DORA: EXPLORING OUTLIER REPRESENTATIONS IN DEEP NEURAL NETWORKS

Kirill Bykov[1, 2, *] **& Mayukh Deb & Dennis Grinwald** [1, 3] **& Klaus-Robert Müller**[1, 3, 4, 5] **& Marina M.-C. Höhne** [1, 2, 3, 6, 7]
[1]Technische Universität Berlin, Machine Learning Group, 10587 Berlin, Germany
[2]Understandable Machine Intelligence Lab, ATB, 14469 Potsdam, Germany
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[4]Korea University, Department of Artificial Intelligence, Seoul 136-713, Korea
[5]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany
[6]Machine Learning Group, UiT the Arctic University of Norway, 9037 Tromsø, Norway
[7]Department of Computer Science, University of Potsdam, 14476 Potsdam, Germany
[*]Corresponding Author: KBykov@atb-potsdam.de

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 itself 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

Deep Neural Networks (DNNs) excel at complex tasks due to the rich and hierarchical representations they learn from input data Bengio et al. (2013), but recent works reported their susceptibility to learn harmful and undesired concepts like biases Guidotti et al. (2018); Jiang and Nachum (2020), Clever Hans effects Lapuschkin et al. (2019), and backdoors Anders et al. (2022). This issue is exacerbated by the semantic opacity of learned abstractions in modern DNNs, which are often trained in a self-supervised manner from potentially limitless amounts of data, making their decision-making strategies unpredictable.

To improve the understanding of the decision-making processes of complex machines and prevent the network from making biased or harmful decisions, it is crucial to explain what representations were learned by the model. One approach to gain insights into a model's prediction strategies is to analyze the relationships among its learned representations. 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

[*]PyTorch implementation of the proposed method could be found by the following link: https://github.com/lapalap/dora.

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.

2 Related Work

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 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" in the training corpus, resulting in *Clever Hans effects* Lapuschkin et al. (2019).

In order to address the concerns about the black-box nature of the 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. The popular approach to explain the purpose of individual components in Computer Vision models , referred to as Activation-Maximization, is to analyze the signals, that maximally activate a particular neuron or any sub-function within the network. These signals, which we will refer to as Activation-Maximization Signals (AMS), illustrate concepts that representations detect in the input data. While natural signals (n-AMS), obtained by selecting a "real" example from an existing data corpus in a data-aware manner, represent a potent tool for explicating the purpose of specific representations Borowski et al. (2020), it is challenging to acquire a subset of natural data that completely represents the diversity of concepts that the model may learn, given that contemporary machine learning models are often trained on large or closed-source datasets Radford et al. (2021). If the corpus used for n-AMS selection does not include all the possible concepts, results could be misleading. For example, as shown in Figure 1, an analysis solely based on natural signals may lead to inaccurate conclusions about the learned concept due to the absence of the true concept in the dataset. To overcome this limitation and be independent of data, Activation-Maximization signals can be generated synthetically using optimization procedures Erhan et al. (2009); Olah et al. (2017); Szegedy et al. (2013).

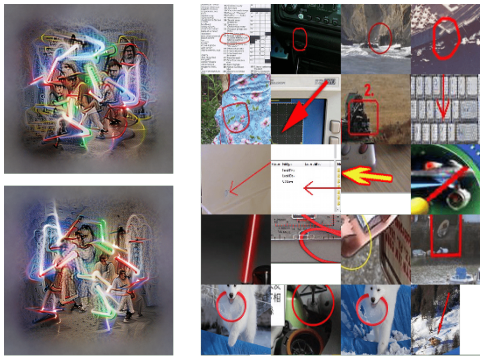

Figure 1: **Failing to explain "Star Wars" representation with n-AMS.** Comparison of the s-AMS (left) and n-AMS (right) collected from the ImageNet Deng et al. (2009) dataset for unit 744 in the last convolutional layer of the CLIP ResNet50 model. Due to the inaccessibility of the training dataset and 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.

3 DORA: Data-Agnostic Representation Analysis

We introduce the DORA framework for analyzing representation spaces of DNNs. It uses the proposed *Extreme-Activation* (EA) distance, a data-independent measure that estimates the similarity between neural representations by analyzing their s-AMS similarities. This distance is then leveraged to produce a visualization of the representation space, referred to as the *representation atlas*, which gives an overview of the learned representations' topological landscape. DORA can also detect outlier representations that may contain representations with undesirable and anomalous concepts.

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 $\mathbb{R}$.*

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. Additionally, we define a function F that maps the image from the input domain to the vector of activations of the individual representations within the layer:*

$$F(x) = [f_1(x), ..., f_k(x)] : \mathbb{D} \longrightarrow \mathbb{R}^k.$$

.

3.1 Extreme-Activation distance

For a specific neural representation f, a synthetic activation-maximization signal (s-AMS) s can be generated to visualize the concepts that maximize the function activation.

Definition 3 (s-AMS) *Let f_i be a neural representation. We define a synthetic Activation-Maximization signal s_i as a solution for the optimization problem*

$$s_i = \arg\max_s f_i(s).$$

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 used 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).

Because different random initializations in the parameter space lead to the convergence of s-AMS generation into different local solutions, the resulting signals can vary. This variability is similar to that observed in n-AMS sampling. To address this, we generated n s-AMS signals for each representation.

