilities within the network without accessing any data. We quantitatively validate the metric's correctness and alignment with human-defined semantic distances. The coherence between the EA distance and human judgment enables us to identify representations whose underlying concepts would be considered unnatural by humans by identifying outliers in functional distance. Finally, we demonstrate the practical usefulness of DORA by analyzing and identifying artifact representations in popular Computer Vision models.

1 Introduction

The ability of Deep Neural Networks (DNNs) to perform complex tasks and achieve *state-of-the-art* performance in various fields can be attributed to the rich and hierarchical representations that they learn Bengio et al. (2013). Far beyond the handcrafted features that were inductively constructed by humans on learning machines in classical Machine Learning methods Marr and Nishihara (1978); Jackson; Fogel and Sagi (1989), Deep Learning approaches exploit the network's freedom for representation learning, which, however, leads to a semantic opacity of learned abstractions. The rapid progress in representation learning only exacerbates the issue of interpretability, as modern DNNs are often trained in a self-supervised manner Jaiswal et al. (2020); LeCun and Misra (2021) and from a potentially limitless amount of data Brown et al. (2020); Bommasani et al. (2021), alleviating human control over the training dataset, and resulting in opaque decision-making strategies.

The increasing popularity of Deep Learning techniques across various fields, coupled with the difficulty of interpreting the decision-making processes of complex models, has led to the emergence of the field of Explainable AI (XAI) (e.g. Montavon et al. (2018); Samek et al. (2019); Xu et al. (2019); Gade et al. (2019); Rudin (2019); Samek et al. (2021)). Research within XAI has revealed that the internal representations that form the basis of DNNs are susceptible to learning harmful and undesired concepts, such as biases Guidotti et al. (2018); Jiang and Nachum (2020), Clever Hans (CH) effects Lapuschkin et al. (2019), and backdoors Anders et al. (2022). The learned artifactual concepts are often unnaturally and semantically distant from the relevant concepts within the dataset, such as watermarks in the PASCAL 2012 image classification task Lapuschkin et al. (2019), Chinese logographic watermarks in ImageNet dataset Li et al. (2022), colored band-aids in skin-cancer detection problem Anders et al. (2022) or tokens in a pneumonia detection problem Zech et al. (2018).

In order to enhance our understanding of the decision-making processes of complex machines and prevent biased or harmful decisions, it is critical to provide an explanation of the representations that are learned by the model. One approach to gaining insights into a model's prediction strategies is to analyze the relationships among its learned representations, which can be quantified using a funtional distance metric. It is important that this metric aligns with human judgment such that the distance between representations reflects the concepts that are learned and is coherent with human perception of the distance between such concepts. This property enables the us to introduce a novel problem of identification of semantically anomalous representations within the network. By assuming that the functional distance metric between representations is aligned with human decision-making, we can use the proposed distance measure to perform an outlier detection analysis in the functional space and identify representations whose concepts are semantically anomalous to the majority and pontentially undesired for the given task.

In this work, we propose *DORA*[*] — the first data-agnostic framework allowing an automatic inspection of the representation space of Deep Neural Networks. DORA leverages the proposed *Extreme-Activation* method that exploits the self-explanation capabilities of the networks and estimates distances between representations, regardless of the availability of the specific data used for training. DORA facilitates the understanding of the associations between neural representations and the visualization of the representation space via *representation atlases.* By assuming that artificial representations, which deviate from the desired decision-making policy, are semantically distant from the relevant representations learned by the network, DORA allows the detection of potentially harmful representations that may lead to unintended learning outcomes. Additionally, DORA can be further used to identify and remove infected data points.

The main contributions of this research are:

- We introduce the Extreme-Activation distance metric for representations in both data-aware and data-agnostic scenarios.
- We propose the data-agnostic DORA framework for analyzing and visualizing the representation space of Deep Neural Networks (DNNs).
- We quantitatively assess the alignment of the proposed distance metrics with human judgement across several semantic baselines and compare them to standard distance measures in controlled scenarios.
- We quantitatively evaluate the ability of various distance metrics to detect semantically anomalous representations in controlled scenarios.
- We demonstrate the applicability of DORA on popular Computer Vision models and demonstrate that in real-world applications, outlier representations may encode undesirable and harmful concepts.

2 Related Work

To address the concerns regarding the black-box nature of complex learning machines Baehrens et al. (2010); Vidovic et al. (2015); Buhrmester et al. (2019); Samek et al. (2021), the field of *Explainable AI (XAI)* has emerged. While some recent research focuses on inducing the self-explaining capabilities through changes in the architecture and the learning process Gautam et al. (2022a;b); Chen et al. (2018); Gautam et al. (2021), the majority of XAI methods (typically referred to as *post-hoc* explanation methods) are decoupled from the training procedure. A dichotomy of post-hoc explanation methods could be performed based on the notion of their explanations, i.e., the model behavior can be either explained on a *local* level, where the decision-making strategy of a system is explained for one particular input sample, or on a *global* data set level, where the aim is to explain the prediction strategy learned by the machine across the data set and investigate the purpose of its individual components in a universal fashion detached from single data points (similar to feature selection Guyon and Elisseeff (2003)).

Local explanation methods, often produce attribution maps, interpreting the prediction by attributing relevance scores to the features of the input signal, highlighting the influential characteristics that affected the prediction the most. Various methods, such as Layer-wise Relevance Propagation (LRP) Bach et al.

---

[*] PyTorch implementation of the proposed method could be found by the following link: `anonimyzed` .

(2015), GradCAM Selvaraju et al. (2019), Occlusion Zeiler and Fergus (2014), MFI Vidovic et al. (2016), Integrated Gradient Sundararajan et al. (2017), have proven effective in explaining DNNs Tjoa and Guan (2020) as well as Bayesian Neural Networks Bykov et al. (2021); Brown and Talbert (2022). To further boost the quality of interpretations, several enhancing techniques were introduced, such as SmoothGrad Smilkov et al. (2017); Omeiza et al. (2019), NoiseGrad and FusionGrad Bykov et al. (2022). Considerable attention also has been paid to analyzing and evaluating the quality of local explanation methods (e.g. Samek et al. (2016); Hedström et al. (2022); Guidotti (2021)). However, while the local explanation paradigm is incredibly powerful in transferring the understanding of the decision-making strategies for a particular data sample, it lacks the ability to provide an overall view of the inner processes of representations in a network.

Global explanation methods aim to interpret the general behavior of learning machines by investigating the role of particular components, such as neurons, channels, or output logits, which we refer to as representations. Existing methods mainly aim to connect internal representations to human understandable concepts, making the purpose and semantics of particular network sub-function transparent to humans. So far, there are already methods, such as Network Dissection Bau et al. (2017; 2018) and Compositional Explanations of Neurons Mu and Andreas (2020) that aim to label representations with class labels from a given dataset, based on the intersection between the class relevant information provided by a binary mask information and the activation map of the respective representation. In contrast, the MILAN method generates a text description of the representation by searching for a text string that maximizes the mutual information with the image regions in which the neuron is active Hernandez et al. (2021).

2.1 Activation-Maximisation methods

The family of Activation-Maximization (AM) Erhan et al. (2009) methods is designed for the global explanation of complex learning machines by identifying the input that maximally activates a particular neuron or layer in the network to visualize the features that have been learned by the neuron or layer. These signals, which we will refer to as Activation-Maximization Signals (AMS), could be either natural signals(n-AMS), found in a *data-aware* fashion by selecting a "real" example from an existing data corpus Borowski et al. (2020), or artificial (synthetic AMS or s-AMS), found in a *data-agnostic* mode by generating a synthetic input through optimization Erhan et al. (2009); Olah et al. (2017); Szegedy et al. (2013).

In comparison to earlier synthetic AM methods, Feature Visualization (FV) Olah et al. (2017) performs optimization in the frequency domain by parametrizing the image with frequencies obtained from the Fourier transformation. This reduces adversarial noise in s-AMS (e.g. Erhan et al. (2009); Szegedy et al. (2013)) — improving the interpretability of the obtained signals. Additionally, the FV method applies multiple stochastic image transformations, such as jittering, rotating, or scaling, before each optimization step, as well as frequency penalization, which either explicitly penalizes the variance between neighboring pixels or applies bilateral filters on the input.

2.2 Spurious correlations

Deep Neural Networks are prone to learn spurious representations — patterns that are correlated with a target class on the training data but not inherently relevant to the learning problem Izmailov et al. (2022). Reliance on spurious features prevents the model from generalizing, which subsequently leads to poor performance on sub-groups of the data where the spurious correlation is absent (cf. Lapuschkin et al. (2016; 2019); Geirhos et al. (2020)). In Computer Vision, such behavior could be characterized by the reliance of the model on an image's background Xiao et al. (2020), object textures Geirhos et al. (2018), or the presence of semantic artifacts in the training data Wallis and Buvat (2022); Lapuschkin et al. (2019); Geirhos et al. (2020); Anders et al. (2022). Artifacts can be added to the training data on purpose as Backdoor attacks Gu et al. (2017); Tran et al. (2018), or emerge "naturally" and might persist unnoticed in the training corpus, resulting in *Clever Hans effects* Lapuschkin et al. (2019).

Recently, XAI methods have demonstrated their potential in revealing the underlying mechanisms of predictions made by models, particularly in the presence of artifacts such as Clever Hans or Backdoor artifacts. Spectral Relevance analysis (SpRAy) aims to provide a global explanation of the model by analyzing local explanations across the dataset and clustering them for manual inspection Lapuschkin et al. (2019).

While successful in certain cases Schramowski et al. (2020), SpRAy requires a substantial amount of human supervision and may not detect artifacts that do not exhibit consistent shape and position in the original images. SpRAY-based Class Artifact Compensation Anders et al. (2022) method significantly reduced the need for human supervision and demonstrated its capability to effectively suppress the artifactual behavior of DNNs, significantly reducing a model's Clever Hans behavior.

2.3 Comparison of representations

The study of representation similarity in DNN architectures is a topic of active research. Numerous methods for comparing network representations have been applied to different architectures, including Neural Networks of varying width and depth Nguyen et al. (2020), Bayesian Neural Networks Grinwald et al. (2022), and Transformer Neural Networks Raghu et al. (2021). Some works Ramsay et al. (1984); Laakso (2000); Kornblith et al. (2019); Nguyen et al. (2022) argue that the representation similarity should be based on the correlation of a distance measure applied to layer activations on training data. Other works Raghu et al. (2017); Morcos et al. (2018) compute similarity values by applying variants of Canonical Correlation Analysis (CCA) Hardoon et al. (2005); Bießmann et al. (2010) on the activations or by calculating mutual information Li et al. (2015). However, all of these methods require the presence of training data.

3 Distance metrics between Neural Representations

The compositional structure of modern neural networks allows neurons, the foundational building blocks for neural processing, to learn complex abstractions. These abstractions, learned without any supervision by humans, lead to the semantic opaqueness of representations — the purpose of the particular representations in DNNs remains unknown to humans. To enhance transparency and gain a better understanding of information processing within a model, studying the interrelationships between neural representations in the model can yield valuable insights.

In the following, we start with the definition of a *neural representation* as a sub-function of a given network that depicts the computation graph, from the input of the model to the output of a specific neuron.

Definition 1 (Neural representation). *We define a neural representation f as a real-valued function $f : \mathbb{D} \to \mathbb{R}$, mapping from the data domain $\mathbb{D}$ to the real numbers $\mathbb{R}$.*

The necessity for representations to be univariate is maintained for the sake of simplicity regarding the explanation. Although neurons in DNNs often produce multidimensional outputs depending on the specific use cases, multidimensional functions could be regarded as a set of individual representations when analyzing the representations. Alternatively, multidimensional outputs could be aggregated without losing transparency: for instance, in the case of convolutional neurons that output activation maps containing the dot product between filter weights and input data at each location, activation maps could be aggregated for the sake of simplifying the explanation of the semantic concept underlying the function. The choice depends on the particular aim and scope of the analysis and does not alter the network itself.

The scalar output of representations often corresponds to the amount of evidence or similarity between certain concepts present in the input and internally learned abstractions. Various sub-functions within the model could be considered as neural representations, ranging from the neurons in the initial layers that are often regarded as elementary edge or color detectors Le and Kayal (2021), to the high-level feature extractors in the final layers and output classification logits. While these representations are interesting to understand, throughout this work, we primarily focus on the high-level abstractions that emerge in the latest layers of networks, such as the feature-extractor layers in well-known computer vision architectures, as they are frequently employed for transfer learning Zhuang et al. (2020).

In DNNs, neural representations are combined into layers — collections of individual neural representations that typically share the same computational architecture and learn abstractions of similar complexity. In the scope of the following work, we mainly focused on the analysis of the relations between representations within one selected layer from the network.

Definition 2 (Layer). *We define a layer $\mathcal{F} = \{f_1, ..., f_k\}$ as a set comprising k individual neural representations.*

To examine the relationships between representations, we can start by examining the behavior of functions on the given dataset. We define a dataset D consisting of N data points, denoted as $D = \{x_1, ..., x_N\}$, which we refer to as the *evaluation dataset* to measure the relationship between two neural representations. We assume that the datapoints $x_1, ..., x_N$ represent i.i.d samples from the global data distribution $\mathcal{D}$. For the analysis of the *data-aware* distance metrics, such as metrics between representations that require the availability of the data, we assume that the activations of representations are standardized on the evaluation dataset, resulting in a mean of 0 and a standard deviation of 1. Practically, this means that for each output of the neural representation, we subtract the mean across the evaluation dataset and divide it by the standard deviation.

For a neural representation f_i and an evaluation dataset D, we define an vector of activations $\mathbf{a}_i = [f_i(x_1), ..., f_i(x_N)]$, and assume that

$$\mu_i := \frac{1}{N} \sum_{t=1}^{N} f_i(x_t) = 0, \quad \sigma_i := \sqrt{\frac{1}{N} \sum_{t=1}^{N} \left(f_i(x_t) - \mu_i\right)^2} = 1. \tag{1}$$

Standardizing the vectors in this way can help to mitigate any differences in scale between the vector components and ensure that each component contributes equally to the distance calculation.

Below, we present three widely recognized metrics that can be utilized to measure the distance between neural representations. These metrics will later serve as a point of reference for comparing our own developed metric.

- **Minkowski distance**:

  $$d_M\left(f_i, f_j\right) = \left(\sum_{t=1}^{N} |f_i(x_t) - f_j(x_t)|^p\right)^{\frac{1}{p}}, \tag{2}$$

  where p determines the degree of the norm, which gauges the sensitivity of the metric to differences between the components of the vectors being compared. In general, larger values of p lead to a greater emphasis on larger differences between the components of the vectors. Conversely, smaller values of p reduce the influence of larger differences, leveraging an increased uniform weighting of all components.

- **Pearson distance**:

  $$d_P\left(f_i, f_j\right) = \frac{1}{\sqrt{2}} \sqrt{1 - \rho_p\left(\mathbf{a}_i, \mathbf{a}_j\right)}, \tag{3}$$

  where $\rho_p(\mathbf{a}, \mathbf{b})$ is the Pearson correlation coefficient between the vectors $\mathbf{a}$ and $\mathbf{b}$.

  The Pearson correlation coefficient is a widely used metric for measuring the linear dependence between two random variables. It is an interpretable measure of similarity, however, it is also sensitive to outliers, which can significantly affect the calculated distance.

- **Spearman distance**:

  $$d_S\left(f_i, f_j\right) = \frac{1}{\sqrt{2}} \sqrt{1 - \rho_s\left(\mathbf{a}_i, \mathbf{a}_j\right)}, \tag{4}$$

  where $\rho_s(\mathbf{a}, \mathbf{b})$ is the Spearman rank-correlation coefficient between vectors $\mathbf{a}$ and $\mathbf{b}$.

  The Spearman correlation is a non-parametric rank-based metric commonly used to measure the monotonic dependence between two random variables. Its main advantage is that it is robust to outliers, can handle ties in the data, and is relatively easy to interpret.

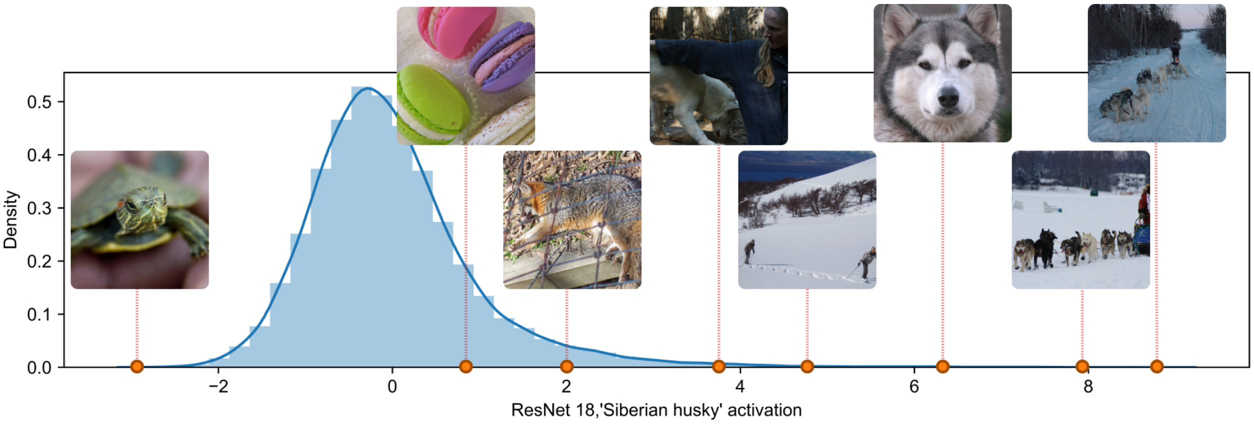


Figure 1: **Distribution of activations for the "Siberian husky" representation.** The figure presented shows the standardized activation distribution of the "Siberian Husky" logit from the ResNet18 model trained on ImageNet. The data was collected across the ImageNet-2012 validation dataset. Additionally, the plot displays various input images along with their corresponding activations. Analyzing the activation of representations can provide crucial insights into the behavior of the model. For instance, we observe that the model achieves extremely high activations when there are multiple dogs in the image, corresponding to the 'Dogsled" class. However, we also observe a potential spurious correlation, where the model assigns high scores to images with a snowy background.

3.1 Data-Aware Extreme-Activation distance

The Minkowski, Pearson, and Spearman distance metrics capture the differences in the general behavior of the representations. However, often the semantics behind representations are characterized by the signals, that extremely activate the function, such as Activation-Maximisation signals. These signals correspond to extreme positive activations and demonstrate the concepts that activate the representations. Although this type of analysis does not fully encompass the intricate nature of neural representations and only focuses on the most activating concepts (ignoring those that deactivate the representation), it is effective due to the widespread use of bounded activation functions, such as ReLU Glorot et al. (2011), where positive activation values indicate the presence of specific patterns in the signal.

