PURE: Turning Polysemantic Neurons Into Pure Features by Identifying Relevant Circuits

Maximilian Dreyer[1], Erblina Purelku[1], Johanna Vielhaben[1],
Wojciech Samek[1,2,3,†], Sebastian Lapuschkin[1,†]
[1] Fraunhofer Heinrich Hertz Institute, [2] Technical University of Berlin,
[3] BIFOLD – Berlin Institute for the Foundations of Learning and Data
[†]corresponding authors: {wojciech.samek | sebastian.lapuschkin}@hhi.fraunhofer.de

Abstract

The field of mechanistic interpretability aims to study the role of individual neurons in Deep Neural Networks. Single neurons, however, have the capability to act polysemantically and encode for multiple (unrelated) features, which renders their interpretation difficult. We present a method for disentangling polysemanticity of any Deep Neural Network by decomposing a polysemantic neuron into multiple monosemantic "virtual" neurons. This is achieved by identifying the relevant sub-graph ("circuit") for each "pure" feature. We demonstrate how our approach allows us to find and disentangle various polysemantic units of ResNet models trained on ImageNet. While evaluating feature visualizations using CLIP, our method effectively disentangles representations, improving upon methods based on neuron activations. Our code is available at https://github.com/maxdreyer/PURE.

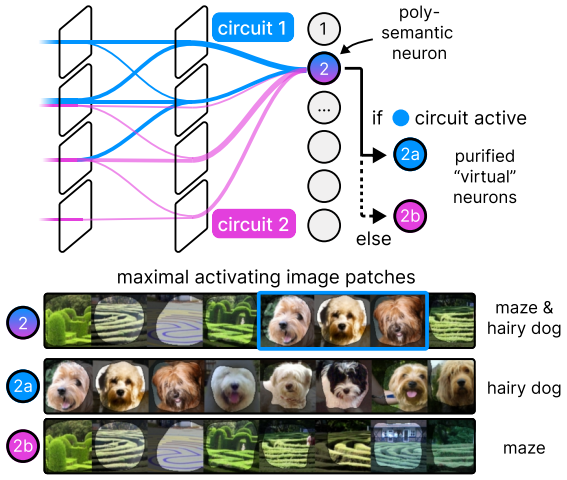

Figure 1. Distinct circuits exist for each feature of a polysemantic neuron. With PURE, we propose to split a polysemantic neuron into multiple pure "virtual" ones, one for each circuit. Here, we disentangle the maximally activating sample (patches) of neuron 2 into its two pure features: "hairy dog" (2a) and "maze" (2b).

1. Introduction

The field of eXplainable Artificial Intelligence (XAI) aims to increase the transparency of Deep Neural Networks (DNNs). Several XAI works study the role of a model's latent neurons and their interactions [18, 19], which recently developed into the sub-field of *mechanistic interpretability*. Neurons are commonly viewed as feature extractors corresponding to human-interpretable concepts [1, 3, 18]. However, neurons can be *polysemantic*, meaning that they extract multiple (unrelated) features, which adds ambiguity to their interpretability. Other XAI works study *circuits*, *i.e.*, distinct sub-graphs of a network performing specific subtasks, which recently became popular for Large Language Models (LLMs) [9, 11, 28]. Notably, the interpretability of a circuit depends on the interpretability of its units, which again can be polysemantic. In applications, such as knowledge discovery or validation of DNNs in safety-critical tasks, high latent interpretability is crucial for XAI usefulness [4, 8]. In this work, we build on the assumption that for each monosemantic ("pure") feature a unique subgraph exists. Identifying the active circuits then allows disentangling a polysemantic unit into multiple "virtual" pure units (with one circuit each), as shown in Fig. 1 where we disentangle a neuron encoding for dog and maze features.

To that end, we introduce Purifying Representations (PURE), a post-hoc approach for increasing interpretability of latent representations by disentangling polysemantic neurons into pure features. PURE is based on discovering the relevant (active) circuit of each semantics, which is identified via a partial backward pass. Through the means of foundational models, we measure a significant increase in interpretability of ResNet [12] models after applying PURE, also improving upon activation-based approaches.