Given a layer $\mathcal{F}$ containing k neural representations defined on the domain $\mathbb{D} \subset \mathbb{R}^d$, where d is the dimension of the input of the model, Extreme-Activation distance could be formulated as follows:

Definition 4 *Let F be a layer with k neural representations and $f_i, f_j \in \mathcal{F}$ be individual representations. The Extreme-Activation (EA) distance between f_i and f_j is given by:*

$$d_{EA}(f_i, f_j) = \frac{1}{\sqrt{2}} \sqrt{1 - \cos\left(\frac{1}{n}\sum_{t=1}^{n} F(s_i^t), \frac{1}{n}\sum_{t=1}^{n} F(s_j^t)\right)},$$

where $\left[s_i^1, ..., s_i^n\right]$ and $\left[s_j^1, ..., s_j^n\right]$ are collections of s-AMS for f_i, f_j, respectively, $\cos(A, B)$ is the cosine similarity between vectors A and B, and n is the parameter of the method, controlling the number of generated s-AMS per representation.

The Extreme-Activation distance between two representations is determined by how similar their s-AMS are. If two representations encode similar concepts, their s-AMS will likely be visually similar. For instance, the s-AMS generated by representations for cat and tiger detectors are expected to show a visual similarity due to inherent similarities between the classes. Conversely, the s-AMS generated by representations for cat and car detectors will likely be visually dissimilar. To quantify s-AMS similarity, our approach uses the network itself to generate embeddings and compares them using cosine similarity. We avoid using external models to maintain approach transparency and prevent introducing additional opaqueness. A cosine value close to 1 indicates common activation-maximization concepts, while a cosine value close to 0.5 indicates independent concepts.

In the EA distance, the embeddings of s-AMS signals are obtained from all representations in the layer F, implying that the distance between two representations $f_i, f_q \in \mathcal{F}$ depends on how all the representations in F respond to their s-AMS. This approach, in which embeddings are collected across the entire layer, is referred to as *layer-wise* (l-w). An alternative approach would be to collect embeddings only from the representations for which the distance is being calculated, i.e. when computing the distance between f_i and f_q, embeddings $A_i, A_q \subset \mathbb{R}^{n\times 2}$ of s-AMS are collected only from the two representations themselves, independent of the other representations in the layer. This approach is referred to as *pair-wise* (p-w), and we compare both approaches in the evaluation section. Unless otherwise specified, the default DORA method uses layer-wise EA distance computation.

3.2 Representation atlas and Outlier detection

The calculation of EA distances between the k neural representations in layer F results in a distance matrix $D \in \mathbb{R}^{k\times k}$. As shown in subsequent sections, this matrix effectively captures the semantic distance between the concepts encoded in the representations. Inspired by Carter et al. (2019), to visually examine the topological landscape of learned representations and identify clusters of semantically similar representations, we pass a pre-computed distance matrix to the UMAP dimensionality reduction algorithm McInnes et al. (2018) to output a two-dimensional map of the representation space, referred to as the *representation atlas*. By utilizing the distance matrix between representations, we can also identify outlier representations that deviate semantically from the majority. We further demonstrate in our analysis that such representations often learn unnatural and unintended concepts.

4 Evaluation

One of the essential considerations to address when comparing various distance metrics between neural representations is its alignment with human judgment. If the underlying concepts of the representations differ semantically from a human perspective, we would expect 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.

We calculated the EA distances in a pair-wise and layer-wise manner between the output logits of image classification networks trained on two widely used computer vision datasets, ILSVRC2012 Deng et al. (2009) and CIFAR-100 Krizhevsky (2009). The baseline distances were obtained by mapping the classification labels to entities in the WordNet taxonomy database Miller (1995). The semantic baseline distances between entities were computed using two distance measures: the *Shortest-Path* distance, which is the length of the shortest path connecting the two entities in the taxonomy, and the *Leacock-Chodorow* distance, a version of the shortest-path distance that scales with the taxonomy depth Leacock and Chodorow (1998). To evaluate the alignment, we used a Mantel Test Mantel (1967), which is commonly used in ecology and evolutionary biology to measure the correlation between two distance matrices. Detailed results and figures could be found in Appendix.

4.1 Hyperparameter selection

The EA metric's quality depends on hyperparameter selection for the s-AMS generation method. We studied the impact of two hyperparameters, n (number of s-AMS samples per representation) and m (number of optimization epochs), on EA distances. We used a ResNet-18 model trained on ImageNet and varied n from 1 to 100 and m from 1 to 500. Our comparison results in Figure 7 show the effect of hyperparameters on the Mantel test statistic ρ for both layer-wise and pair-wise settings, using

the shortest-path semantic baseline. We observe that m has a greater influence on coherence with ground truth than n, achieving optimal alignment with $m = 500$, and that the layer-wise computation outperforms pair-wise in terms of the correlation score to the semantic baseline.

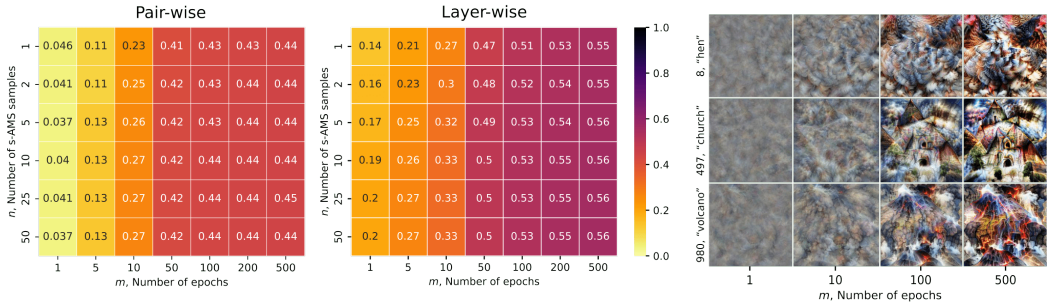

Figure 2: **The influence of hyperparameter selection on the correlation with semantic distance.** The Mantel test results for the comparison of pair-wise and layer-wise EA metrics with the shortest-path semantic distance baseline are shown on the left, and the impact of the number of optimization steps on s-AMS generation is depicted on the right for 3 ImageNet logit representations obtained from ResNet18 He et al. (2016). The results (the higher the score, the better) indicate that the number of epochs for s-AMS generation, controlled by the parameter m, has a greater impact than the number of samples per representation n. The layer-wise EA also outperforms the pair-wise metric, implying that incorporating a higher number of descriptors in the s-AMS embedding is beneficial.