Given the evaluation dataset D and a neural representation f_i, we define a collection of natural Activation-Maximisation signals (n-AMS) as follows:

Definition 3 (n-AMS)**.** *Let* f_i *be a neural representation, and* $D = \{x_1, ..., x_N\} \subset \mathbb{D}$ *be an evaluation dataset with* N *datapoints. Assume that the dataset* D *could be split in* n *disjoint blocks* $D = \bigcup_{i=1}^{n} D_t, D_t = \left\{x_{td+1}, ..., x_{(t+1)d+1}\right\}, \forall t \in \{0, ..., n-1\}$ *of length* d*.*

We define a collection of natural Activation-Maximisation signals (n-AMS) as $S_i = \left\{s_1^i, ..., s_n^i\right\}$*, where*

$$s_t^i = \arg\max_{x \in D_t} f_i\left(x\right), \forall t \in \{0, ..., n-1\}. \tag{5}$$

The collection of n-AMS is determined by two parameters: n, which denotes the number of sampled signals, and d, referred to as the *depth*, which represents the size of the subset from which the signal is obtained. Note that we could examine the highest activation signal of the whole dataset by setting $n = 1$ and $d = N$, however, interpreting the representation's semantics by using only one signal might be misleading. Figure 1 illustrates the distribution of activations of the "Siberian husky" logit from the ResNet18 model trained on ImageNet He et al. (2016) across all the images from the ImageNet-2012 Deng et al. (2009) validation dataset, where we can observe that the most activating signal corresponds to the "Dogsled" class. In light of this, we aim to sample several n-AMS from separate data subsets.

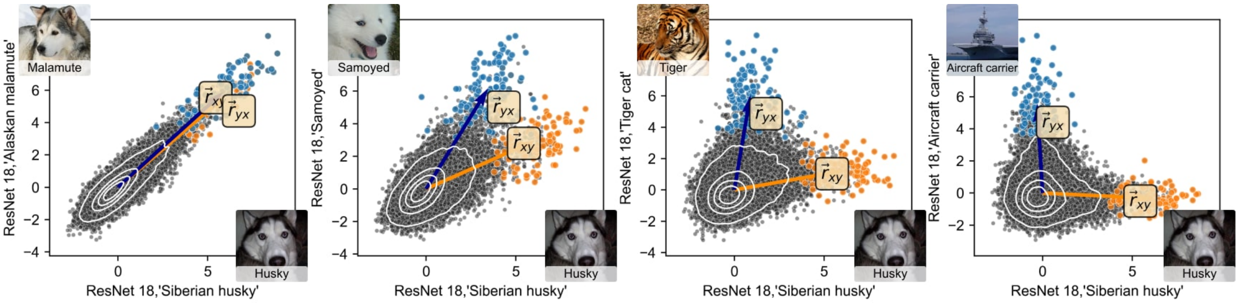


Figure 2: **Joint activation and pair-wise RAVs of different ImageNet representations.** Through four scatterplots, we can visualize the joint activations and pairwise RAVs of four neural representations, “Alaskan malamute”, “Samoyed”, “Tiger cat”, and “Aircraft carrier”, with representation “Siberian husky”, all taken from the ResNet18 output logit layer. We can observe that the angle between RAVs reflects the visual similarity between classes, representations were trained to learn: the RAVs of “Siberian husky” and “Alaskan malamute” are almost collinear due to the high visual similarity between the two dog breeds, while the RAVs of “Siberian husky” and “Aircraft carrier” are orthogonal, indicating their visual dissimilarity.

The relationship between two neural representations can be assessed by examining how each representation is activated by the n-AMS values of the other. For this, we first introduce the *representation activation vectors* (RAVs).

Definition 4. *Let $\mathcal{F} = \{f_1, ..., f_k\}$ be a layer including k neural representations, and $\mathcal{S} = \{S_1, ..., S_k\}$ be a collection of n n-AMS for each representation in the layer. For $\forall a, b \in \{1, ..., k\}$ we define $\mathbf{a}_b^a = [f_b(s_1^a), ..., f_b(s_n^a)]$, where $s_t^a \in S_a, \forall t \in \{1, ..., k\}$ is a vector of activations of neural representation f_b computed across the collection of the n-AMS of representation f_a. Additionaly, we introduce $\mu_b^a = \frac{1}{n}\sum_{t=1}^{n} f_b\left(s^a\right)$ as mean activation of f_b given the n-AMS of f_a.*

For any two representations $f_i, f_j \in \mathcal{F}$, we define their pair-wise representation activation vectors (RAVs) r_{ij}, r_{ji} as:

$$r_{ij} = \begin{pmatrix} \mu_i^i \\ \mu_j^i \end{pmatrix}, \quad r_{ji} = \begin{pmatrix} \mu_i^j \\ \mu_j^j \end{pmatrix}. \tag{6}$$

In addition, for each neural representation $f_i \in \mathcal{F}$, we define the corresponding layer-wise RAV as follows:

$$r_{i*} = \begin{pmatrix} \mu_1^i \\ \vdots \\ \mu_k^i \end{pmatrix}. \tag{7}$$

The concept behind RAVs is to capture how one representation’s n-AMS are perceived by another. To achieve this, we gather n n-AMS for each representation within the layer, then use the model to infer them and collect activations across the representations from the layer and average the embeddings. While pair-wise RAVs encode information about how two neural representations respond to each other’s stimuli, layer-wise RAVs utilize all representations in the layer as descriptors. The angle between RAVs can serve as a measure of the semantic similarity between representations. If two representations encode similar concepts, their n-AMS will probably be visually similar, resulting in colinear RAVs, since both representations will be activated by each other’s n-AMS. Conversely, if the representations encode different concepts, their n-AMS will successively depict different concepts, resulting in the orthogonality of RAVs.

To illustrate the concept of *representation activation vectors*, we calculated n-AMS for five distinct neural representations extracted from the output layer of the ImageNet pre-trained ResNet18 model. These representations corresponded to the classes “Siberian husky”, “Alaskan malamute”, “Samoyed”, “Tiger cat”, and “Aircraft carrier”, which were selected manually to demonstrate the decreasing visual similarity between the classes and the “Siberian husky” class. Using the ImageNet-2012 validation dataset, we computed the signals with a sample size of $n = 100$ and a subset size of $d = 500$. Figure 2 presents a scatter plot of activation values across all data points and pair-wise RAVs. Our results indicate that the angle between these vectors increases with the visual dissimilarity between the classes.

In this regard, we propose a novel distance metric between representations, the Extreme-Activation distance, which assesses the similarity between two neural representations based on the angle between RAVs and is defined as follows:

Definition 5 (Extreme-Activation distance)**.** *Let* f_i, f_j *be two neural representations, and* r_{ij}, r_{ji} *be their pair-wise RAVs. We define a pair-wise Extreme-Activation distance as*

$$d^{p}_{EA_n}(f_i, f_j) = \frac{1}{\sqrt{2}}\sqrt{1 - \cos(r_{ij}, r_{ji})}, \tag{8}$$

where $\cos(A, B)$ *is the cosine of the angle between vectors* A, B*.*

Additionally, given a layer $\mathcal{F} = \{f_1, ..., f_i, ..., f_j, ..., f_k\}$ *with* k *neural representations, we define (layer-wise) the Extreme-Activation distance between* f_i, f_j *as*

$$d^{l}_{EA_n}(f_i, f_j) = \frac{1}{\sqrt{2}}\sqrt{1 - \cos(r_{i*}, r_{j*})}. \tag{9}$$

3.2 Synthetic Extreme-Activation distance

Although data-aware distance metrics can offer insight into the relationships between representations, their dependence on the data can be viewed as a limitation. Modern machine learning models are often trained on closed-source or very large datasets, making it difficult to obtain the exact dataset the model was trained on. When the evaluation dataset differs from the training dataset, the correctness of the computed distances can no longer be guaranteed. For instance, if the evaluation dataset lacks some concepts present in the training dataset, distance measures may be misleading, as illustrated in Figure 3, where analysis based solely on natural signals leads to erroneous conclusions about the learned concept due to the absence of the true concept in the dataset. Furthermore, in this case we cannot be certain that distances computed over the evaluation dataset reflect the general behavior of the functions learned from the training dataset, rather than reflecting biases in the evaluation dataset. For example, even though a relatively high Pearson correlation (0.51) was reported between the ResNet18 representations of "tennis ball" and "toy terrier" computed over the ImageNet validation dataset, it is unclear whether this is a bias introduced in the ImageNet validation dataset or general behavior of the model (e.g. same correlation could be also observed in the ImageNet training dataset).

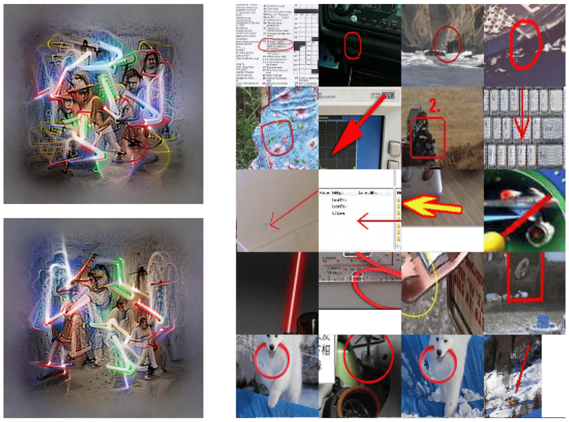

s-AMS n-AMS

Figure 3: **Failing to explain “Star Wars” representation with ImageNet n-AMS.** Comparison of the s-AMS (left) and n-AMS (right) collected from the ImageNet dataset for unit 744 in the last convolutional layer of the CLIP ResNet50 model. Due to the inaccessibility of the training dataset and lack of specific images due to copyright restrictions, n-AMS struggle to illustrate the concept of the “Star Wars” neuron. Illustrated signals were obtained from OpenAI Microscope.

We suggest a data-agnostic approach to address the limitations of data-aware distance measures. This approach is not reliant on the data and involves using the Extreme-Activation distance, which is computed based on synthetic Activation-Maximization signals (s-AMS). These synthetic signals are generated through an optimization procedure by the model itself and do not require external generative models or data.

Definition 6 (s-AMS)**.** *Let f_i be a neural representation. A synthetic Activation-Maximisation signal $\tilde{s}^i$ is defined as follows:*

$$\tilde{s}^i = \arg\max_{\tilde{s}} f_i(\tilde{s}).$$

Generating s-AMS for a neural representation is a non-convex optimization problem Nguyen et al. (2019) that typically employs gradient-based methods Erhan et al. (2009); Nguyen et al. (2015); Olah et al. (2017). Starting from a random noise parametrization of input signals, the gradient-ascend procedure searches for the optimal set of signal parameters that maximize the activation of a given representation. Early methods employed standard pixel parametrization Erhan et al. (2009), while modern approaches used Generative Adversarial Network (GAN) generators Nguyen et al. (2016) or Compositional Pattern Producing Networks (CPPNs) Mordvintsev et al. (2018); Stanley (2007). In this study, we use the Feature Visualization method Olah et al. (2017) for s-AMS generation, which parametrizes input signals by frequencies and maps them to the pixel domain using Inverse Fast Fourier Transformation (IFFT). This method is popular for its simplicity and independence from external generative models, as well as for its ability to be human-interpretable Olah et al. (2020); Goh et al. (2021); Cammarata et al. (2020).

The optimization procedure for s-AMS generation has several adjustable hyperparameters, including the optimization method and transformations applied to signals during the procedure. One critical parameter is the number of optimization steps (epochs) m, which is analogous to the parameter d for n-AMS generation.

Since different random initializations in the parameter space can lead to the convergence of s-AMS generation into different local solutions, the resulting s-AMS can vary. This variability is similar to the variability observed when sampling n-AMS. To address this, we generated n s-AMS signals for each representation f_i, defining $\tilde{S}_i = \{\tilde{s}^i_1, ..., \tilde{s}^i_n\}$ as a collection of n s-AMS for the given representation.

Definition 7 (Synthetic Extreme-Activation distance)**.** *Let f_i, f_j be two neural representations, and let $\tilde{S}_i, \tilde{S}_j$ be the collections of their respective s-AMS. Similarly to Definition 4, we define $\tilde{\mu}^a_b = \frac{1}{n}\sum_{t=1}^n f_b(\tilde{s}^a_t)$.*

We define a pair of synthetic representation activation vectors $\tilde{r}_{ij}, \tilde{r}_{ji}$, such that

$$\tilde{r}_{ij} = \begin{pmatrix} \tilde{\mu}^i_i \\ \tilde{\mu}^i_j \end{pmatrix}, \quad \tilde{r}_{ji} = \begin{pmatrix} \tilde{\mu}^j_i \\ \tilde{\mu}^j_j \end{pmatrix}. \tag{10}$$

In addition, for each neural representation $f_i \in \mathcal{F}$, we define a layer-wise synthetic RAV as:

$$\tilde{r}_{i*} = \begin{pmatrix} \tilde{\mu}^i_1 \\ \vdots \\ \tilde{\mu}^i_k \end{pmatrix}. \tag{11}$$

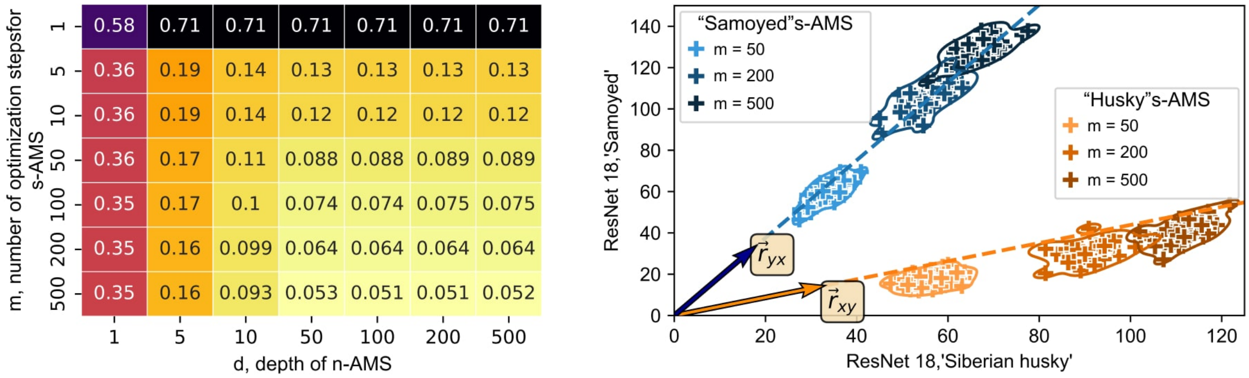
Furthermore, we define a pair-wise synthetic Extreme-Activation distance as

$$d^p_{EA_s}(f_i, f_j) = \frac{1}{\sqrt{2}}\sqrt{1 - \cos(\tilde{r}_{ij}, \tilde{r}_{ji})}, \tag{12}$$

and layer-wise synthetic Extreme-Activation as

$$d^l_{EA_s}(f_i, f_j) = \frac{1}{\sqrt{2}}\sqrt{1 - \cos(\tilde{r}_{i*}, \tilde{r}_{j*})}. \tag{13}$$

To distinguish between Extreme-Activation distances based on n-AMS and s-AMS, we denote them as EA$_n$ and EA$_s$ distances, respectively. Notably, EA$_n$ distance is computed using standardized activations. However, due to the data-agnostic nature of the EA$_s$ distance, standardization cannot be performed without accessing the evaluation dataset; hence, we use the raw representation's activations. Although this could be considered as a limitation, since the EA$_s$ distance is not shift-invariant, we found in our practical experiments that the angles between synthetic and natural RAVs are typically maintained.



m, number of optimization stepsfor s-AMS \ d, depth of n-AMS	1	5	10	50	100	200	500
1	0.58	0.71	0.71	0.71	0.71	0.71	0.71
5	0.36	0.19	0.14	0.13	0.13	0.13	0.13
10	0.36	0.19	0.14	0.12	0.12	0.12	0.12
50	0.36	0.17	0.11	0.088	0.088	0.089	0.089
100	0.35	0.17	0.1	0.074	0.074	0.075	0.075
200	0.35	0.16	0.099	0.064	0.064	0.064	0.064
500	0.35	0.16	0.093	0.053	0.051	0.051	0.052

Figure 4: **Similarity and angle preservation between EA$_n$ and EA$_s$ distance measures.** The left part of the figure shows the RMSE (lower is better) between pair-wise EA$_n$ and EA$_s$ distances on the output layer of the ResNet18 network, with a fixed parameter $n = 50$ for both metrics, while varying parameters d, corresponding to the subset size in n-AMS sampling, and m, number of epochs for s-AMS generation. On the right part of the figure, the distributions of pair-wise activations of s-AMS signals are visualized with different parameters m for two neural representations, namely “Samoyed” and “Siberian husky”, overlayed with the direction of natural RAVs computed with $n = 50$ and $m = 1000$.

EA$_s$ conserves the angle between natural RAVs

Although both n-AMS and s-AMS activate specific neural representations maximally, the adversarial nature of synthetic signals needs to be considered. In our experiments, we observed that while the generated s-AMS are far from the original *natural* image domain, the angles between natural and synthetic RAVs are consistent, providing additional evidence to the utility of the EA$_s$ distance metric.

To evaluate the angle conservation quantitatively, we employed a ResNet18 pre-trained on the ImageNet dataset and computed EA$_n$ and EA$_s$ distances between the output logit representations, i.e., all 1000 ImageNet classes. For this experiment, we fixed the number of signals to $n = 50$ for both distance metrics, while varying the parameter d for n-AMS generation and the parameter m for s-AMS generation. With $\mathcal{F} = \{f_1, \dots, f_k\}$ corresponding to the ResNet18 output layer with $k = 1000$ neural representations, we measured the root mean square error between pairwise EA$_n$ and EA$_s$ distances:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left(d^p_{EA_n}(f_i, f_j) - d^p_{EA_s}(f_i, f_j)\right)^2}{k(k-1)/2}}, \tag{14}$$

where $k(k-1)/2$ corresponds to the number of all unique pairs of two different functions from a set of k functions.

Figure 4 illustrates the similarity between the computed EA$_n$ and EA$_s$ distances between the representations of the 1000 ImageNet classes. In the left part of the figure, which shows RMSE between the two distance measures for different parameters, we observe that for each parameter m for EA$_s$ distance, the lowest error is achieved with an EA$_n$ distance with high values of d. This indicates that the EA$_s$ distance captures the angle between RAVs corresponding to the top activating images. Additionally, we observed that increasing the parameter m is beneficial to lowering the RMSE between natural and synthetic measures. Furthermore, the right part of the figure shows the direction of natural RAVs and activations of s-AMS for “Samoyed” and “Siberian husky” representations from ResNet18. From this figure, we can observe that the angle between natural RAVs and synthetic RAVs is conserved.

4 DORA: Data-agnOstic Representation Analysis

In the following, we introduce the DORA (Data-agnostic Representation Analysis) framework for analyzing representation spaces of DNNs. The proposed DORA analysis utilizes the data-independent *Extreme-Activation* (EA) distance measure to investigate the relationships between neural representations, providing insights into the model decision-making processes. Here we outline several potential applications.

4.1 Investigating Neural Associations

The functional distance metric can be utilized to investigate various learned associations in neural representations. As the training data inherently contains various correlations, the models learn such correlations, which affects their behavior and decision-making strategies. While some associations may be harmless and based on the visual similarity of the concepts (e.g., the similarity between dog breeds, such as “Alaskan malamute” and “Siberian husky”), others might be damaging to the model’s generalizability or introduce potential biases (e.g., the connection between immigration and Latin America neurons or terrorism and Middle East neurons reported in the CLIP model Goh et al. (2021)). Analyzing the functional distance between representations, particularly employing the EA_s distance, can help to discover such associations. For instance, in ResNet18, we found that the “Fountain” and “Fireboat” logit representations have a small functional distance, possibly due to the shared concept of the water jet. Moreover, we found a low functional distance between “Steam Locomotive” and “volcano,” which might be due to the shared concept of smoke clouds. Examining such associations can aid in auditing the model and uncovering previously unknown spurious correlations, thus increasing the transparency of the models.

4.2 Visualizing representation space with Representation Atlases

Inspired by Carter et al. (2019), the visual examination of the functional diversity within one layer can be done by employing the dimensionality reduction method on a given distance matrix between representations. Such visualization, referred to as the *representation atlas*, allows researchers to visually examine the topological landscape of learned representations and identify clusters of semantically similar representations. In the scope of this paper, we employed the widely used UMAP dimensionality reduction algorithm McInnes et al. (2018), which has established itself in recent years as an effective method for visualizing relationships between data points. Figure 5 depicts the representation atlas of the output logit layers of the ResNet18 model trained on ImageNet. Each point in the figure corresponds to an individual neural representation among the 1000 representations in the output layer. The color of each point reflects the WordNet hypernym, a high-level synset, that corresponds to the learned concept of the particular representation. The UMAP visualization, based on the computed EA_s distances, reveals the clusters of semantically similar representations that are preserved, which can be observed in Figure 5.

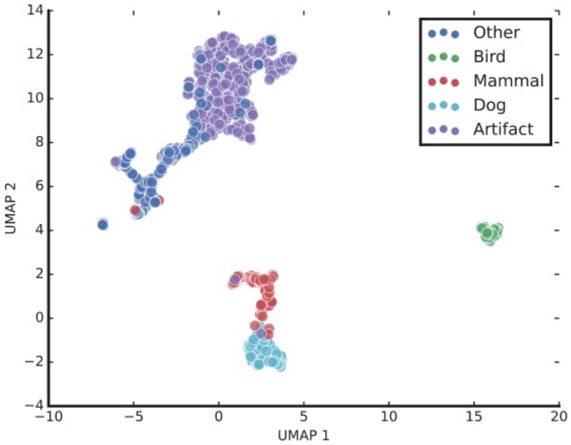


Figure 5: **Representation Atlas of ResNet18 Output Layer.** Illustration of a UMAP visualization of the layer-wise Euclidean distances (EA_s) between the output logit representations from a ResNet18 model trained on ImageNet. Each point in the visualization represents an individual neural class representation, colored by the respective WordNet hypernym.

In comparison with other dimensionality reduction methods, such as t-SNe Van der Maaten and Hinton (2008) and PCA Jolliffe and Cadima (2016), UMAP is scalable, exhibits a faster computation time McInnes et al. (2018); Trozzi et al. (2021); Becht et al. (2019); Wu et al. (2019), and has fewer parameters to tune compared to other dimensionality reduction methods. Qualitatively, compared to the other methods, UMAP was reported to improve visualizations and accurately represent the data structure on the projected components Trozzi et al. (2021); Becht et al. (2019); Wu et al. (2019).

4.3 Identifying outlier representations

Although Deep Neural Networks have demonstrated high effectiveness in various applications, they are prone to learning unintended artifacts and spurious correlations from the data, resulting in unnatural and semantically different features from the concepts within the dataset. This undesirable behavior persists throughout the model, making internal representations susceptible to learning malicious concepts. Analysis of the functional EA_s distance between representations within a specific layer of the network can reveal a set of functions that deviate from the majority of functions within the layer. Due to the high degree of alignment of functional distance and human judgment, we might expect that such outlier representations in the functional space encode semantically anomalous concepts. While some outlier representations could learn unique, natural concepts relevant to the task, in our practical experiments, we observed that such representations often encode undesired concepts, constituting the *shortcut learning* or *Clever Hans* behavior.

5 Evaluation

One of the key considerations to address when comparing various distance metrics between neural representations is their consistency with human perception. If the underlying concepts of the representations differ semantically from a human perspective, we would expect from our functional distance measure to reflect this difference. To quantitatively assess the alignment of various distance metrics, we compare the computed distances generated by these metrics with human-defined distance metrics between concepts in scenarios where the latter are available.

Therefore, we employed two widely used computer vision datasets, ILSVRC2012 Deng et al. (2009) and CIFAR-100 Krizhevsky (2009). For each of these datasets, human-defined semantic distances were obtained by mapping the classification labels to entities in the WordNet taxonomy database Miller (1995), a lexical database that organizes English words into a taxonomy of synonym sets, or synsets. In this taxonomy, each synset represents a group of words that are synonyms or have the same meaning. WordNet organizes these synsets into a hierarchy, with more specific concepts being nested under more general ones. For the ImageNet dataset, class labels were mapped automatically due to the cross-connection with WordNet synsets, while for CIFAR-100 labels were matched manually.

Given the WordNet taxonomy in a form of an undirected graph $\mathcal{G} = (V, E)$ with root $r \in V$, the baseline semantic distances between entities from the WordNet database were computed using the following three distance measures:

- **Shortest-Path distance**

  Given two vertices $c_i, c_j \in V$ the distance between vertices is determined by the length of the shortest path that connects the two entities in the taxonomy.

  $$d_{SP}(c_i, c_j) = l(c_i, c_j),$$

  where $l(c_i, c_j)$ is the function, corresponding to the minimal number of edges that need to be traversed to get from c_i to c_j.

- **Leacock-Chodorow distance Leacock and Chodorow (1998)**

  Given two vertices $c_i, c_j \in V$ the distance between vertices is determined by a logarithm of the shortest-path distance with additional scaling by the taxonomy depth:

  $$d_{LC}(c_i, c_j) = \log \frac{l(c_i, c_j) + 1}{2T} - \log \frac{1}{2T},$$

  where $T = \max_{c \in V} l(r, c)$ is the taxonomy depth.

- **Wu-Palmer distance Wu and Palmer (1994)**

  Given two vertices $c_i, c_j \in V$ the Wu-Palmer distance is defined as:

  $$d_{SP}(c_i, c_j) = 1 - 2\frac{l(r, lcs(c_i, c_j))}{l(r, c_i) + l(r, c_j)},$$

where $lcs(c_i, c_j)$ is the Least Common Subsumer Pedersen et al. (2004) of two concepts c_i and c_j.

Furthermore, we have utilized the textual labels from both ImageNet and CIFAR100 datasets and calculated the Word2Vec Mikolov et al. (2013) similarity between class labels as an extra semantic benchmark to evaluate the alignment.

- **Word2Vec distance**

  Given textual labels t_i, t_j of two concepts c_i, c_j, we define Word2Vec distance as

  $$d_{W2V} = \frac{1}{\sqrt{2}} \sqrt{1 - \cos_{W2V}(t_i, t_j)},$$

  where $\cos_{W2V}(A, B)$ is the cosine of the angle between Word2Vec embeddings of the words A, B.

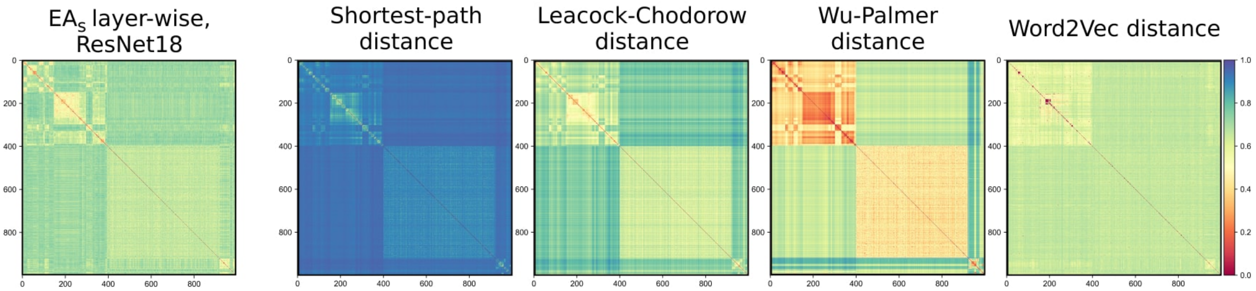


Figure 6: **Visualization of Semantic Baselines and EA$_s$ Distances for ImageNet Classes**: From left to right: the EA$_s$ distance metric computed for the output logits of the ImageNet pre-trained ResNet18 model, Shortest-Path, Leacock-Chodorow, Wu-Palmer distances from WordNet taxonomy, and Word2Vec distance.

Figure 6 illustrates baseline distances between ImageNet classes, alongside the distance matrix computed using our proposed EA$_s$ metric for the ResNet18 output logits. To evaluate the alignment between the proposed distance metric and a human-defined baseline, we employed the Mantel Test Mantel (1967), which is often employed in ecology and evolutionary biology to measure the correlation between two distance matrices. The test calculates the correlation coefficient ρ, which indicates the strength of the relationship between the two matrices, and the p-value of the test, which describes the statistical significance of the correlation.

It is essential to note that while we evaluate the alignment based on human-defined semantic benchmarks, optimizing such metrics should not be the ultimate objective when proposing new distance metrics between representations. This is because DNNs can naturally employ different decision-making strategies than humans, and these differences may not always be attributed to spurious correlations. For instance, taxonomy-based approaches might be sub-optimal compared with attributing freedom to the models to train for the desired tasks Binder et al. (2012). Conversely, in Computer Vision, network representations are expected to be aligned to some extent due to the correlations between visual and semantic similarity of classes Brust and Denzler (2019); Deselaers and Ferrari (2011).

5.1 Hyperparameter selection

Incorporating distance metrics for representations frequently relies on a set of hyperparameters. This section examines the selection of parameters in terms of their ability to attain optimal alignment with the semantic baselines. In our experiments, we employed a pre-trained ResNet18 model on ImageNet, and the ImageNet-2012 validation set with 50,000 images across 1,000 classes for computing the data-aware distances, including Minkowski, Pearson, Spearman, and EA$_n$ distances. To ensure consistency, we standardized the activations of the representations by centering them around zero and scaling their standard deviation to 1 for each of the 1,000 neural representations across the dataset, as previously described.

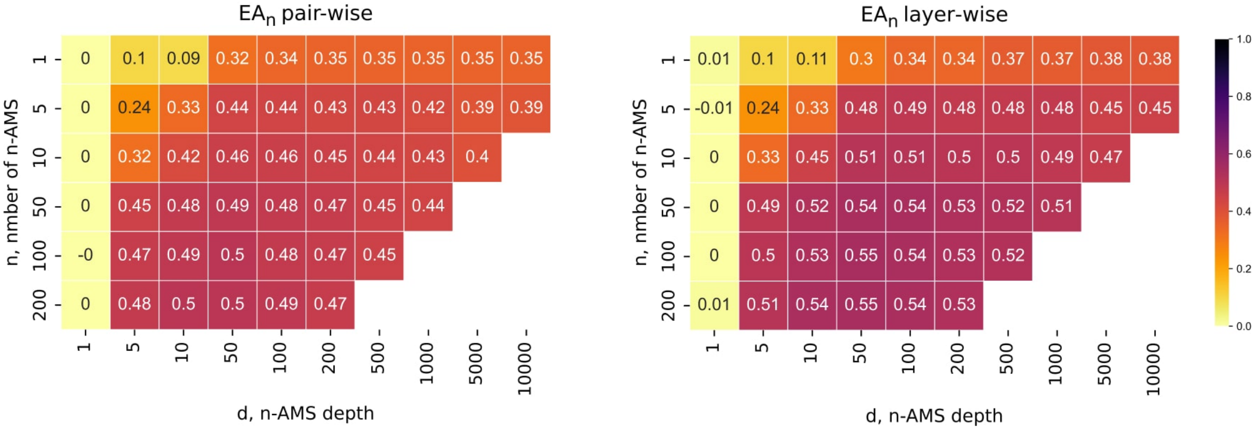


Figure 8: **Impact of parameter selection in EA$_n$ distance on alignment with semantic baselines.** To assess alignment with four semantic baselines, we calculated data-aware EA distance on the output logits of the ResNet18 network while varying the parameters n and d for both pair-wise and layer-wise options. The average Mantel correlation statistic across four semantic baselines is reported at each cell.

Minkowski distance

To investigate how different values of the parameter p affect the coherence to semantic baselines, we varied the parameter and evaluated the alignment with four semantic baselines for each case. Figure 7 shows the effect of parameter selection on the Mantel test statistic. We observed that the optimal average value of the statistic across the four baselines was achieved for $p = 2$. However, for future experiments, we selected the second-best parameter choice with $p = 1$ due to the natural connection between Euclidean distance and Pearson distance. We also observed that higher values of p generally result in lower alignment, possibly due to sensitivity to the large amplitudes of individual data points.

EA$_n$ distance

Data-Aware Extreme Activation distance is influenced by two key parameters: n, which denotes the number of n-AMS signals gathered, and d, which represents the size of the subset collected from each signal. To investigate the impact of parameter selection, we varied these parameters for both pair-wise and layer-wise modes. Figure 8 shows the average Mantel correlation statistic across four semantic baselines for each hyperparameter choice. Our observations reveal that, in general, increasing the number of collected n-AMS, irrespective of the parameter d, has a positive impact on the alignment. However, the optimal depth d is achieved when n-AMS are taken from subsets of $d = 50$ datapoints.

EA$_s$ distance

In the data-agnostic version of the Extreme-Activation distance, the choice of hyperparameters depends on the s-AMS generation method used. In our study, we employed the Feature Visualisation method to generate s-AMS, and we identified two critical hyperparameters: n, which is the number of

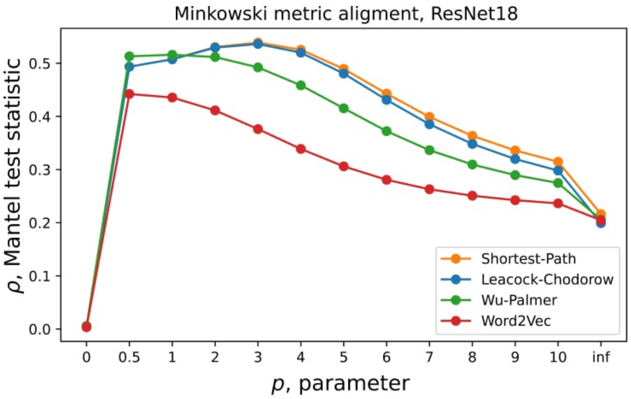


Figure 7: **Impact of Parameter Selection in Minkowski Distance on Alignment with Semantic Baselines.** To assess the alignment with respect to the four semantic baselines, we calculated the Minkowski distance on the output logits of the ResNet18 network while varying the parameter p. The Mantel correlation statistic was reported for each semantic baseline at each parameter value.

generated s-AMS per representation, and m, which is the number of optimization epochs per signal. Figure 9 depicts the impact of the EA_s distance measure's hyperparameter selection on semantic baseline alignment. We observed that while increasing the number of generated s-AMS generally has a positive effect, this effect is negligible compared to the positive impact of increasing the number of optimization epochs per representation. This is likely due to the generation algorithms' convergence to better local optima, resulting in improved visual preciseness of the images, as shown on the right side in Figure 9.

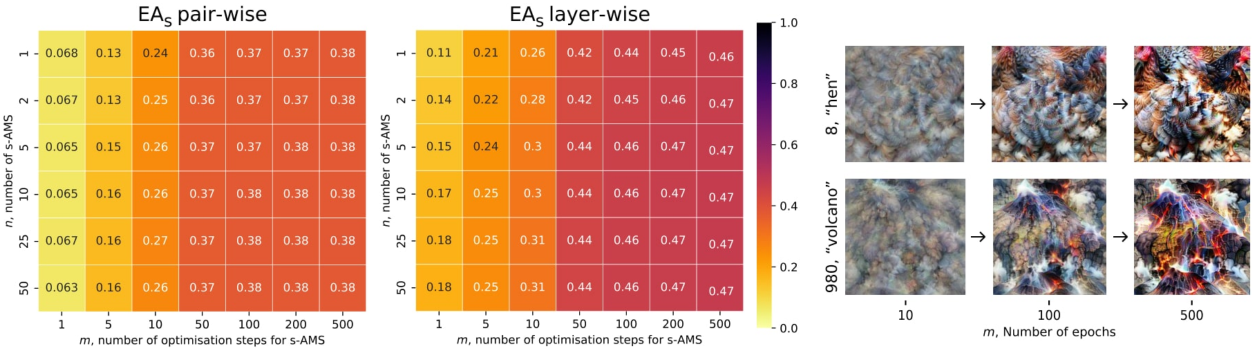


Figure 9: **Impact of the parameter selection in EA_s distance on alignment with semantic baselines.** To evaluate the alignment between the EA_s distance and four semantic baselines, we computed the data-agnostic EA distance using the ResNet18 network's output logits while varying the hyperparameters n and m for both pair-wise and layer-wise options. For each cell, we reported the average Mantel correlation statistic across the four semantic baselines. The effect of the hyperparameter m, which corresponds to the number of optimization steps taken for s-AMS generation, on two neural representations from the ResNet18 output logit layer is shown on the right.