4.2 Alignment with ground truth

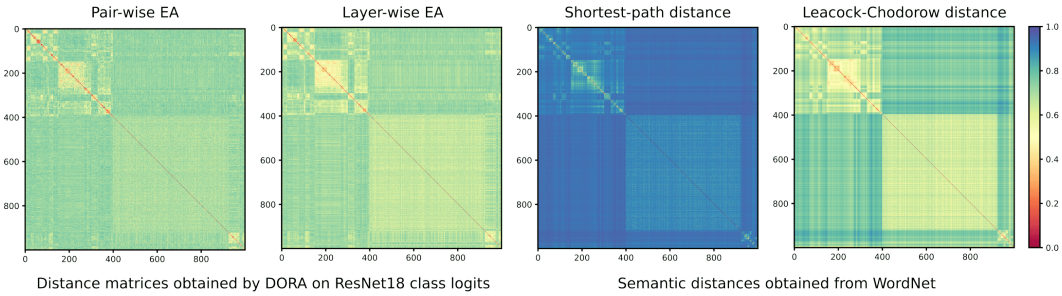

Figure 3: **Similarity between EA distances and taxonomy baseline distances.** The EA distances ($n = 50, m = 500$) between 1000 logit representations from the output layer of ImageNet-trained ResNet18, both in pair-wise and layer-wise fashion, are illustrated on the left, with semantic baseline distances between labels obtained from the WordNet taxonomy illustrated on the right. We can observe a connection between the baseline semantic distances and the distances obtained via the Extreme-Activation metric in a data-agnostic manner by the visual similarity of distance matrices .

In this experiment, we assess the alignment of the EA distance with the semantic baseline, across different datasets and architectures. We used eight different architectures for ImageNet and CIFAR100 and performed the Mantel test between obtained EA distance matrices between representations in the output logit layers (computed with optimal hyperparameters, both pair-wise and layer-wise) and semantic baseline. Results (shown in Table 1) illustrate statistical significance for all the experiments and, again, better alignment of layer-wise EA distance (up to 0.56 correlation) compared to the pair-wise. Figure 8 illustrates the similarity between data-agnostic EA procedure and semantic baselines for the ImageNet pre-trained ResNet18 He et al. (2016) output logit layer.

5 Experiments

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. We

applied the DORA framework to investigate representations in the feature extractor layer of 3 popular pre-trained models, namely ResNet-18 He et al. (2016), MobileNetV2 Sandler et al. (2018) and DenseNet-121 Huang et al. (2017), that 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).

By utilizing Local Outlier Factor (LOF) Breunig et al. (2000) outlier detection method with the contamination parameter $p = 0.01$ (corresponding to the top 1% of representations), we observed that in all 3 networks, reported outliers appear to function as watermark detectors, capable of detecting either Chinese or Latin text patterns. Since ImageNet does not have a category for watermarks, these representations may be considered as Clever-Hans artifacts and can lead to undesired decision-making Lapuschkin et al. (2019); Anders et al. (2022). To confirm the watermark-detecting abilities of these representations, we generated two binary classification datasets that distinguish between watermarked and non-watermarked images, and we evaluated sensitivity using AUC ROC. To ensure the detection of characters and not specific words or phrases (unlike CLIP models Goh et al. (2021)), we used probing datasets with randomly generated characters. Our results indicate that not only the reported outliers but also neighboring representations in EA distance are impacted by artifactual behavior.

Figure 4: **Cluster of Clever-Hans representations in the ResNet-18 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, which is the representation with the highest outlier score, 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.

The results of the analysis of the ResNet-18 average pooling layer are shown in Figure 9, illustrating the cluster of Clever-Hans representations found, along with their s-AMS and AUC ROC performance on the binary classification problem. 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)).

A similar analysis of the DenseNet-121 feature extractor layer (channel representation was used for the analysis by averaging the activation maps), yielded neurons 768 and 427, to be a Chinese and Latin detector, respectively. 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 DenseNet-121 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, neuron 427, retained its original semantic concept and still was reported as an outlier in the fine-tuned model. 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. The results indicate that the Clever-Hans effect persisted after fine-tuning, possibly due to small Latin text characters in the dataset.

6 Discussion and Conclusion

The proposed DORA framework is simple and data-independent, allowing the inspection of any trained neural network without access to the training data. It employs the self-explanatory capabilities of Computer Vision networks to estimate distances within the network itself. We found that representations that deviate from desired decision-making strategies typically manifest themselves as outliers in the representation space, which we can identify by our proposed method DORA. This can be used to analyze datasets for sensitivity to certain artifact concepts encoded in these outlier representations. We also demonstrated that such outlier representations can persist after transfer learning, highlighting the potential for pre-trained models in safety-critical areas to contain undesired behavior even after fine-tuning on a new dataset.

While we have demonstrated the broad applicability of DORA, there are still some challenges that need to be addressed. The main limitation is the assumption that malicious or CH-behaviour in the representations is not systematic. In other words, DORA may not be able to detect infected representations if this behavior is prevalent across a large number of representations, as it would no longer be perceived as anomalous behavior. Another limitation is the potential semantic multimodality of representations Goh et al. (2021), which DORA attempts to mitigate by computing several s-AMS per representation. However, this may not be sufficient to uncover all of the concepts that a representation is capable of detecting.