5.2 Evaluating the alignment with human judgment

In this experiment, we quantitatively assess the alignment of the discussed distance metrics with the human-defined distance measures across different datasets and architectures. To this end, we employed eight different architectures for two datasets, ImageNet and CIFAR100. For ImageNet, we employed ResNet18 He et al. (2016), AlexNet Krizhevsky et al. (2017), ViT Dosovitskiy et al. (2020), BEiT Bao et al. (2021), Inception V3 Szegedy et al. (2016), DenseNet 161 Huang et al. (2017), MobileNet V2 Sandler et al. (2018), ShuffleNet V2 Ma et al. (2018), while for CIFAR-100, we used ResNet 18, ResNet 9, MobileNet V2, ShuffleNet V1, and V2, as well as NASNet Qin and Wang (2019), SqueeeNet Iandola et al. (2016) and VGG 11 Simonyan and Zisserman (2014).

We computed functional distances with optimal hyperparameters found in Section 5.1, including Minkowski $p = 1$, Pearson, Spearman, EA_n with $n = 50, d = 200$, and EA_s with $n = 3, m = 500$, on the output logit layer for each model. We then compared each distance matrix with four semantic baselines: Shortest-Path, Leacock-Chodorow, Wu-Palmer distances from WordNet taxonomy, and Word2Vec distance. This comparison yielded four Mantel test statistics per distance metric. The results of the evaluation are presented in Table 1 for ImageNet-trained models and in Table 2 for CIFAR100 models, where we averaged the four Mantel correlation test statistics for each model and distance metric. Our analysis indicates that the layer-wise EA_n metric's distance is generally more favorable due to its stronger linear relationship with all four baseline metrics. Furthermore, we observed that the data-agnostic EA_s metric is on par with data-aware metrics in terms of coherence with the semantic baselines.

5.3 Evaluating Anomaly-Identification capabilities

Alignment of the distance metrics between neural representations and the human judgment of the distance between concepts opens an interesting possible application — based on the functional distance, we can identify representations, that are semantically anomalous to the majority of learned representations. While

Table 1: **Alignment of Distance Metrics in ImageNet Trained Models**: Each cell represents the average Mantel test statistic across four semantic baselines: Shortest-Path, Leacock-Chodorow, Wu-Palmer distances, and Word2Vec distance. All results demonstrate statistical significance with $p < 0.001$.

	Minkowski	*Pearson*	*Spearman*	EA_n		EA_s	
	$p = 1$			*p-w*	*l-w*	*p-w*	*l-w*
ResNet18	0.49	0.50	0.48	0.49	0.55	0.38	0.47
BeIT	0.32	0.36	0.29	0.44	0.50	0.39	0.47
MobilenetV2	0.46	0.46	0.45	0.47	0.52	0.40	0.50
DenseNet161	0.46	0.47	0.44	0.49	0.54	0.32	0.39
ShuffleNetV2	0.21	0.21	0.19	0.29	0.30	0.19	0.16
InceptionV3	0.31	0.34	0.32	0.38	0.49	0.22	0.27
AlexNet	0.52	0.53	0.52	0.52	0.55	0.42	0.45
ViT	0.53	0.54	0.52	0.54	0.58	0.48	0.53
Mean	**0.41**	**0.43**	**0.40**	**0.45**	**0.50**	**0.35**	**0.40**

Table 2: **Alignment of Distance Metrics in CIFAR100 Trained Models**: Each cell represents the average Mantel test statistic across four semantic baselines: Shortest-Path, Leacock-Chodorow, Wu-Palmer distances, and Word2Vec distance. All results demonstrate statistical significance with $p < 0.001$.

	Minkowski	*Pearson*	*Spearman*	EA_n		EA_s	
	$p = 1$			*p-w*	*l-w*	*p-w*	*l-w*
ResNet9	0.32	0.37	0.33	0.41	0.52	0.27	0.30
ShuffleNetV2	0.49	0.52	0.49	0.53	0.59	0.43	0.47
MobileNetV2	0.50	0.51	0.49	0.52	0.59	0.40	0.44
ResNet18	0.43	0.47	0.45	0.48	0.57	0.30	0.37
ShuffleNet	0.48	0.51	0.49	0.52	0.58	0.42	0.46
VGG11	0.30	0.31	0.31	0.36	0.43	0.23	0.23
NasNet	0.48	0.51	0.48	0.52	0.59	0.36	0.41
SqueezeNet	0.50	0.52	0.51	0.53	0.59	0.45	0.51
Mean	**0.44**	**0.46**	**0.44**	**0.48**	**0.56**	**0.36**	**0.40**

these representations may simply learn unique individual concepts, we demonstrate in further experiments that in real-life scenarios they might correspond to the undesired concepts from spurious correlations in the training data that diverge from the typical (intended) decision-making strategy.

To assess the usefulness of the alignment between distance metrics and human-defined semantic baseline, we conducted the experiment, where we measured the ability of the distance metrics to detect semantically anomalous representations. For this purpose, we conducted a toy experiment by training a ResNet18 He et al. (2016) network on a combination of two conceptually different datasets. The combined dataset comprised the Tiny Imagenet Le and Yang (2015), containing 200 ImageNet classes, and the MNIST handwritten-numbers dataset Deng (2012), containing 10 handwritten numbers, resulting in a total of 210 classes. MNIST images were upsampled to the size of $3 \times 64 \times 64$ pixels to match the size of images in Tiny ImageNet. After training on the combined dataset in the image classification task, we computed functional distances between the output logits and evaluated the ability of different Outlier Detection methods to detect MNIST logit, given the computed distance matrices only. For this, we utilized five different Outlier Detection methods: the Angle-based Outlier Detector (ABOD) Kriegel et al. (2008), Feature Bagging (FB) Lazarevic and Kumar (2005), Isolation Forest (IF) Liu et al. (2008), Local Outlier Factor (LOF) Breunig et al. (2000) and One-class SVM (OCSVM) Schölkopf et al. (2001). The performance of the Outlier Detection (OD) methods was evaluated using the AUC ROC metric for the classification between Tiny ImageNet representations, and the ones from MNIST classes. To ensure stability in light of the stochastic nature of some outlier detection methods, the results of the outlier detection were repeated 100 times with different random states.

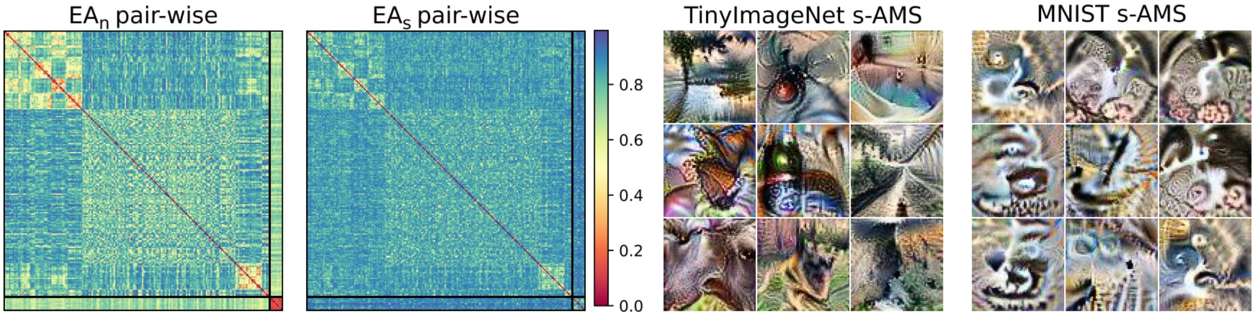


Figure 10: **Anomaly-Detection Evaluation Experiment Visualized.** From left to right, pair-wise n-AMS EA distance table between the output logits of the network trained on the combined dataset, the s-AMS EA distance table, s-AMS for Tiny ImageNet logits, and s-AMS for MNIST logits. MNIST representations are highlighted on both distance matrices in the bottom right corner, revealing a block structure in both distance metrics that suggests a high degree of functional differences between Tiny ImageNet representations and semantically distinct MNIST representations. On the left, we can visually observe differences between the s-AMS of Tiny ImageNet and MNIST representations.

We utilized the same hyperparameter configuration for distance computation as described in Section 5.2. The effectiveness of EA distances, both natural and synthetic, in distinguishing between representations of Tiny ImageNet and MNIST is demonstrated in Figure 10, as evidenced by the block structure of the distance matrices. This behavior can be attributed to the visual dissimilarities between the classes, where Tiny ImageNet classes exhibit natural and diverse features that are typical for natural images, while MNIST images consist of white digits on a black background. In the case of synthetic EA distance, the ability to detect MNIST representations is based on the visual differences in the s-AMS, which are depicted in the right-hand portion of Figure 10. The s-AMS-based EA distance measure depends on the network's ability to perceive self-generated s-AMS, and we can observe distinct dissimilarities between the patterns of s-AMS for Tiny ImageNet classes, which contain high-level natural concepts, and the more data-specific patterns for MNIST classes, which illustrate the network's perception of white-on-black handwritten digits and letters.

The results of the described experiment are presented in Table 3, which indicate that, in general, all distance metrics are capable of detecting MNIST representations. However, the EA distance metrics are more effective in detecting semantically different representations, where the pairwise EA metric is the most effective.

Table 3: **Detection Performance of Distance Metrics for Semantically Different Representations.** The table displays the average AUC ROC binary classification accuracy of the Outlier Detection methods across 100 re-trials, in the task of detecting MNIST representations among the combined Tiny ImageNet and MNIST representations, specifically in the output layer of the trained network.

	Minkowski $p=1$	*Pearson*	*Spearman*	EA_n *p-w*	EA_n *l-w*	EA_s *p-w*	EA_s *l-w*
ABOD	0.56	0.63	0.58	0.91	1.00	0.82	0.71
FB	0.97	0.99	0.81	1.00	1.00	0.89	0.87
IF	0.83	0.87	0.64	0.94	0.70	0.76	0.61
LOF	0.65	0.53	0.55	0.67	0.96	1.00	0.87
OCSVM	1.00	1.00	0.95	1.00	0.67	1.00	0.72
Mean	**0.80**	**0.80**	**0.71**	**0.90**	**0.87**	**0.89**	**0.76**

6 Experiments: Finding outlier representations

As previously demonstrated, the DORA framework facilitates the visualization of a topological map of representations in a designated layer and is able to identify outlier representations. In this section, we aim to investigate the latent representations of widely-used computer vision architectures and demonstrate that the outlier representations found by DORA in real-life scenarios may align with undesirable Clever-Hans concepts and deviate from the intended decision-making approach.

6.1 ImageNet pre-trained networks

Pre-trained networks on ImageNet have become an essential component in the field of Computer Vision. Their capability to recognize a diverse set of objects and scenes makes them particularly useful as a starting point for a wide range of computer vision tasks. They are frequently utilized for fine-tuning to specific tasks or as a feature extractor, where the images are encoded by the networks for further computations Zhuang et al. (2020); Weiss et al. (2016).

In the following we explore the feature extractor representations of three widely-used pre-trained models: ResNet18 He et al. (2016), MobileNetV2 Sandler et al. (2018), and DenseNet121 Huang et al. (2017). Using LOF outlier detection, we found latent layers with representations that appear to be watermark detectors, e.g., detecting Chinese and Latin text patterns. As ImageNet does not have a specific category for watermarks, these representations could be seen as Clever-Hans artifacts and deviate from desired decision-making Lapuschkin et al. (2019); Anders et al. (2022). To verify these representations can detect watermarks, we created two binary classification datasets, for Chinese and Latin watermarks, containing normal images and identical images, with inserted random watermarks, evaluating the sensitivity of individual representations using the AUC ROC classification measure. To ensure the detection of characters and not specific words/phrases (unlike CLIP models Goh et al. (2021)), the probing datasets were generated with random characters (for more details we refer to the Appendix). Our results show that not only the reported outliers but also neighboring representations in EA distance are affected by artifactual behavior. Lastly, we find that this behavior persists during transfer learning, posing a risk for safety-critical fields like medicine.

ImageNet ResNet18

We applied DORA to analyze the Average Pooling layer, which consists of the last 512 high-level representations of the "feature extractor" that are commonly used without further modification during transfer learning. Following the DORA approach, we calculated EA layer-wise distance with $n = 5$ s-AMS per each representation and with $m = 500$, based on our findings in the section 5.1. After calculating the EA distances, we used the LOF method with a contamination parameter $p = 0.01$ (corresponding to the top 1% of representations) and the number of neighbors was set to 20 (the default value used in the `sklearn` package Pedregosa et al. (2011)).

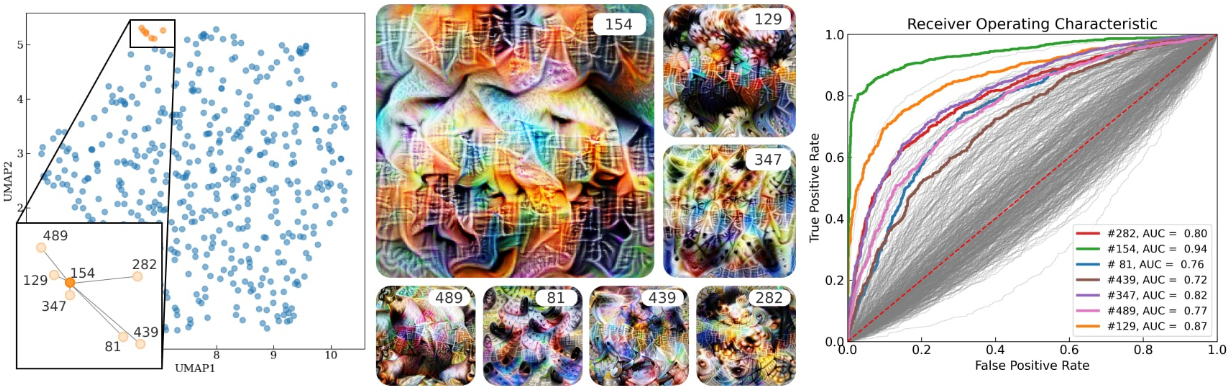


Figure 11: **Cluster of Clever-Hans representations in the ResNet18 feature extractor.** From left to right: representation atlas of the ResNet 18 average pooling layer with the highlighted cluster of Clever-Hans representations (left), s-AMS of the representations in the cluster (middle), and AUC ROC sensitivity scores for the detection of images with Chinese watermarks in the binary classification problem(right), where colored curves correspond to the behavior of representations in the cluster and gray curves for other representations. From the s-AMS of neuron 154, we can observe symbolic patterns resembling Chinese logograms learned by the neuron as well as by its closest neighbor neurons. We can observe that the outlier neuron 154 exhibits the highest AUC value (green curve), followed by its nearest neighbors.

DORA identified five outlier representations, namely neurons 7, 99, 154, 160, 162, and 393. The outlier neuron 154, displayed a specific, recognizable pattern in s-AMS that could be perceived as the presence of Chinese logograms. By probing the network on a binary classification problem between images watermarked with Chinese logograms vs normal images, Neuron 154 showed a strong detection rate (AUC ROC of 0.94) towards the class with watermarked images, providing significant evidence that this representation is susceptible to the Clever-Hans effect. Further analysis of neighboring representations in EA distance showed that they also exhibit similar behavior. The results of the analysis of the ResNet 18 average pooling layer are shown in Figure 11, illustrating the cluster of Clever-Hans representations found, along with their s-AMS and AUC ROC performance on the binary classification problem. Additional information on the dataset generation and the identified outlier representations can be found in the appendix. Furthermore, the high sensitivity of these representations in terms of their ability to detect artifacts in the data suggests a possible application for using such representations to identify artifacts in training data. Note that in general, the presence of such artifacts could indeed pose serious risks and may lead to a degradation in classifier performance (see Anders et al. (2022)).

In the further investigation of the model, we inferenced s-AMS signals of representations in the reported CH-cluster and obtained their predictions by the model. Among the selected signals, the model predominantly predicted an affiliation of these signals with the classes “carton”, “swab”, “apron”, “monitor” and “broom”, which is in line with the reported spurious correlation of the “carton” class and Chinese watermarks Li et al. (2022). Upon computing the corresponding s-AMS signals for these logits, we were able to confirm their association with CH-behaviour, as they displayed clear, visible logographic patterns, specific to Chinese character detectors, in their corresponding s-AMS. Corresponding signals and additional information could be found in Appendix.

ImageNet MobileNetV2

We used DORA with the same parameters as in the previous experiment ($n = 5$ s-AMS per each representation and $m = 500$ epochs for s-AMS generation) to analyze the “features” layer of MobileNetV2 network Sandler et al. (2018), which consists of 1280 channels with 7×7 activation maps. The analysis was performed on channels by averaging the resulting activation maps of neurons. We calculated the EA distances between representations and applied the LOF method with a contamination parameter of 0.01 which yielded 13 outlier representations. Upon visual inspection of the s-AMS of these representations, we observed distinct patterns

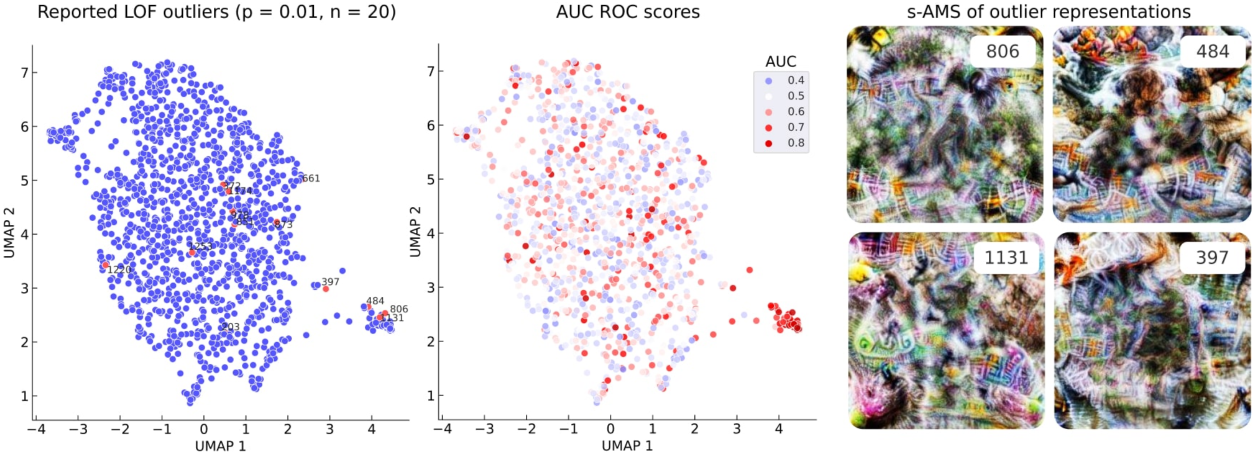


Figure 12: **Cluster of Clever-Hans representations in the MobileNet V2 feature extractor.** The left figure illustrates the outlier representations as identified by the LOF OD method, overlaid on the DORA representation atlas. The middle figure displays the sensitivity of the neural representations to Chinese watermarks, where the highly-sensitive cluster of neurons can be clearly observed in the bottom-right part of the atlas, including 3 reported outlier representations. The right graph illustrates the s-AMS of several of the reported outlier neurons, which exhibit a distinctive logographic pattern typical of Chinese character detectors.

specific to Chinese character detectors in neurons 397, 484, 806, and 1131. Figure 12 illustrates the s-AMS of these neurons, as well as the sensitivity of neurons in the Chinese-character detection task. We can observe that the neighbors of these neurons (397, 484, 806, 1131) are sensitive to CH artifacts and form a distinctive cluster visible in the representation atlas.

ImageNet DenseNet 121

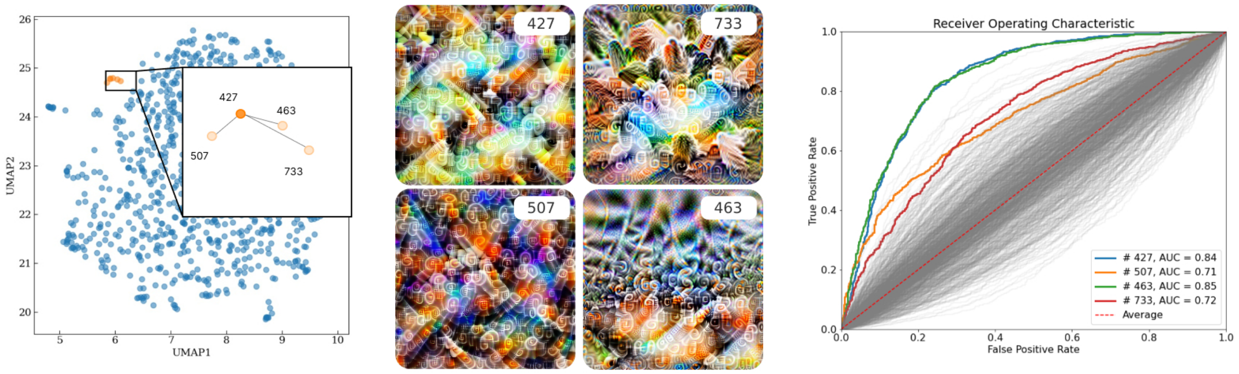


Figure 13: **DenseNet121 — Latin text detector.** Applying DORA to the last layer of the feature extractor of DenseNet121 yields, among others, Neuron 427 as an outlier, which corresponds to the upper left of the 4 feature visualizations. From neuron 427 as well as from its three closest neighbors (shown left), we can observe semantic concepts resembling Latin text characters. The AUC values were computed using the average channel activations on the Latin probing dataset. As shown, the AUCs are high for the representation outliers found by DORA, compared to most of the other representations, which indicates that they indeed learned to detect Latin text patterns.

We conducted a similar analysis on the last layer of the feature extractor of the ImageNet pre-trained DenseNet121 model, which consists of 1024 channel representations with 7×7 activation maps. We calculated $n = 5$ s-AMS per representation with $m = 150$ optimization steps for quicker experimentation. The LOF outlier detection method with a contamination parameter of $p = 0.01$ identified 10 outlier representations. One of these, neuron 768, was found to be a Chinese character detector (more information can be found in the Appendix). By increasing the contamination parameter to $p = 0.035$ (corresponding to the top 3.5% or 35 representations), we also identified neuron 427, which is susceptible to the detection of Latin text and watermarks. Figure 13 illustrates the representation atlas, highlighting representation 427 along with several neighboring representations, namely neurons 733, 507, and 463, which also exhibit a high detection rate for unintended concepts.

Clever Hans representations survive transfer learning

Given the widespread use of pre-trained models in safety-critical areas, it is essential that the artifacts embodied in a pre-trained model are made ineffective or unlearned during the transfer learning task (see also Anders et al. (2022)). To this end, we examined the effect of fine-tuning the pre-trained DenseNet121 model on the CheXpert challenge Irvin et al. (2019), which benchmarks classifiers on a multi-label chest radiograph dataset. Despite the modification of all model parameters during fine-tuning, neurons 427 and 768, which were Latin and Chinese characters detectors in the pre-trained model, retained their original semantic information and remained outliers after applying DORA. We studied neuron 427's ability to detect Latin text and found that it had an AUC value of 0.84 in the pre-trained model and 0.81 in the fine-tuned model, as shown in Figure 14. Similar behavior was observed with neuron 768, indicating that the Clever-Hans effect persisted after fine-tuning.

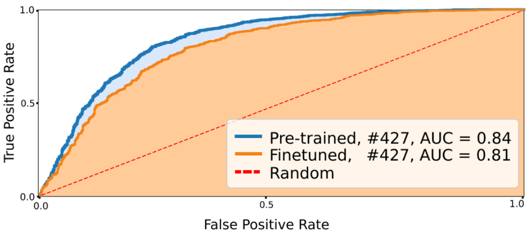


Figure 14: **Persistent Latin text detector**. Neuron 427 in the DenseNet121 network learns to detect Latin text during pre-training and does not unlearn this behavior after fine-tuning on the CheXpert dataset, as shown by the ROC detection curves. The AUC values of the neuron activations on images corrupted with Latin watermarks are high after pre-training and persist after fine-tuning.

6.2 CLIP ResNet50

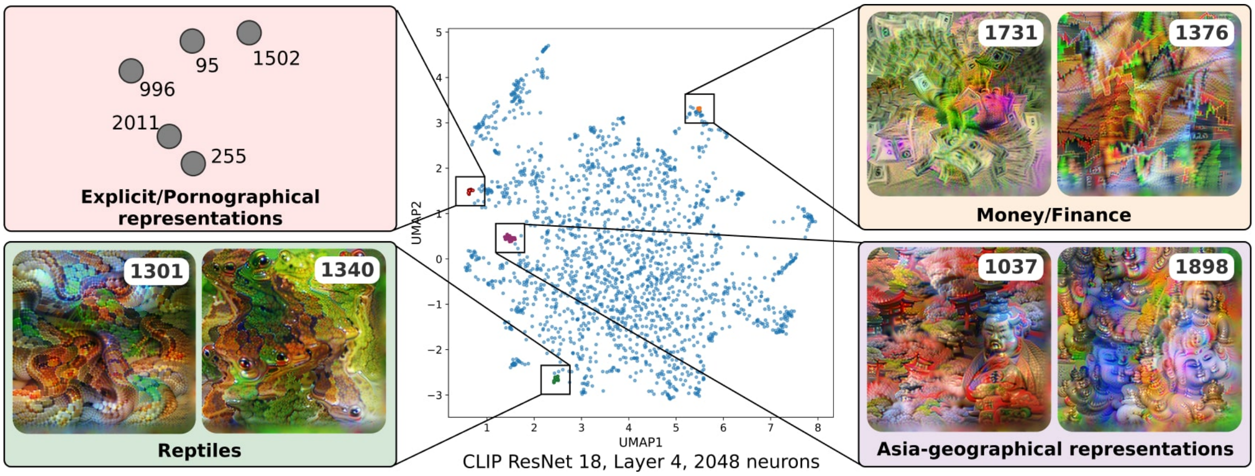


Figure 15: **Representation atlas of CLIP ResNet50 "layer 4".** Representation atlas for CLIP ResNet50 "layer 4", where several clusters of representations are highlighted. Activation-Maximisation signals associated with the Explicit/Pornographic representations were omitted due to the presence of explicit concepts in the signals.

CLIP (Contrastive Language-Image Pre-training) models predict relationships between text and images, trained using contrastive learning objective Dai and Lin (2017); Hjelm et al. (2018) on large datasets and fine-tuned on tasks such as image classification Agarwal et al. (2021) or text-to-image synthesis, where CLIP models also often serve as text encoders (e.g. Stable Diffusion Rombach et al. (2022)).

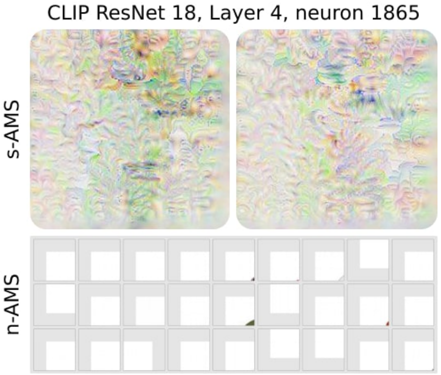


Figure 16: **AMS for reported outlier representation**. LOF identified neuron 1865 as the strongest outlier. Analysis of s-AMS and ImageNet n-AMS indicate that it primarily detects white images/backgrounds, which is atypical compared to other high-level representations in the same layer.

In this experiment, we explore the representation space of the CLIP ResNet50 model Radford et al. (2021) focusing on the last layer of its image feature extractor ("layer 4"). The training dataset was not publicly disclosed, but it is reported to be much larger than standard computer vision datasets like ImageNet, resulting in greater variability of concepts compared to ImageNet networks. We used DORA on 2048 channel representations from "layer 4", generating $n = 3$ signals per representation with $m = 512$ and using similar settings as (Goh et al., 2021).

Analysis of the outlier representations with contamination parameter $p = 0.0025$ yielded 6 outlier neurons, namely 631, 658, 838, 1666, 1865, and 1896. Representation 1865 – neuron with the highest outlier score – was found to detect the unusual concept of white images/background, as shown by synthetic and natural (collected from `OpenAI Microscope`) AMS in the Figure 16. However, the other outlier representations could not be concluded to be undesirable as they seemed to detect rare but natural concepts. Further details and analysis of the other outlier representations can be found in the Appendix.

After computing the representation atlas for "layer 4", we manually investigated several distinctive clusters. Figure 15 illustrates the representation atlas alongside several reported clusters of semantically similar representations. With our analysis, we found a cluster of Explicit/Pornographic representations. Furthermore, we were able to confirm the presence of geographical neurons, as reported in (Goh et al., 2021) and we noted that representations from neighboring geographical regions, such as India, China, Korea, and Japan, were located close to one another. Additional information and more detailed visualizations can be found in the Appendix.

During our analysis of the "layer 4" representations in the CLIP model, we confirmed the semantic multimodality of these representations Goh et al. (2021). This behavior is characterized by a multimodal distribution of AMS, whether natural or synthetic and demonstrates the representations' ability to be activated by visually and semantically distinct concepts. An example from Goh et al. (2021) illustrates this by noting that the CLIP network activates the "pizza" concept when presented with either a visual image of a pizza or the written word "pizza" within an image. Such multimodality leads to high variance in both natural and synthetic RAVs, as depicted in Figure 17 through pairwise n-AMS for two CLIP model representations, highlighting the multimodal distribution of activations across these signals.

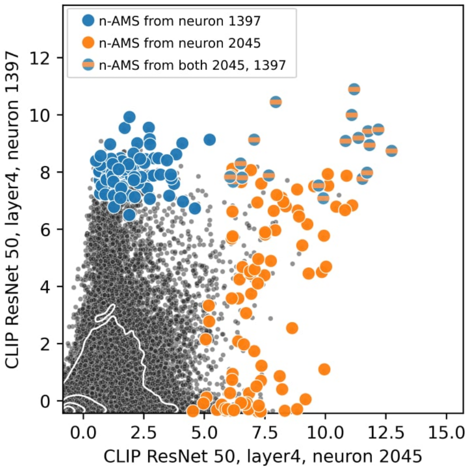


Figure 17: **Multimodality of the representations in the CLIP model**. The figure illustrates the joint distribution of activations across two representations from the CLIP ResNet50 model. The n-AMS from ImageNet are reported for both representations. As can be seen, the semantic multimodality of the representations manifests itself in the multimodal distribution of the n-AMS activations.

7 Discussion and Conclusion

Learned representations in Deep Neural Networks embody the task-relevant (see Braun et al. (2008); Montavon et al. (2011)) essence of the training data. Since it is not uncommon for datasets to contain artifacts, spurious correlations, or biases, it is more than ever essential to inspect these models using explainable artificial intelligence (XAI) methods to avoid undesirable or even harmful behavior. So far, this has mostly been done by applying local XAI methods, which require access to the data to explain the prediction of the model at hand. While such methods are powerful in discovering *local* decision-making strategies given the data-sample, they are limited with respect to uncovering global strategies, that complex models employ. Up until our work, to the best of our knowledge, no method existed to identify representations that have learned unintended or malicious concepts.

In our work, we introduced a novel problem of identifying semantically anomalous representations within a network using a functional distance metric that is aligned with human perception. We developed the DORA framework, which is straightforward and does not depend on the dataset, enabling the examination of any trained neural network without the need for access to the training data. This framework utilizes the self-explanatory nature of Computer Vision networks to estimate distances within the network. Our results demonstrated that semantically anomalous representations often encode undesirable concepts, such as watermarks in the context of the ImageNet classification problem. Such representations could be used to analyze the dataset and identify data points that contain unwanted artifacts. This analysis of representations is crucial because we also observed that artifactual representations persist even after fine-tuning, highlighting potential risks for safety-critical applications due to the widespread use of transfer learning.

Although we have demonstrated the broad applicability of DORA, there exist several limitations that require attention.

- **Systematic artifactual behavior**

  The proposed approach assumes that undesired behavior in representations is not systematic. Consequently, DORA may not be able to identify infected representations if such behavior is widespread across a large number of representations, as it would no longer be considered anomalous.

- **Semantic multimodality of representations**

  Another limitation pertains to the potential semantic multimodality of representations Goh et al. (2021), which DORA aims to mitigate by calculating multiple s-AMS per representation. However, this approach may not unveil all the concepts a representation can capture. Although this behavior was observed only in the CLIP experiment for ImageNet- and CIFAR-100-trained networks, semantic unimodality is generally observed, and additional research is necessary to address this issue.

- **Interpretation of outlier representations**

  Although the EA distance metric can be computed without access to training data, using only s-AMS for interpreting the concept behind an outlier representation may be insufficient for understanding why it was identified as an outlier. Therefore, additional analysis that may require data, such as visualizing n-AMS, is necessary.

In summary, we showed the functionality and usefulness of the DORA framework for explaining the representation spaces of Computer Vision models. Such an approach could be used for exploring associations between representations, visualization of the representation space using representation atlases, and finding artifactual aspects in representation space. We demonstrated that the introduced Extreme-Activation distance is aligned with human judgment and is an interpretable metric for measuring the relationships between neural representations. In future work, we will apply the proposed solution broadly in the sciences, medicine, and other technical domains, such as NLP, where discovering artifacts and biases in the representations is of great value.

References

GitHub - weiaicunzai/pytorch-cifar100: Practice on CIFAR100— github.com. `https://github.com/weiaicunzai/pytorch-cifar100`, 2020. [Accessed 08-Jan-2023].

S. Agarwal, G. Krueger, J. Clark, A. Radford, J. W. Kim, and M. Brundage. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77:261–295, 2022.

S. Bach, A. Binder, G. on, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

H. Bao, L. Dong, and F. Wei. BEIT: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba. GAN dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

F. Bießmann, F. C. Meinecke, A. Gretton, A. Rauch, G. Rainer, N. K. Logothetis, and K.-R. Müller. Temporal kernel CCA and its application in multimodal neuronal data analysis. *Machine Learning*, 79(1):5–27, 2010.

A. Binder, K.-R. Müller, and M. Kawanabe. On taxonomies for multi-class image categorization. *International Journal of Computer Vision*, 99(3):281–301, 2012.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc., 2009.

E. Bisong and E. Bisong. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64, 2019.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

J. Borowski, R. S. Zimmermann, J. Schepers, R. Geirhos, T. S. Wallis, M. Bethge, and W. Brendel. Natural images are more informative for interpreting cnn activations than state-of-the-art synthetic feature visualizations. In *NeurIPS 2020 Workshop SVRHM*, 2020.

M. L. Braun, J. M. Buhmann, and K.-R. Müller. On relevant dimensions in kernel feature spaces. *The Journal of Machine Learning Research*, 9:1875–1908, 2008.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of data*, pages 93–104, 2000.

K. E. Brown and D. A. Talbert. Using explainable AI to measure feature contribution to uncertainty. In *The International FLAIRS Conference Proceedings*, volume 35, 2022.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

C.-A. Brust and J. Denzler. Not just a matter of semantics: The relationship between visual and semantic similarity. In *German Conference on Pattern Recognition*, pages 414–427. Springer, 2019.

V. Buhrmester, D. Münch, and M. Arens. Analysis of explainers of black box Deep Neural Networks for Computer Vision: A survey. *arXiv preprint arXiv:1911.12116*, 2019.

K. Bykov, M. M.-C. Höhne, A. Creosteanu, K.-R. Müller, F. Klauschen, S. Nakajima, and M. Kloft. Explaining Bayesian Neural Networks. *arXiv preprint arXiv:2108.10346*, 2021.

K. Bykov, A. Hedström, S. Nakajima, and M. M.-C. Höhne. NoiseGrad—enhancing explanations by introducing stochasticity to model weights. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6132–6140, 2022.

N. Cammarata, G. Goh, S. Carter, L. Schubert, M. Petrov, and C. Olah. Curve detectors. *Distill*, 5(6):e00024–003, 2020.

S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah. Exploring Neural Networks with activation atlases. *Distill.*, 2019.

C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin. This looks like that: deep learning for interpretable image recognition. *arXiv preprint arXiv:1806.10574*, 2018.

J. Da. A corpus-based study of character and bigram frequencies in chinese e-texts and its implications for chinese language instruction. In *Proceedings of the fourth International Conference on new technologies in teaching and learning Chinese*, pages 501–511. Citeseer, 2004.

B. Dai and D. Lin. Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*, 30, 2017.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

L. Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *CVPR 2011*, pages 1777–1784. IEEE, 2011.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113, 1989.

K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly. Explainable AI in industry. In *Proceedings of the 25th ACM SIGKDD International Conference on knowledge discovery & data mining*, pages 3203–3204, 2019.

S. Gautam, M. M.-C. Höhne, S. Hansen, R. Jenssen, and M. Kampffmeyer. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. *arXiv preprint arXiv:2108.12204*, 2021.

S. Gautam, A. Boubekki, S. Hansen, S. Salahuddin, R. Jenssen, M. Höhne, and M. Kampffmeyer. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35:17940–17952, 2022a.

S. Gautam, M. M.-C. Höhne, S. Hansen, R. Jenssen, and M. Kampffmeyer. Demonstrating the risk of imbalanced datasets in chest x-ray image-based diagnostics by prototypical relevance propagation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022b. doi: 10.1109/ISBI52829.2022.9761651.