In summary, we showed the functionality and usefulness of DORA for finding artifactual aspects in representation space for both controlled and real-world environments. Additionally, the representation atlas, computed by DORA was shown to be a powerful tool for understanding the topological landscape of representations, allowing to visually illustrate the semantic similarities between representations. Note, that although we have introduced DORA as an automatic tool, if necessary, the final decision on the degree of harmfulness of any outlier representations needs to be subject to human scrutiny. In this sense, DORA substantially facilitates human intervention and reduces it to a minimum, however, for safety-critical applications human supervision will still be necessary. 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.

G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International joint Conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

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 RELATED WORKS

In order to address the concerns about the black-box nature of the 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. (2022b); 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 scope 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* level, where the aim is to explain the prediction strategy learned by the machine across the population 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. (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 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 (e.g., 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. Methods such as Network Dissection Bau et al. (2017) and Compositional Explanations of Neurons Mu and Andreas (2020) 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 representation, while 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). Similar approaches were used for the analysis of representations in Generative adversarial networks (GANs) Bau et al. (2018).

A.1.1 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 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" 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

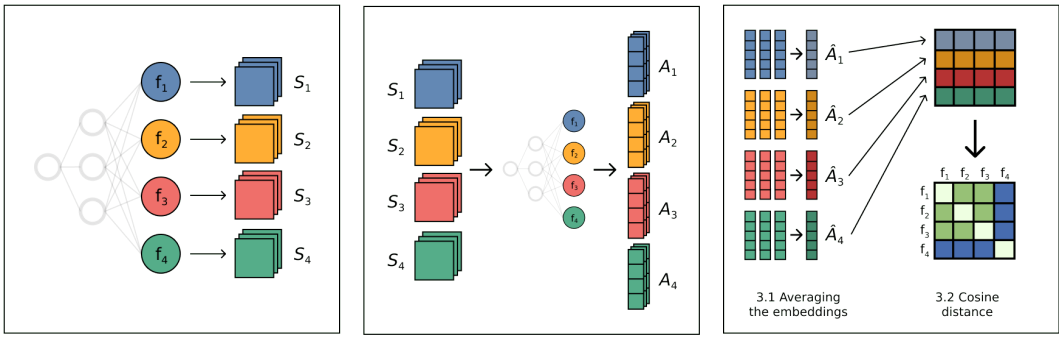

Figure 5: **Estimation of the (layer-wise) Extreme-Activation distance:** 1. Generation of s-AMS for a set of neurons (left), 2. Collection the embeddings of the generated s-AMS of Step 1 (middle), 3. Finding semantic outliers in the activation space (right).

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.

A.1.2 Comparison of representations

The study of representation similarity in DNN architectures is a topic of active research. Many methods for comparing network representations have been applied to various 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 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 activations or by calculating mutual information Li et al. (2015). However, these methods lack the capability to perform a comparison between individual neurons and their efficacy is contingent upon the presence of training data.

A.2 Extreme-Activation distance

Given a layer F containing k neural representations defined on the domain $\mathbb{D} \subset \mathbb{R}^d$, where d is the dimension of the input of the model, computation of Extreme-Activation distance could be summarized in three steps, as illustrated in Figure 5:

1. **Generation of s-AMS**
   For each neural representation $f_i \in \mathcal{F}, \forall i \in [1, ...k]$, a collection of n s-AMS is generated:
   $$S_i = \left[s_i^1, ..., s_i^n\right], \forall i \in [1, ...k],$$
   where $s_i^j \subset \mathbb{R}^d$ is the j-th s-AMS sample for representation f_i. The parameter n controls the number of generated signals, which are generated non-deterministically.
2. **Collection of s-AMS embeddings**
   After s-AMS sets $S_i \subset \mathbb{R}^{n \times d}, \forall i \in [1, ..., k]$ are collected for all representations in F, signals are successively inferenced by the model, and their corresponding activations across representations in F are saved. For each set of s-AMS signals S_i we obtain a collection of vectors
   $$A_i = \left[F(s_i^1), ..., F(s_i^n)\right] \subset \mathbb{R}^{n \times k}, \forall i \in [1, ...k],$$

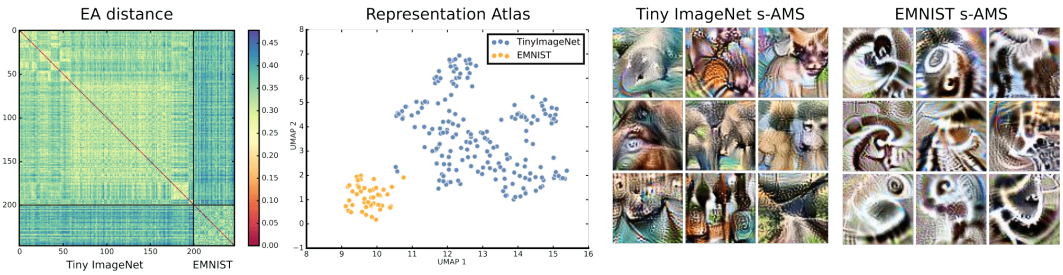

Figure 6: **Illustration of the performance of DORA on a toy example.** From left to right: EA distance metric between the output logits of the network trained on the combined dataset, UMAP visualization of the representational space, and s-AMS for Tiny ImageNet logits and for EMNIST logits. Visual differences between s-AMS for ImageNet and EMNIST classes can be observed and measured by the activations of the representations, resulting in a clear and visible cluster structure of the distance matrix.

where $F(s_i^j)$ correspond to the k-dimensional embedding of the $j-th$ s-AMS sample generated for representation f_i.