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on Artificial Intelligence and Statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

D. Grinwald, K. Bykov, S. Nakajima, and M. M.-C. Höhne. Visualizing the diversity of representations learned by bayesian neural networks. *arXiv preprint arXiv:2201.10859*, 2022.

T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

R. Guidotti. Evaluating local explanation methods on ground truth. *Artificial Intelligence*, 291:103428, 2021.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 (Mar):1157–1182, 2003.

D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical Correlation Analysis: An overview with application to learning methods. *Neural computation*, 16:2639–64, 01 2005. doi: 10.1162/0899766042321814.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

A. Hedström, L. Weber, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne. Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations. *arXiv preprint arXiv:2202.06861*, 2022.

E. Hernandez, S. Schwettmann, D. Bau, T. Bagashvili, A. Torralba, and J. Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2021.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. Mong, S. Halabi, J. Sandberg, R. Jones, D. Larson, C. Langlotz, B. Patel, M. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, 07 2019.

P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson. On feature learning in the presence of spurious correlations. *arXiv preprint arXiv:2210.11369*, 2022.

P. Jackson. Introduction to expert systems. URL `https://www.osti.gov/biblio/5675197`. [Accessed 16-Feb-2023].

A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 702–712. PMLR, 26–28 Aug 2020. URL `https://proceedings.mlr.press/v108/jiang20a.html`.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 444–452, 2008.

A. Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

A. Laakso. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, 13, 05 2000. doi: 10.1080/09515080050002726.

S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2912–2920, 2016.

S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10:1096, 2019.

A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge discovery in data mining*, pages 157–166, 2005.

M. Le and S. Kayal. Revisiting edge detection in convolutional neural networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.

Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *Stanford CS 231N*, 7(7):3, 2015.

C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

Y. LeCun and I. Misra. Self-supervised learning: The dark matter of intelligence, 2021. URL `https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/`. [Accessed 08-Jan-2023].

Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

Z. Li, I. Evtimov, A. Gordo, C. Hazirbas, T. Hassner, C. C. Ferrer, C. Xu, and M. Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others, 2022.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation Forest. In *2008 8-th IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.

N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Res.*, 27:175–178, 1967.

S. Marcel and Y. Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1485–1488, 2010.

D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294, 1978.

L. McInnes, J. Healy, N. Saul, and L. Grossberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 09 2018. doi: 10.21105/joss.00861.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

G. Montavon, M. L. Braun, and K.-R. Müller. Kernel analysis of deep networks. *Journal of Machine Learning Research*, 12(78):2563–2581, 2011.

G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

A. Morcos, M. Raghu, and S. Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018.

A. Mordvintsev, N. Pezzotti, L. Schubert, and C. Olah. Differentiable image parameterizations. *Distill*, 2018. doi: 10.23915/distill.00012. https://distill.pub/2018/differentiable-parameterizations.

J. Mu and J. Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020.

A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, 2016.

A. Nguyen, J. Yosinski, and J. Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 55–76. Springer, 2019.

A. M. Nguyen, J. Yosinski, and J. Clune. Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 annual conference on genetic and evolutionary computation*, pages 959–966, 2015.

T. Nguyen, M. Raghu, and S. Kornblith. Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

T. Nguyen, M. Raghu, and S. Kornblith. On the origins of the block structure phenomenon in neural network representations. *arXiv preprint arXiv:2202.07184*, 2022.

C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

C. Olah, N. Cammarata, C. Voss, L. Schubert, and G. Goh. Naturally occurring equivariance in neural networks. *Distill*, 5(12):e00024–004, 2020.

D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam. Smooth Grad-Cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*, 2019.

T. Pedersen, S. Patwardhan, J. Michelizzi, et al. Wordnet:: Similarity-measuring the relatedness of concepts. In *AAAI*, volume 4, pages 25–29, 2004.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct): 2825–2830, 2011.

X. Qin and Z. Wang. Nasnet: A neuron attention stage-by-stage net for single image deraining. *arXiv preprint arXiv:1912.03151*, 2019.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for deep understanding and improvement. *arXiv preprint arXiv:1706.05806*, 2017.

M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

J. Ramsay, J. Berge, and G. Styan. Matrix correlation. *Psychometrika*, 49:403–423, 09 1984. doi: 10.1007/BF02306029.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

W. Samek, A. Binder, G. on, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2016.

W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 10 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7.

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

D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

K. O. Stanley. Compositional pattern producing networks: A novel abstraction of development. *Genetic programming and evolvable machines*, 8:131–162, 2007.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

E. Tjoa and C. Guan. A survey on Explainable Artificial Intelligence (XAI): Toward medical XAI. *IEEE transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2020.

B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor attacks. *Advances in Neural Information Processing Systems*, 31, 2018.

F. Trozzi, X. Wang, and P. Tao. Umap as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: a comparison study. *The Journal of Physical Chemistry B*, 125(19):5022–5034, 2021.

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

M. M.-C. Vidovic, N. Görnitz, K.-R. Müller, G. Rätsch, and M. Kloft. Opening the black box: Revealing interpretable sequence motifs in kernel-based learning algorithms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 137–153. Springer, 2015.

M. M.-C. Vidovic, N. Görnitz, K.-R. Müller, and M. Kloft. Feature importance measure for non-linear learning algorithms. *arXiv preprint arXiv:1611.07567*, 2016.

D. Wallis and I. Buvat. Clever hans effect found in a widely used brain tumour mri dataset. *Medical Image Analysis*, 77:102368, 2022.

Y. Wang. CIFAR-100 Resnet PyTorch 75.17% Accuracy — kaggle.com. `https://www.kaggle.com/code/yiweiwangau/cifar-100-resnet-pytorch-75-17-accuracy`, 2021. [Accessed 08-Jan-2023].

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

R. Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

D. Wu, J. Y. Poh Sheng, G. T. Su-En, M. Chevrier, J. L. Jie Hua, T. L. Kiat Hon, and J. Chen. Comparison between umap and t-sne for multiplex-immunofluorescence derived single-cell data from tissue sections. *BioRxiv*, page 549659, 2019.

Z. Wu and M. Palmer. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*, 1994.

K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF International Conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

Z. Yuan, Y. Yan, M. Sonka, and T. Yang. Large-scale Robust Deep AUC Maximization: A new surrogate loss and empirical studies on medical image classification. pages 3020–3029, 10 2021. doi: 10.1109/ICCV48922.2021.00303.

J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11): e1002683, 2018.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A Appendix

A.1 Evaluation

In the evaluation, two datasets were used: ILSVRC2012 (ImageNet 2012) Deng et al. (2009) and CIFAR-100 Krizhevsky (2009). For ImageNet, we employed eight different pre-trained models: ResNet18 He et al. (2016), AlexNet Krizhevsky et al. (2017), Inception V3 Szegedy et al. (2016), DenseNet 161 Huang et al. (2017), MobileNet V2 Sandler et al. (2018), ShuffleNet V2 Ma et al. (2018), obtained from the `torchvision-models` package Marcel and Rodriguez (2010), as well as ViT Dosovitskiy et al. (2020) and BEiT Krizhevsky et al. (2017), obtained from the `pytorch-vision-models` library Wightman (2019). For the CIFAR-100 dataset, we trained seven networks: ResNet 18, MobileNet V2, ShuffleNet V1, and V2, NASNet Qin and Wang (2019), SqueeeNet Iandola et al. (2016), and VGG 11 Simonyan and Zisserman (2014), using the `Pytorch-cifar100` GitHub repository git (2020), while the ResNet9 network was trained using a publicly available Kaggle notebook Wang (2021).

The semantic baseline distances between concepts for both datasets were obtained using the `NLTK` package Bird et al. (2009). There is a cross-connection between class labels and WordNet entities for ILSVRC2012, as the classes are inherently connected with WordNet synsets. For CIFAR-100, we manually connected the labels to synsets by matching class label names with WordNet synset names. For 98 classes, WordNet synsets were found. For the remaining two classes, “aquarium fish” and “maple tree”, WordNet synsets for “fish” and “maple” were used, respectively, due to the absence of a direct name match.

A.2 Experiments

A.2.1 Probing dataset

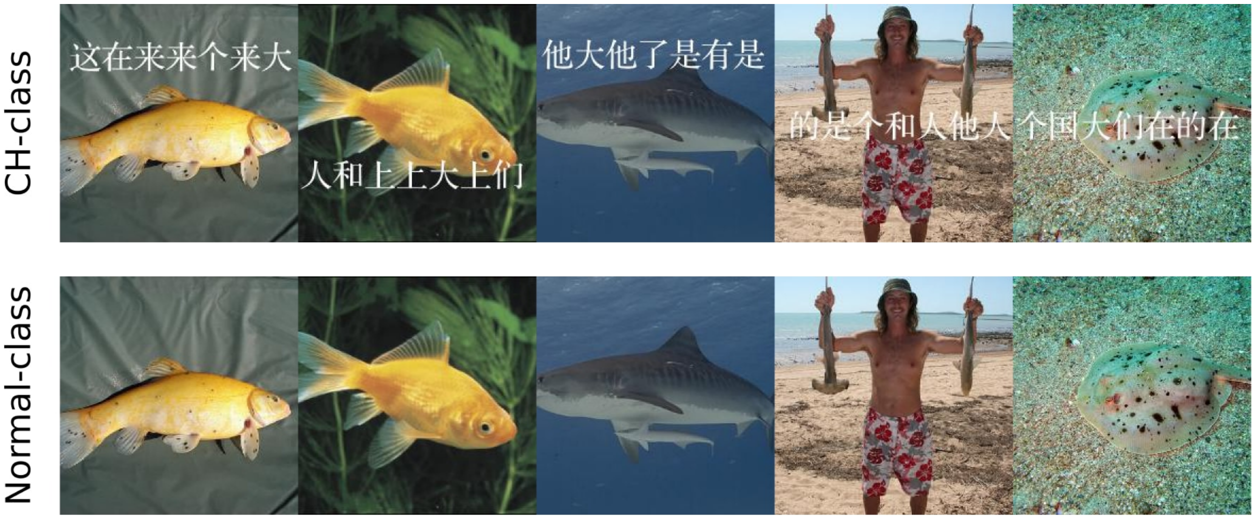


Figure 18: **Illustration of the Probing Dataset.** The figure depicts images from the probing dataset utilized to evaluate the representation’s capacity to distinguish between watermarked (CH) and non-watermarked (normal) images. The watermarked class images are identical to the normal class images, except for the addition of a random test string at a random location on the image.

To assess the ability of the identified representations to detect undesirable concepts, we created two probing datasets for the binary classification of Chinese and Latin text detection. We modified one class of images by adding specific watermarks while leaving the other class unchanged. We used a baseline dataset of 998 ImageNet images † to create 2 probing datasets (Chinese and Latin) by inserting random textual watermarks, as shown in Figure 18. For the Chinese-characters detection problem, the watermarks were generated by

---

† Images were obtained from `https://github.com/EliSchwartz/imagenet-sample-images`, with the exception of two images (of the class "carton" and "terrapin") that already exhibit watermarks.

randomly selecting 7 out of the 20 most commonly used Chinese characters Da (2004), and a similar process was followed using the English alphabet for the Latin text detection problem. The font size for all watermarks has been set to 30, while the image dimensions remain standard at 224 × 224 pixels. AUC ROC was used as the performance metric to evaluate the representations' ability to differentiate between watermarked and normal classes. The true labels provided by the two datasets were used, where class 1 represents images with a watermark and class 0 represents images without. We computed the scalar activations for all images from both classes for a specific neural representation and then calculated the AUC ROC classification score based on the differences in activations using the binary labels. A score of 1 indicates a perfect classifier, consistently ranking watermarked images higher than normal ones, while a score of 0.5 indicates a random classifier.

A.2.2 ImageNet ResNet18

In the following, we provide additional details on the ResNet18 He et al. (2016) experiment, discussed in the main paper. The model was downloaded from the Torchvision library Marcel and Rodriguez (2010) and s-AMS were generated with parameters $n = 5$ and $m = 500$ using the `DORA` package.

Figure 19 illustrates the cluster of reported representations in the average pooling layer of the model, specifically neurons 154, 129, 347, 489, 81, 439, and 282, along with the sensitivity of other neurons to Chinese watermarks. It can be seen that representations close to the reported cluster also exhibit sensitivity towards malicious concepts. For additional context, Figure 23 shows the natural Activation-Maximisation signals (n-AMS) for the reported representations, obtained using 1 million subsamples of the ImageNet 2012 train dataset. The presence of Chinese watermarks in the n-AMS further supports our hypothesis of the Clever-Hans nature of these representations.

To examine which output class logits may be compromised by CH behavior, we used the s-AMS of the reported neurons to obtain class predictions on these signals. Figure 25 shows several s-AMS for the reported representations along with the network's predictions for the corresponding data points. We observed that certain classes, such as "carton" (478), "apron" (411), "swab, swob, mop" (840), "monitor" (664), and "broom" (462) were frequently predicted with high scores. When we computed the s-AMS for selected output logits, we found similar Chinese patterns, similar to those observed in the reported neurons of the average pooling layer (see Figure 25). These results suggest that such artifacts learned by the network pose a potential threat to applications due to the network's tendency to classify images with added watermarks as belonging to one of these classes.

A.2.3 DenseNet 121

The DORA framework was employed to investigate the pre-trained DenseNet121 on the ImageNet dataset Huang et al. (2017). Specifically, attention was focused on the last layer of the feature extractor, which comprised 1024 channel representations. The study primarily examined two outliers detected by DORA: neuron 768 and neuron 427, along with some of their nearest neighbors in the EA distance. Following an analysis of the s-AMS for both neurons, specific symbolic patterns were observed, which were characteristic of character detectors. Neuron 768 was identified as a Chinese-character detector, while neuron 427 was identified as a Latin text detector. Figure 13 in the main paper and Figure 21 depict these neurons, along with their closest neighbors in EA distance, which exhibited similar properties. The hypothesis was further supported by visualizing the n-AMS across the ImageNet dataset, as demonstrated in Figure 22.

As mentioned in Section 4.4.2, we find that the outliers found by DORA are maintained during fine-tuning on another dataset, e.g. the CheXpert challenge. The CheXpert challenge benchmarks various deep learning models on the task of classifying multilabel chest radiographs and additionally provides human experts, e.g. radiologists, with performance metrics for comparison. The data set itself consists of 224,316 training, 200 validation, and 500 test data points. The current best approach in terms of AUC-ROC score uses an ensemble of five DenseNet121's Huang et al. (2017) that were pre-trained on the ImageNet dataset and fine-tuned by optimizing a special surrogate loss for the AUC-ROC score Yuan et al. (2021). The training code can be found in this public repository `https://github.com/Optimization-AI/LibAUC/`. We choose to finetune one DenseNet121 using this approach on a downsampled version of the CheXpert data with a resolution of 256x256x3. The converged model yields an AUC-ROC score of 87.93% on the validation dataset. Having

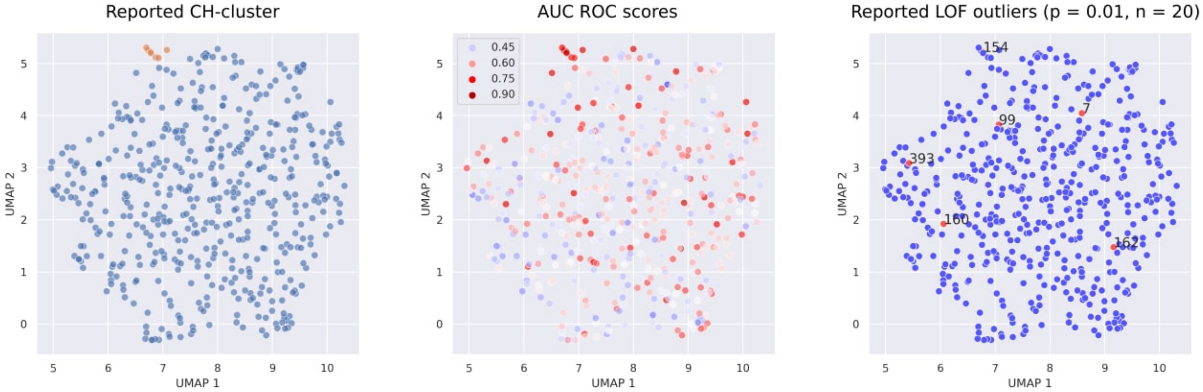


Figure 19: **Detailed illustration of the cluster of malicious representations found.** All of the figures illustrate the representation atlas of the average pooling layer of ResNet18, calculated using the DORA distance metric. From left to right: illustration of the reported Chinese detector cluster, the sensitivity of different representations for detecting Chinese watermarks, and a set of reported outliers among the representations using the LOF method. From the middle figure, it can be observed that the cluster of reported representations exhibits high sensitivity towards the artifactual concept of the desired task, and the closer the representations are to the cluster in the representation atlas, the more they are able to detect malicious concepts in the data.

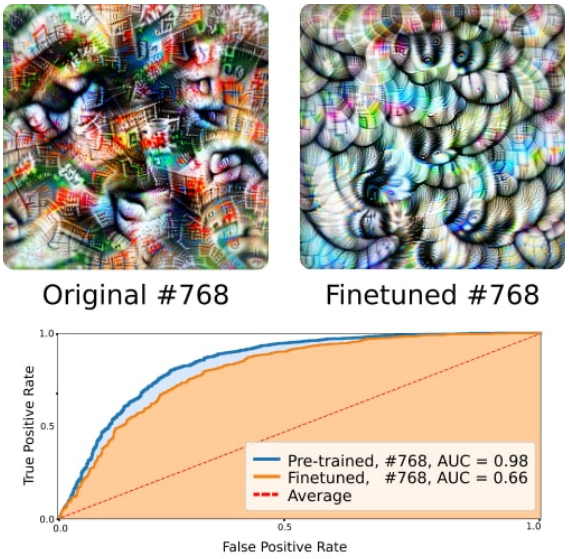


Figure 20: **Survived Chinese-characters detector**. Neuron 768 learns to detect Chinese logographic symbols during pre-training (top left) and does not unlearn this behavior during fine-tuning on the CheXpert dataset (top right). The AUC values of the neurons' activation on images corrupted with Chinese watermarks are still high after pre-training.