3. **Computing cosine similarity between average embeddings**
   Finally, for every representation $f_i \in \mathcal{F}, \forall i \in [1, ...k]$, we average the embeddings, corresponding to the generated signals

$$\hat{A}_i = \frac{1}{n}\sum_{j=1}^{n} F(s_i^j) \subset \mathbb{R}^k, \forall i \in [1, ..., k],$$

   resulting in a single embedding vector for each representation in F. The EA distance between two representations $f_i, f_q \subset \mathcal{F}$ is then defined as the square root of the cosine distance between the k-dimensional vectors of the averaged embeddings:

$$d_{EA}(f_i, f_q) = \frac{1}{\sqrt{2}}\sqrt{1 - \cos\left(\hat{A}_i, \hat{A}_q\right)},$$

   where $\cos(A, B)$ is the cosine similarity between vectors A and B.

A.3 Toy example

To demonstrate the capabilities of our proposed framework, we conducted a simple 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 EMNIST dataset Cohen et al. (2017), an extension of the MNIST dataset.

- Tiny-ImageNet Le and Yang (2015) is a small version of the ImageNet dataset, consisting of a subset of images from the ImageNet dataset Deng et al. (2009) and is often used as a more computationally efficient alternative for testing and developing new image classification algorithms. The Tiny-ImageNet dataset contains 200 classes, with 500 images in each class, for a total of 100,000 images. The images are downscaled to 64 x 64 pixels in size and are labeled with one of the 200 class labels.
- EMNIST Cohen et al. (2017) is a dataset of handwritten characters and digits that is widely used in machine learning and computer vision research. The EMNIST dataset was developed as an extension of the original MNIST Deng (2012) dataset, which only contained images of digits, and has proven to be a valuable resource for researchers working in the fields of pattern recognition and machine learning. For this particular application, we employed 47 different classes from the "balanced" split of the dataset, obtained from `torchvision` library Marcel and Rodriguez (2010). For each of the classes, images were resized to 64 x 64 pixels with 3 color channels to share the same dimensions with TinyImageNet, and the number of images per class was set to 200.

The results obtained by DORA are shown in Figure 6. Due to the visual differences, incorporated in s-AMS of different output representations, our proposed framework is able to clearly distinguish between logits corresponding to classes from the two different datasets, which we can observe from the visible block structure of the computed EA distance matrix and from the representation atlas. DORA is based on the network's ability to perceive self-generated s-AMS and we can observe a clear difference between the patterns of s-AMS for Tiny ImageNet classes, containing high-level natural concepts, and the more data-specific patterns for EMNIST classes, illustrating the network's perception of white-on-black handwritten digits and letters.

The ResNet18 model was utilized for training, initialized with ImageNet pre-trained weights, and the output layers were modified to accommodate the altered input size and the number of classes. The dataset, comprising 109400 images, was divided into training (103930 images), testing (2735 images), and validation (2735 images) sets. The network was trained for 50 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 256, resulting in a model that achieved an accuracy of 0.552 on the validation set.

A.4 Evaluation

To demonstrate the correctness of the proposed distance metric, we conducted a quantitative comparison between the EA distance metric and the semantic baseline, in the scenario, where the baseline for the distance between representations is available. For this purpose, we calculated the EA distances in both a pair-wise and layer-wise manner between the output logits of image classification networks trained on two widely used computer vision datasets, ILSVRC2012 Deng et al. (2009) and CIFAR-100 Krizhevsky (2009). Baseline 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. The baseline semantic distances between entities from the WordNet database were computed using two distance measures:

- **Shortest-Path distance:** the distance between two classes is determined by the length of the shortest path that connects the two entities in the taxonomy.

$$d_{SP}(c_i, c_j) = l(c_i, c_j), \tag{1}$$

  where $l(c_i, c_j)$ is the length of the shortest path between classes c_i, c_j.
- **Leacock-Chodorow distance:** a version of the shortest-path distance with additional scaling by the taxonomy depth Leacock and Chodorow (1998):

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

While it has been reported that taxonomy-based approaches can be suboptimal to taxonomy-free methods Binder et al. (2012), the visual similarity of concepts is strongly coherent with the semantic similarity. In recent years, the relationship between visual and semantic similarities in Computer Vision was studied and a significant linkage has been shown Brust and Denzler (2019). In Deselaers and Ferrari (2011), the authors confirm the assumption that for the ImageNet dataset visual similarity between categories grows with semantic similarity.

In order to test whether the proposed Extreme-Activation distance metric is able to preserve the semantic distance between representations, we performed a comparison between distance matrices by a 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.

A.5 Hyperparameter selection

The quality of distances obtained with the EA metric depends on the selection of hyperparameters. In the following experiment, we sought to understand the influence of two hyperparameters: n,