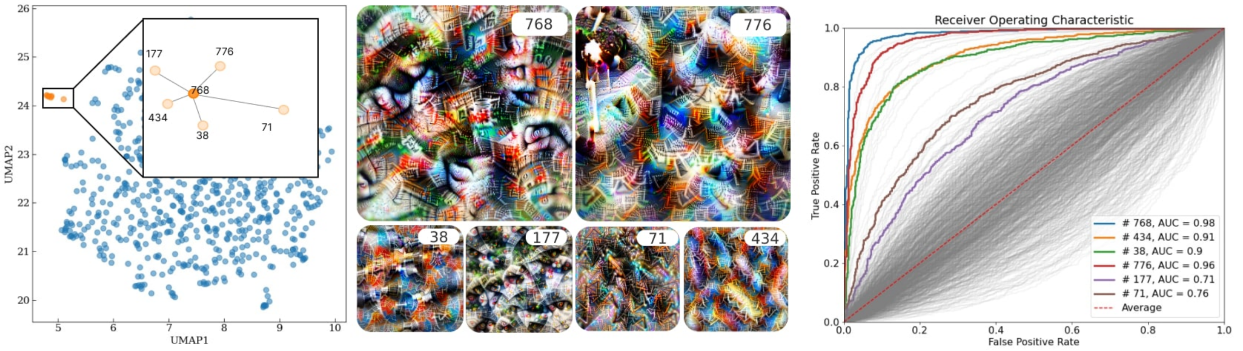
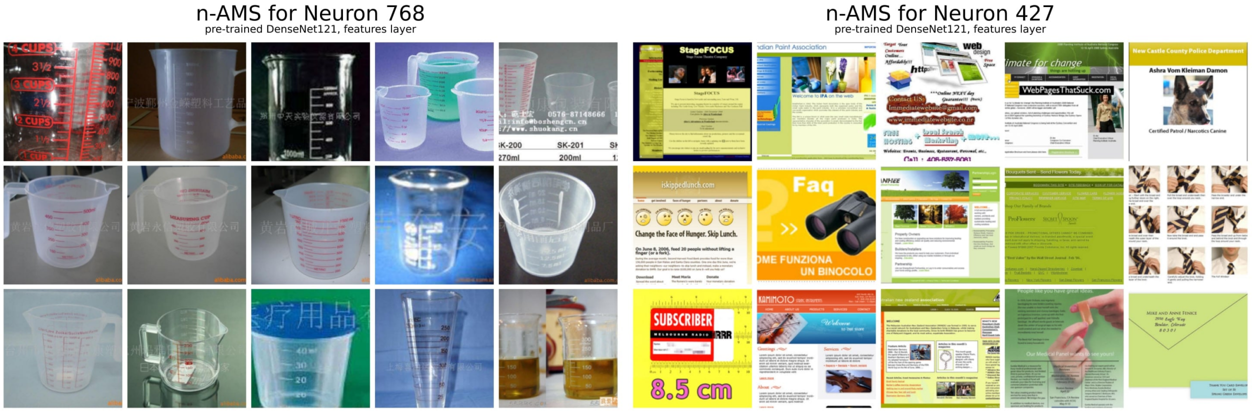


Figure 21: **DenseNet121 — Chinese-characters detector.** Applying DORA to the last layer of the feature extractor of DenseNet121 yields, among others, Neuron 768, which corresponds to the upper left of the 6 feature visualizations. From Neuron 768 as well as from its five closest neighbors (shown left), we can observe semantic concepts resembling Chinese logograms. The AUC values were computed using the channel activations on a data set that was corrupted with watermarks written in Chinese. As shown, the AUCs are high for the representation outliers found by DORA, compared to most of the other representations, which indicates that they indeed learned to detect Chinese logograms.


n-AMS for Neuron 768

pre-trained DenseNet121, features layer

n-AMS for Neuron 427

pre-trained DenseNet121, features layer

(a) Neuron 768

(b) Neuron 427

Figure 22: **n-AMS for different DenseNet121 neurons.** Illustration of the 15 n-AMS signals for the Chinse watermark detector (neuron 768) and the Latin text detector(neuron 427) in the “features” layer of DenseNet121.

the finetuned DenseNet121 and the outlier neuron 768 at hand we show the Feature Visualizations and the AUC-ROC curves for both the pre-trained and fine-tuned channel on an ImageNet subset with both uncorrupted and corrupted images with Chinese watermarks in Figure 20.

A.2.4 CLIP ResNet 50

The s-AMS for the CLIP ResNet 50 was computed using the same parameters as Goh et al. (2021) with the Lucent library. The number of optimization steps m was set to 512. The analysis was conducted on representations (channels) from the “layer 4” layer of the model. (Details on the s-AMS generation parameters can be found at `https://github.com/openai/CLIP-featurevis` and Lucent library at `https://github.com/greentfrapp/lucent`)

Star Wars representation

Table 4: **Clusters of CLIP "layer4" representations.** This table presents several interesting clusters and the indexes of the corresponding representations that were examined through manual inspection of the natural and synthetic AMS.

Cluster	*Representations*
Explicit/Pornographic	95, 255, 996, 1502, 2011
Money/Finance	785, 1376, 1731
Reptiles	230, 250, 417, 521, 652, 654, 694, 1008, 1234, 1301, 1340, 1364, 1445, 1598
Fish/Aquarium	1193, 1384
Asia-geographic	13,165, 235, 536, 780, 931, 1037, 1261, 1247,1423, 1669,1761,1874, 1898

Figure 3 shows the limitations of the n-AMS approach when the data corpus for analysis differs from the training dataset. Figure 26 further illustrates n-AMS collected from ImageNet and Yahoo Creative Commons Thomee et al. (2016) datasets via `OpenAI Microscope`. Text Feature Visualization Goh et al. (2021) supports our hypothesis that the model is a detector of Star Wars-related concepts.

Outlier representations

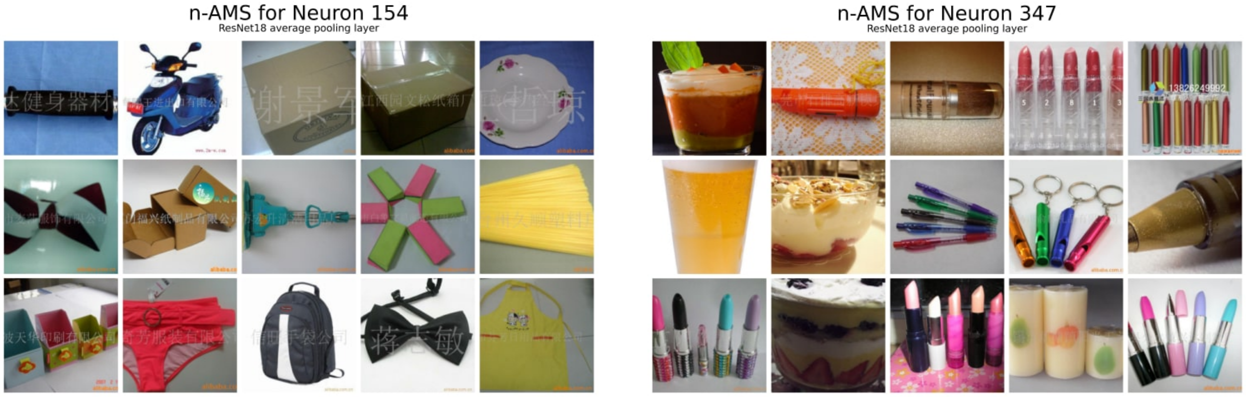
Analysis of the representations space of the CIP model yielded a number of potential candidates to be considered outlier representations, namely neurons 631, 658, 838, 1666, 1865, and 1896. In Figure 27 we illustrate 3 s-AMS signals, alongside n-AMS images, collected from the ImageNet dataset per each reported representation, collected using `OpenAI Microscope`. While it is hard to explain the anomalous nature of neurons 631, 658, 838, 1666, and 1896, we can clearly observe how different the concept of neuron 1865 is.

Clsuters of representations

We manually examined several distinctive classes of representations in "layer 4" of the CLIP model after computing the representation atlas for the channel representations. Table 4 summarizes the results of our analysis and shows interesting clusters found along with the associated neurons. Figure 28 shows synthetic and natural AMS, providing evidence for the assignment of neurons to their respective clusters.

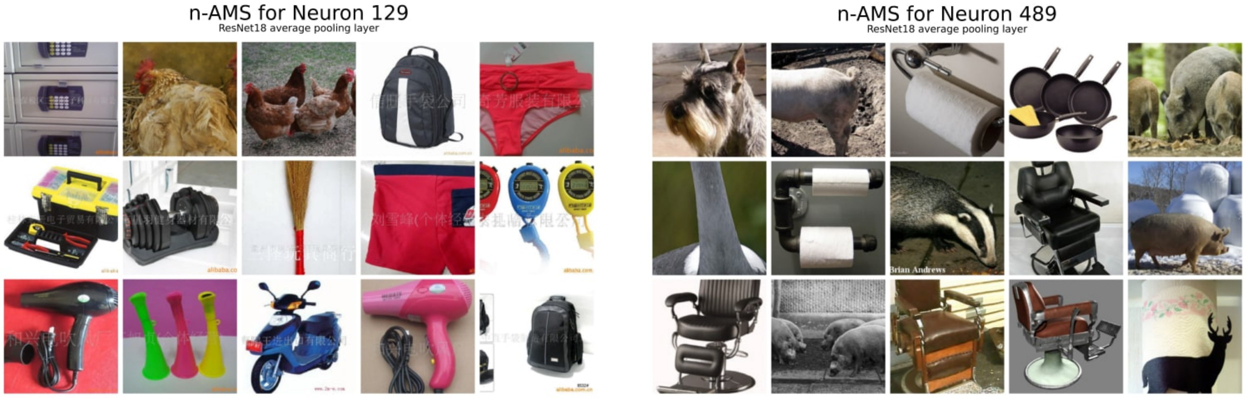
A.3 Experimental setup

All described experiments, if not stated otherwise, were performed on the Google Colab Pro Bisong and Bisong (2019) environment with the GPU accelerator.


n-AMS for Neuron 154
ResNet18 average pooling layer
n-AMS for Neuron 347
ResNet18 average pooling layer

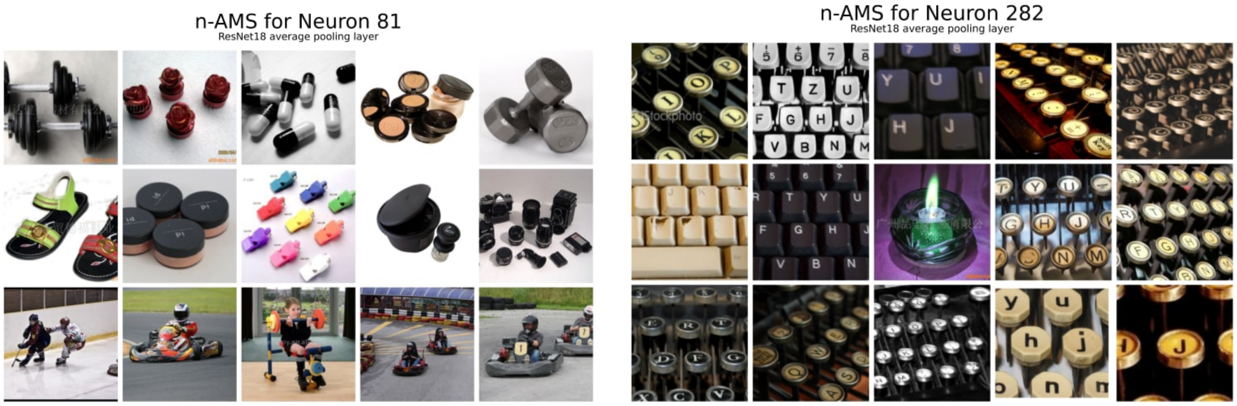
(a) Neuron 154

(b) Neuron 347


n-AMS for Neuron 129
ResNet18 average pooling layer
n-AMS for Neuron 489
ResNet18 average pooling layer

(c) Neuron 129

(d) Neuron 489


n-AMS for Neuron 81
ResNet18 average pooling layer
n-AMS for Neuron 282
ResNet18 average pooling layer

(e) Neuron 81

(f) Neuron 282

Figure 23: **n-AMS for different ResNet18 neurons, reported in the cluster of malicious representations.** The figure shows the 15 n-AMS signals for various neurons in the "avgpool" layer of the ResNet18 network, which were identified as being in the cluster of malicious representations. The signals were calculated using a subset of 1 million images from the ImageNet 2012 training dataset. It can be observed that among the top natural activation maximization signals, there are images of Chinese watermarks, supporting the hypothesis that these neurons have learned undesirable concepts.

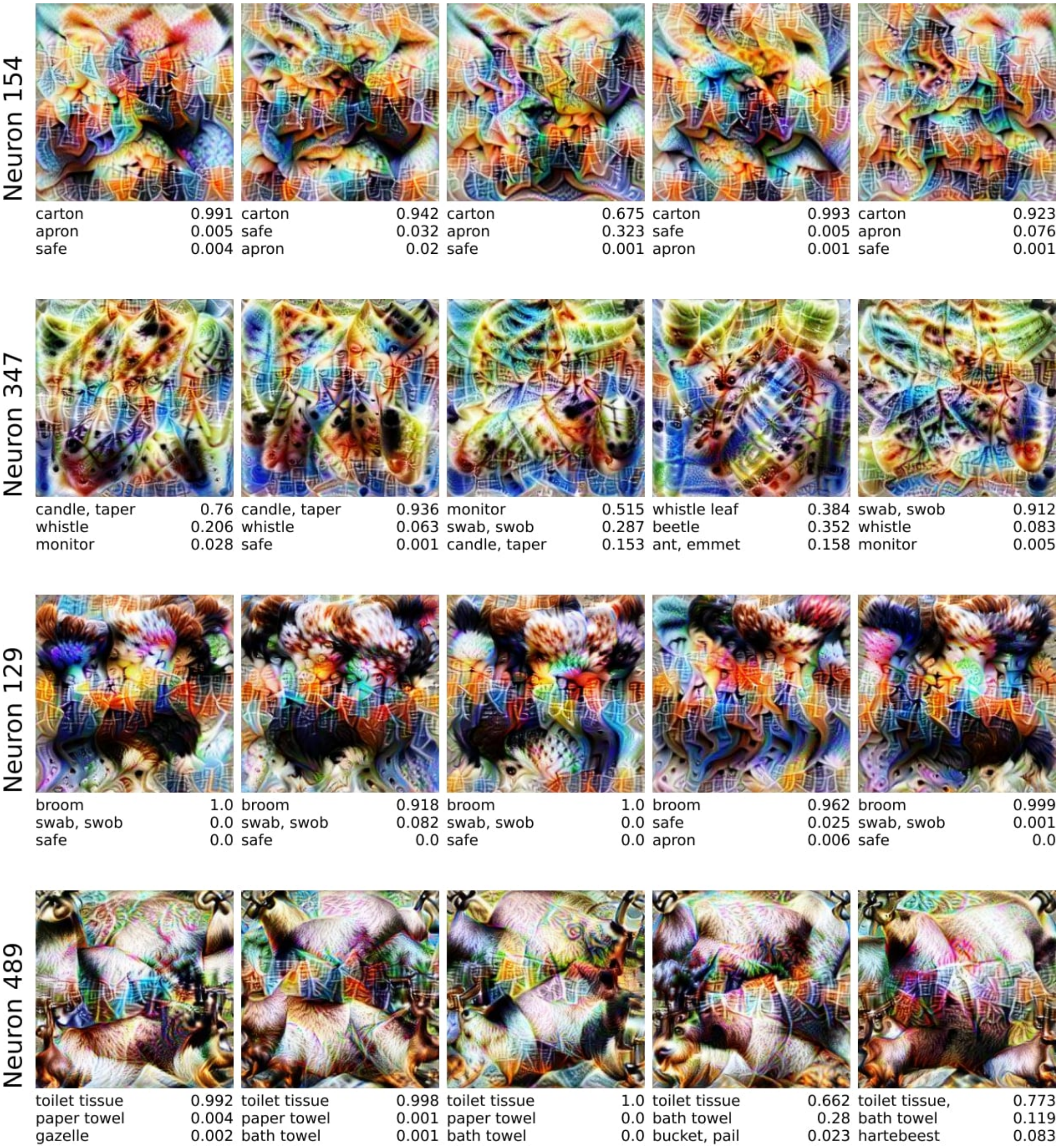

Neuron 154
carton 0.991 apron 0.005 safe 0.004
carton 0.942 safe 0.032 apron 0.02
carton 0.675 apron 0.323 safe 0.001
carton 0.993 safe 0.005 apron 0.001
carton 0.923 apron 0.076 safe 0.001
Neuron 347
candle, taper 0.76 whistle 0.206 monitor 0.028
candle, taper 0.936 whistle 0.063 safe 0.001
monitor 0.515 swab, swob 0.287 candle, taper 0.153
whistle leaf 0.384 beetle 0.352 ant, emmet 0.158
swab, swob 0.912 whistle 0.083 monitor 0.005
Neuron 129
broom 1.0 swab, swob 0.0 safe 0.0
broom 0.918 swab, swob 0.082 safe 0.0
broom 1.0 swab, swob 0.0 safe 0.0
broom 0.962 safe 0.025 apron 0.006
broom 0.999 swab, swob 0.001 safe 0.0
Neuron 489
toilet tissue 0.992 paper towel 0.004 gazelle 0.002
toilet tissue 0.998 paper towel 0.001 bath towel 0.001
toilet tissue 1.0 paper towel 0.0 bath towel 0.0
toilet tissue 0.662 bath towel 0.28 bucket, pail 0.023
toilet tissue, 0.773 bath towel 0.119 hartebeest 0.083

Figure 24: **s-AMS and model predictions for reported neurons in ResNet18.** Figure illustrates the s-AMS signals for four different reported neurons in the average pooling layer of ImageNet-trained ResNet18, along with the model's predictions for the top three classes with their respective softmax scores.

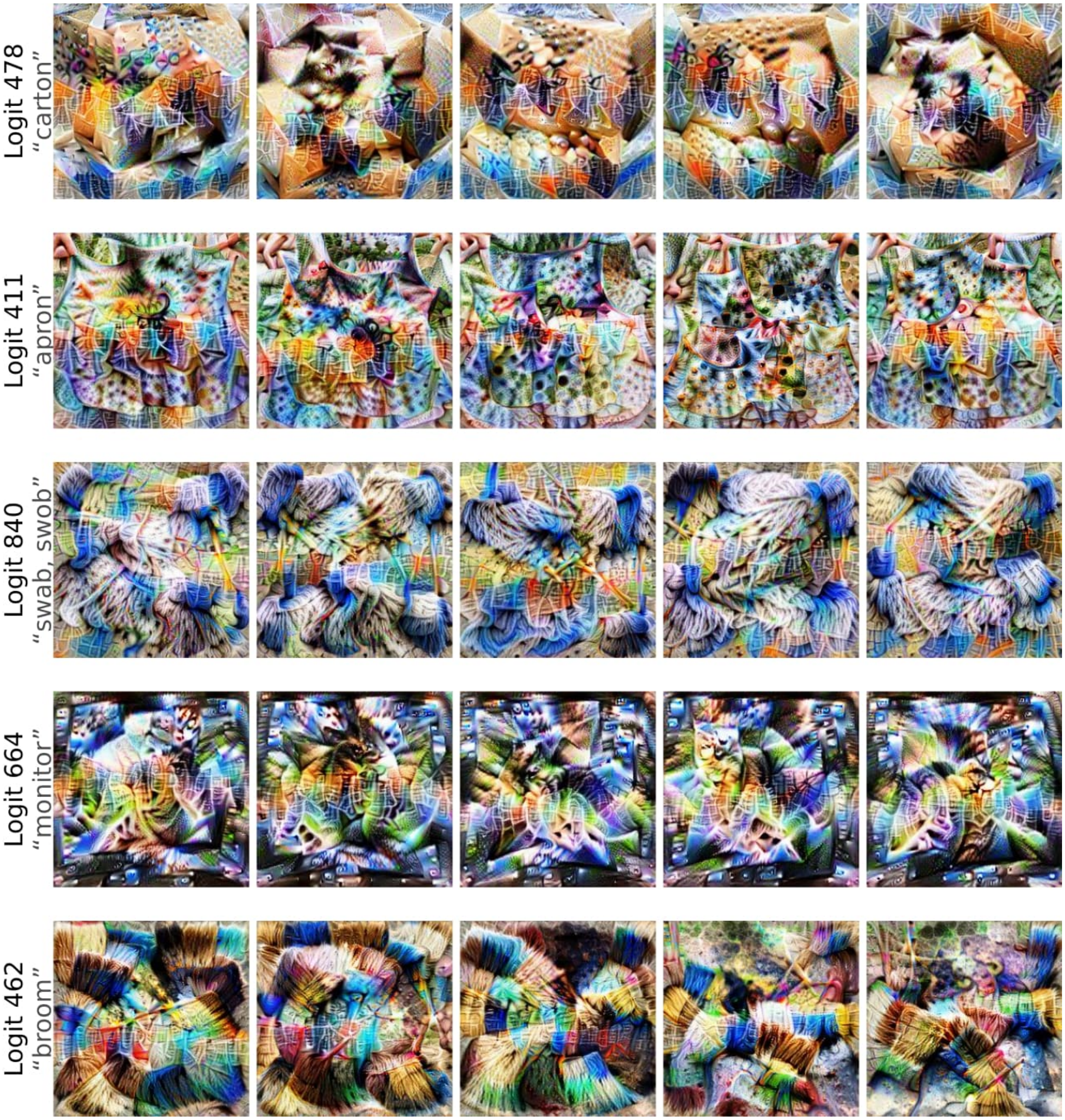
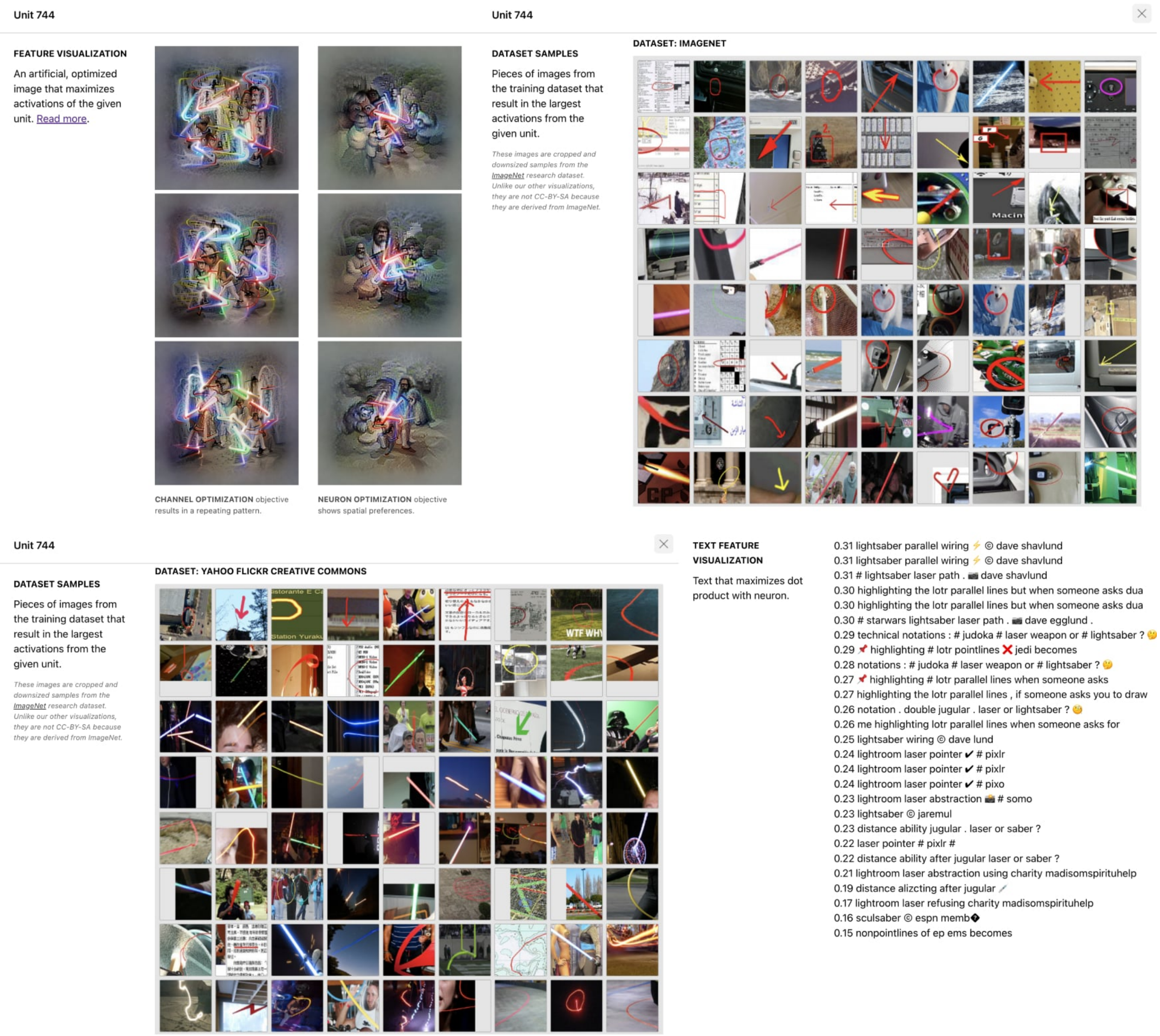


Figure 25: **s-AMS for several ResNet18 logits.** Figure shows s-AMS for the output logit representations of ResNet18. Similar to the reported neurons from the average pooling layer, the logits display logographic patterns, logographic patterns specific to Chinese character detectors, suggesting that these classes may be particularly affected by CH behavior.



Unit 744

FEATURE VISUALIZATION

An artificial, optimized image that maximizes activations of the given unit. Read more.

CHANNEL OPTIMIZATION objective results in a repeating pattern.

NEURON OPTIMIZATION objective shows spatial preferences.

Unit 744

DATASET SAMPLES

Pieces of images from the training dataset that result in the largest activations from the given unit.

These images are cropped and downsized samples from the ImageNet research dataset. Unlike our other visualizations, they are not CC-BY-SA because they are derived from ImageNet.

DATASET: IMAGENET

Unit 744

DATASET SAMPLES

Pieces of images from the training dataset that result in the largest activations from the given unit.

These images are cropped and downsized samples from the ImageNet research dataset. Unlike our other visualizations, they are not CC-BY-SA because they are derived from ImageNet.

DATASET: YAHOO FLICKR CREATIVE COMMONS

TEXT FEATURE VISUALIZATION

Text that maximizes dot product with neuron.

0.31 lightsaber parallel wiring ⚡ © dave shavlund
0.31 lightsaber parallel wiring ⚡ © dave shavlund
0.31 # lightsaber laser path . 📷 dave shavlund
0.30 highlighting the lotr parallel lines but when someone asks dua
0.30 highlighting the lotr parallel lines but when someone asks dua
0.30 # starwars lightsaber laser path . 📷 dave egglund .
0.29 technical notations : # judoka # laser weapon or # lightsaber ? 🤔
0.29 📌 highlighting # lotr pointlines ❌ jedi becomes
0.28 notations : # judoka # laser weapon or # lightsaber ? 🤔
0.27 📌 highlighting # lotr parallel lines when someone asks
0.27 highlighting the lotr parallel lines , if someone asks you to draw
0.26 notation . double jugular . laser or lightsaber ? 🧐
0.26 me highlighting lotr parallel lines when someone asks for
0.25 lightsaber wiring © dave lund
0.24 lightroom laser pointer ✔ # pixlr
0.24 lightroom laser pointer ✔ # pixlr
0.24 lightroom laser pointer ✔ # pixo
0.23 lightroom laser abstraction 📸 # somo
0.23 lightsaber © jaremul
0.23 distance ability jugular . laser or saber ?
0.22 laser pointer # pixlr #
0.22 distance ability after jugular laser or saber ?
0.21 lightroom laser abstraction using charity madisomspirituhelp
0.19 distance alizcting after jugular 💉
0.17 lightroom laser refusing charity madisomspirituhelp
0.16 sculsaber © espn memb�
0.15 nonpointlines of ep ems becomes

Figure 26: **CLIP ResNet Neuron 744.** The figure shows s-AMS and n-AMS for neuron 744 in the "layer 4" layer of the model, computed for 2 different data corpora. The observed signals and explanations from Text Feature Visualization confirm that the neuron can detect Star Wars-related concepts. Results obtained from `OpenAI Microscope`.

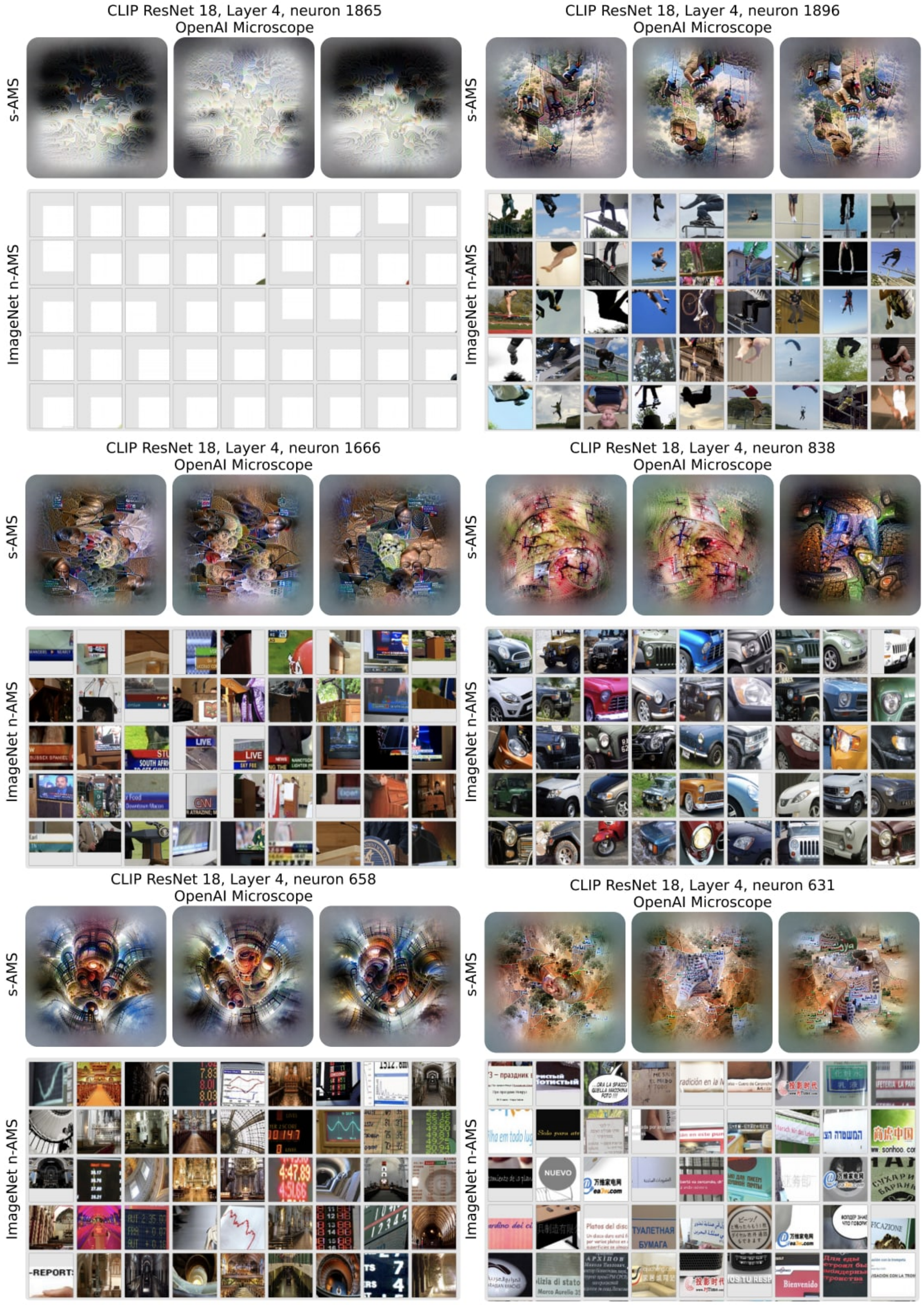

CLIP ResNet 18, Layer 4, neuron 1865
OpenAI Microscope
CLIP ResNet 18, Layer 4, neuron 1896
OpenAI Microscope
s-AMS
ImageNet n-AMS
CLIP ResNet 18, Layer 4, neuron 1666
OpenAI Microscope
CLIP ResNet 18, Layer 4, neuron 838
OpenAI Microscope
CLIP ResNet 18, Layer 4, neuron 658
OpenAI Microscope
CLIP ResNet 18, Layer 4, neuron 631
OpenAI Microscope

Figure 27: **s-AMS and n-AMS for reported outlier neurons.** Figure illustrates s-AMS and n-AMS for the reported outlier neurons in the "layer 4" layer of the CLIP ResNet 50 model, collected from `OpenAI Microscope`.

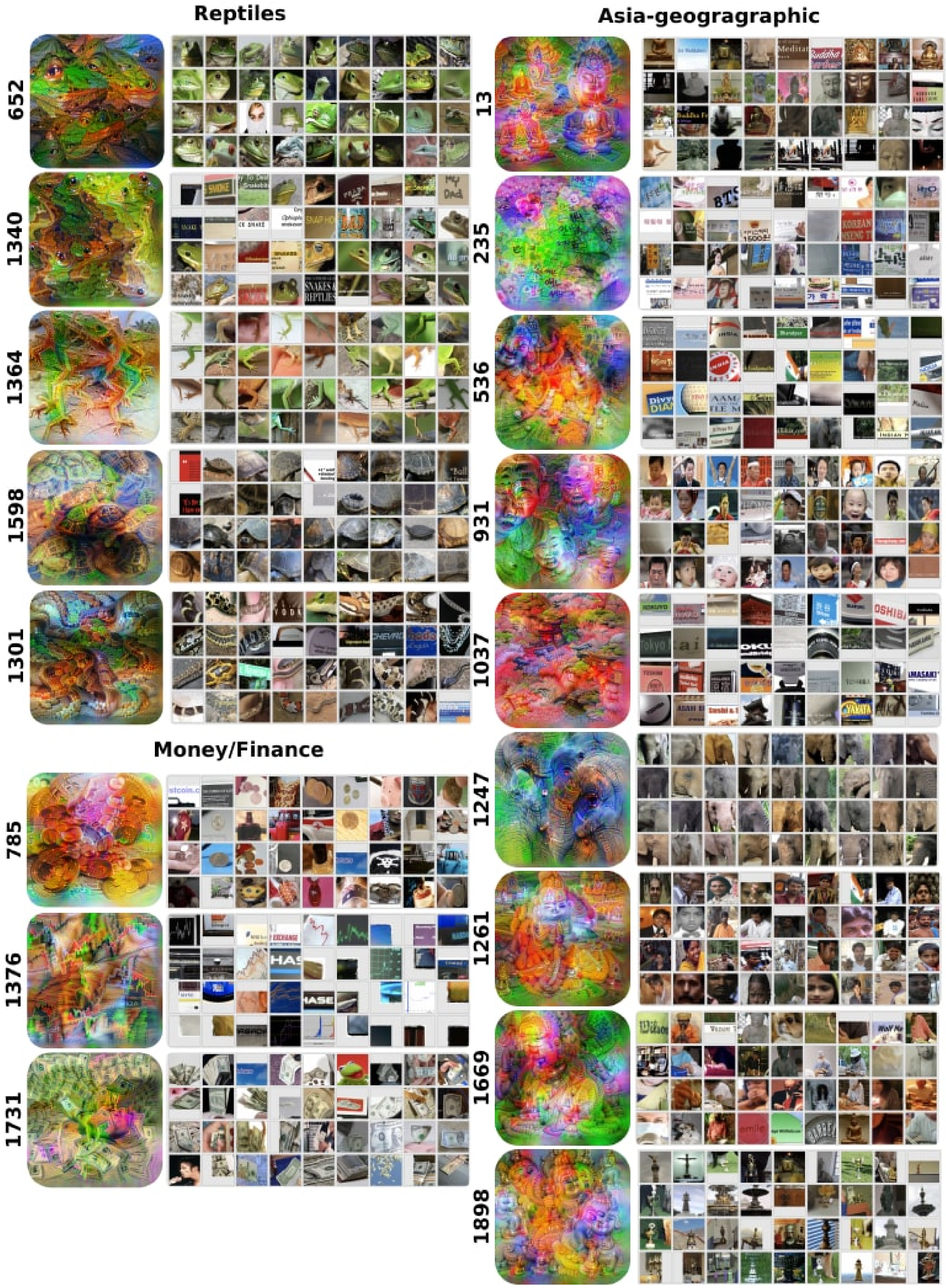


Figure 28: **s-AMS and n-AMS for the neurons in the reported clusters.** Figure shows s-AMS and n-AMS for representations assigned to different reported clusters. s-AMS were generated, while n-AMS (Activation-Maximization images from ImageNet dataset) were collected via `OpenAI Microscope`. Representations of explicit/pornographic content were excluded due to the presence of obscene images.