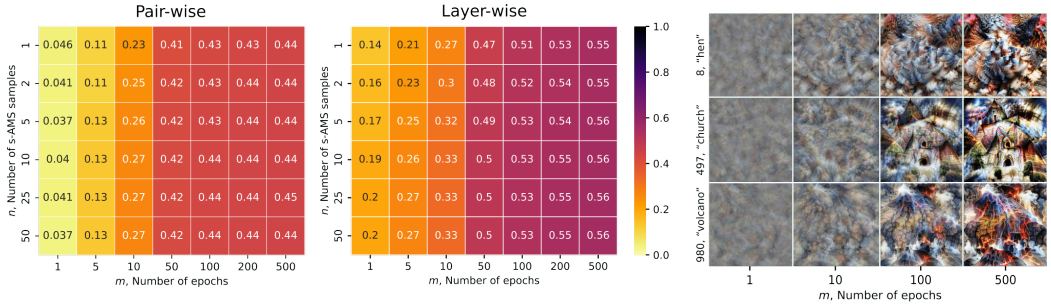

Figure 7: **The influence of hyperparameter selection on the correlation with semantic distance.** The Mantel test results for the comparison of pair-wise and layer-wise EA metrics with the shortest-path semantic distance baseline are shown on the left, and the impact of the number of optimization steps on s-AMS generation is depicted on the right for 3 ImageNet logit representations obtained from ResNet18 He et al. (2016). The results (the higher the score, the better) indicate that the number of epochs for s-AMS generation, controlled by the parameter m, has a greater impact than the number of samples per representation n. The layer-wise EA also outperforms the pair-wise metric, implying that incorporating a higher number of descriptors in the s-AMS embedding is beneficial.

which controls the number of s-AMS samples generated per representation, and m, which controls the number of optimization steps (epochs) for the Feature Visualization method. To conduct this experiment, we utilized the standard ResNet-18 model He et al. (2016) trained on the ImageNet dataset, which is available through the Torchvision library. The number of samples n ranged from 1 to 100, and the number of optimization epochs m ranged from 1 to 500. The results of our comparison are shown in Figure 7 for both pair-wise and layer-wise settings, illustrating the effect of hyperparameters on Mantel test statistic ρ using the shortest-path semantic baseline for comparison. From these results, we can observe that the number of epochs m has a greater influence on the correlation with ground truth than the number of samples n – the coherence with the ground truth increases as the number of optimization steps increases, and optimal performance is already achieved with only a few s-AMS samples per representation. Additionally, we observe that the layer-wise DORA method, which uses all neurons as descriptors for s-AMS, outperforms the pair-wise method. EA distances obtained from both layer-wise and pair-wise approaches are illustrated in Figure 8, together with semantic distance matrices, obtained from the respective taxonomy Miller (1995).

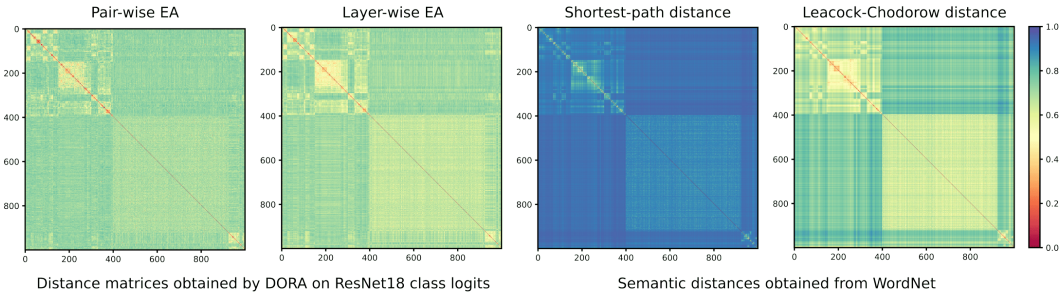

Figure 8: **Similarity between EA distances and taxonomy baseline distances.** The EA distances ($n = 50, m = 500$) between 1000 logit representations from the output layer of ImageNet-trained ResNet18, both in pair-wise and layer-wise fashion, are illustrated on the left, with semantic baseline distances between labels obtained from the WordNet taxonomy illustrated on the right. We can observe a connection between the baseline semantic distances and the distances obtained via Extreme-Activation metric in a data-agnostic manner by the visual similarity of distance matrices .

A.6 Evaluation with ground truth

In this experiment, we quantitatively assess the ability of our proposed method to conserve the semantic distance between representations across different datasets and architectures. To this end,

Table 1: **Correlation between EA distance matrices and semantic distance baselines.** Correlation coefficients ρ (higher is better) obtained from the Mantel test between pair-wise (p-w) and layer-wise (l-w) EA distances and shortest-path (d_{SP}) and Leacock-Chodorow (d_{LC}) semantic distances. All results show statistical significance with $p < 0.001$.

ImageNet	d_{SP}		d_{LC}	
	p-w	*l-w*	*p-w*	*l-w*
ResNet 18	0.44	**0.56**	0.40	**0.53**
AlexNet	0.48	**0.51**	0.46	**0.52**
ViT	0.50	**0.53**	0.49	**0.55**
BEiT	0.43	**0.50**	0.39	**0.48**
Inception V3	0.24	**0.27**	0.20	**0.23**
DenseNet161	0.37	**0.44**	0.32	**0.40**
MobileNet V2	0.47	**0.59**	0.43	**0.58**
ShuffleNet V2	**0.21**	0.14	**0.17**	0.10

CIFAR-100	d_{SP}		d_{LC}	
	p-w	*l-w*	*p-w*	*l-w*
ResNet 9	**0.26**	0.26	**0.27**	0.22
ShuffleNet V2	0.34	**0.35**	0.29	**0.30**
MobileNet V2	**0.30**	0.28	**0.25**	0.25
ResNet 18	0.25	**0.28**	0.21	**0.24**
ShuffleNet	0.34	**0.36**	0.30	**0.33**
VGG 11	0.21	**0.21**	0.19	**0.20**
NasNet	0.28	**0.29**	0.25	**0.26**
SqueezeNet	0.34	**0.34**	**0.28**	0.28

we used 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). The results of the evaluation can be found in Table 1, where the Mantel correlation test statistic ρ is stated for both the layer-wise (l-w) and pair-wise (p-w) versions between the computed EA distance and the two semantic distance baselines. From the results, we can observe that the distance obtained by the layer-wise EA metric is more favorable over the pair-wise one due to its stronger linear relationship with both baseline metrics. Additionally, we observe that correlations are higher for ImageNet than for CIFAR-100. This can be attributed to the complexity of the dataset, particularly to the size of the image domain (224×224 pixels in ImageNet versus 32×32 in CIFAR-100). As a sanity check, an additional experiment was carried out, where EA distances were computed over random noise signals for all networks and both datasets. The results revealed that the test statistics were approximately zero and statistical significance was not demonstrated at a significance level of $\alpha = 0.05$ for all the models.

A.7 Evaluating Anomaly-identification capabilities

The ability of EA distances to maintain semantic distances allows the DORA framework to not only detect and display clusters of semantically similar representations via representation atlases but also identify anomalous representations that semantically differ from the majority. While these representations may simply learn unique individual concepts, we demonstrate in further experiments that they might also have learned undesired concepts from spurious correlations in the training data that diverge from the typical (intended) decision-making strategy.

To assess the ability of DORA to detect anomalous representations, we conducted the experiment by inserting random representations in the network layer and evaluating the ability of different Outlier Detection (OD) methods to identify unnatural functions. We employed an ImageNet pre-trained ResNet18 and introduced a set of additional 25 random representations to the average pooling layer of the network, containing 512 learned representations, resulting in a total of 537 representations. These 25 representations were not learned but were constructed as linear combinations of existing representations from various layers within the network. The experiment was conducted in five different scenarios, in which inserted representations were constructed as linear combinations of representations from different layers of the network, namely the "maxpool" (layer preceding the "layer1"), "layer1", "layer2", "layer3", and "layer4". The weights for the linear combinations were randomly drawn from a standard normal distribution. For each scenario, we estimated the EA distance both pair-wise (p-w) and layer-wise (l-w) and used the resulting distances in five different Outlier Detection methods: the Angle-based Outlier Detector (ABOD) Kriegel et al. (2008), Feature

Table 2: **Performance of DORA in detecting random representations.** Each cell represents the average AUC ROC of the Outlier Detection methods, for detecting randomly generated representations, based on different layers of the ResNet18 network.

	"maxpool"		"layer1"		"layer2"		"layer3"		"layer4"	
	p-w	*l-w*	*p-w*	*l-w*	*p-w*	*l-w*	*p-w*	*l-w*	*p-w*	*l-w*
ABOD	0.79	**0.98**	0.74	0.93	0.77	0.80	0.81	0.62	1.00	0.58
FB	0.95	0.92	0.95	**0.96**	0.62	**0.89**	0.60	0.59	**1.00**	0.73
IF	0.76	0.83	0.70	0.63	0.58	0.59	0.63	0.57	**1.00**	0.73
LOF	0.66	0.72	0.68	0.69	0.84	0.63	**0.98**	0.78	**1.00**	0.63
OCSVM	0.62	0.84	0.65	0.79	0.78	0.77	0.88	0.78	0.61	0.70

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 existing (learned) representations in the layer, and randomly generated ones. To ensure stability, the OD results were repeated 10 times with different random states, and the linear combinations for the added representations were computed five times for each scenario. Classification performances for each OD method were averaged per scenario.

Table 2 presents the results of the described experiment, which demonstrate that all of the Outlier Detection methods performed well in terms of detecting the randomly generated representations across all scenarios. It is worth noting that the pair-wise EA distance (p-w) performed slightly better for the detection of random representations. Additionally, the results indicate that different layers of the network have an impact on the detection of random representations, as the Outlier Detection methods performed better when detecting representations generated from higher-level concepts, such as "layer2", "layer3", and "layer4", likely due to the fact that the analysis is performed on the average pooling layer, the last layer of the feature extractor of the network. The high detection rate of randomly initialized representations illustrates the ability of DORA to maintain semantic similarity between learned representations and find semantically anomalous representations.

A.8 Experiments

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.

A.8.1 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. While, in theory, character-detectors might align with specific classes (e.g. "keyboard"), our results suggest that the emergence of character-detector representations is due to a large number of watermarked images in the training dataset that are not equally distributed across ImageNet classes. As ImageNet does not have a 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 between watermarked and normal images, evaluating sensitivity using AUC ROC. 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 $n = 5$ s-AMS per each representation, with $m = 500$, based on our findings in the section A.5. 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)).

DORA identified five outlier representations, namely neurons 7, 99, 154, 160, 162, and 393. The outlier with the highest outlier score, 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 sensitivity (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 9, 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)).

Figure 9: **Cluster of Clever-Hans representations in the ResNet 18 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, which is the representation with the highest outlier score, 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.

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". 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

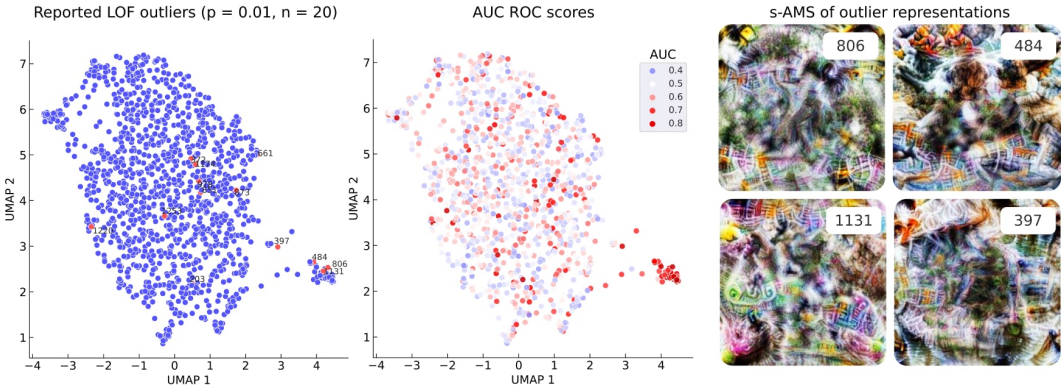

Figure 10: **Cluster of Clever-Hans representations in the MobileNet V2 feature extractor.** The left figure illustrates the outlier representations as identified by the LOF OD method, overlaid on the DORA representation atlas. The middle figure displays the sensitivity of the neural representations 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.

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 specific to Chinese character detectors in neurons 397, 484, 806, and 1131. Figure 10 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

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 11 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.

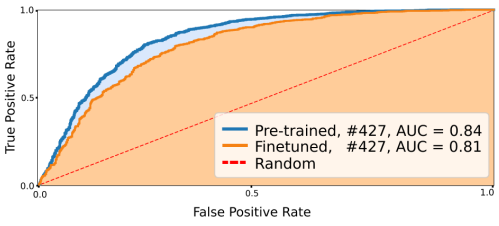

Figure 12: **Persistent Latin text detector**. Neuron 427 in the DenseNet121 network learns to detect Latin text during pre-training and does not unlearn this behavior after fine-tuning on the CheXpert dataset, as shown by the ROC detection curves.

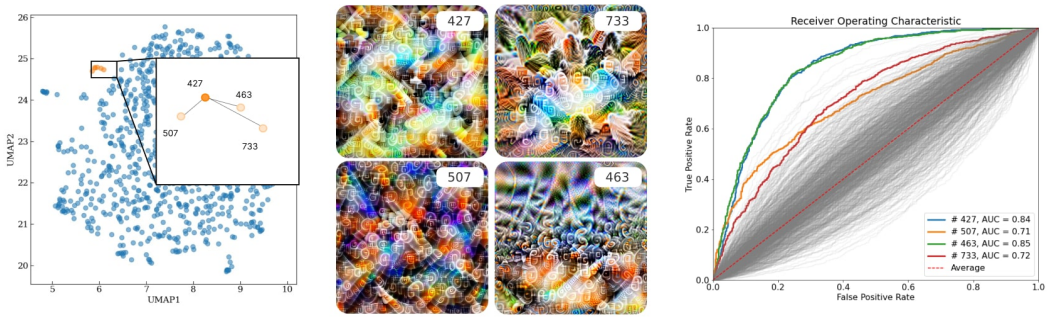

Figure 11: **DenseNet121 — Latin text detector.** Applying DORA to the last layer of the feature extractor of DenseNet121 yields, among others, Neuron 427 as an outlier, which corresponds to the upper left of the 4 feature visualizations. From neuron 427 as well as from its three closest neighbors (shown 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.

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 12. The results indicate that the Clever-Hans effect persisted after fine-tuning, possibly due to small Latin text characters in the dataset.

A.8.2 CLIP RESNET50

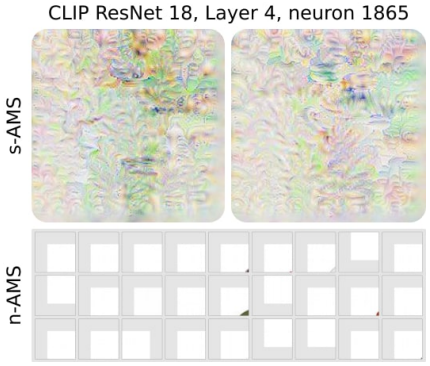

Figure 13: **AMS for reported outlier representation**. LOF identified neuron 1865 as the strongest outlier. Analysis of s-AMS and ImageNet n-AMS indicate that it primarily detects white images/backgrounds, which is atypical compared to other high-level representations in the same layer.

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)).

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 13. 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.

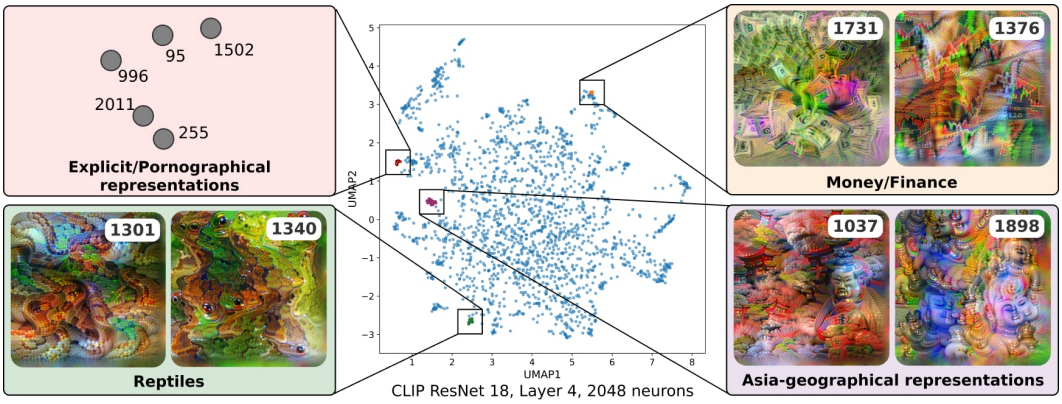

Figure 14: **Representation atlas of CLIP ResNet50 "layer 4".** Representation atlas for CLIP ResNet50 "layer 4", where several clusters of representations are highlighted. Activation-Maximisation signals associated with the Explicit/Pornographic representations were omitted due to the presence of explicit concepts in the signals.

After computing the representation atlas for "layer 4", we manually investigated several distinctive clusters. Figure 14 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.