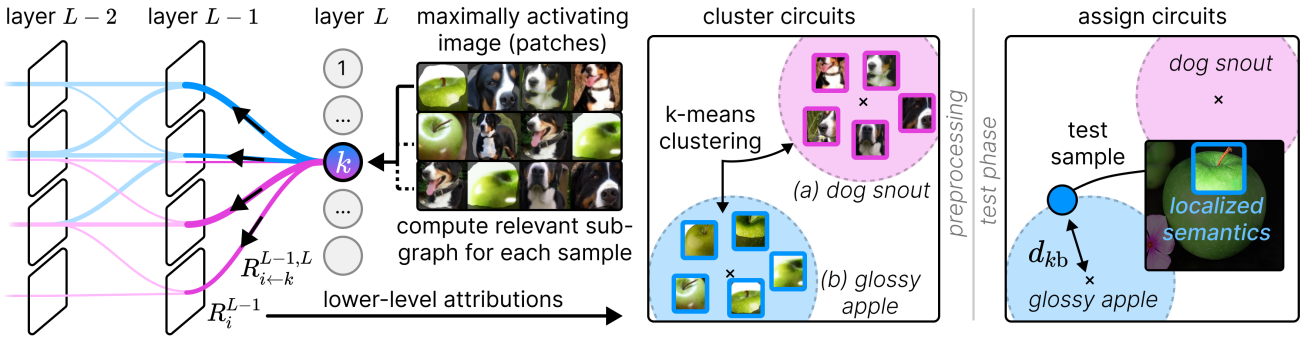

Figure 2. PURE detects circuits using lower-level neuron attributions for the n_{ref} most activating input samples in a preprocessing step. For polysemantic neurons, we assume distinct active circuits for each semantics, which are found through clustering attributions with k-means. During test phase, the active circuit can be assigned post-hoc for any new test sample by identifying the closest circuit.

2. Related Work

Various works have shown that neurons in DNNs encode for distinct features that can often be interpreted by humans [1, 3, 5]. However, besides redundancies in representations, an occurring problem is the polysemanticity of neurons. As such, interpretation of the latent space is difficult. For concept-based explanations, feature visualizations are confusing, or ambiguous as it might be unclear, which semantics are actually present. Further, semantics can be overlooked, *e.g.*, when one feature is more dominant than another [17]. As a way out, some works propose to find more meaningful directions [10, 14] or subspaces in latent space [27], but they often require pre-defined concepts or reconstruct the latent space only partially. To resolve poly-semanticity on a neuron level, O'Mahony *et al.* [20] propose to find directions (*i.e.*, a linear combination of neurons) based on latent activations. Instead of activations (that depend on *all* present input features), PURE is based on neuron-specific circuits which is more specific to the role of a neuron and leads to an improved disentanglement.

Circuits, in general, are viewed as the sub-graphs of a neural network architecture [28] that perform a specific task. They further consist of a set of linked features and the weights between them [19]. In recent work, circuit analysis [24] has been extended beyond convolution-based architectures [6] to, *e.g.*, transformer-based models [9, 11, 28]. The discovery of circuits has been partially automatized both for computer vision [23] and language models [7].

3. Method

For PURE, we view circuits as directed acyclic graphs that consist of nodes R^l_j for neuron j in layer l and edges $R^{l-1,l}_{i\leftarrow j}$ connecting nodes j and i of adjacent layers. We hypothesize that polysemanticity of a neuron arises because multiple circuits share one node, as illustrated in Fig. 2. PURE disentangles these circuits from a neuron perspective by clustering neuron k's functional connectivity with neurons of lower layers. This process involves two steps: (1) computing circuits, and (2) replacing a shared neuron with multiple virtual neurons for each circuit through clustering.

Step 1) Computing Circuits: To compute the edges of a circuit, we *explain* the activation of a neuron and attribute lower-level neurons. This naturally fits the idea of backpropagation-based feature attribution methods, specifically LRP [2]. LRP allows to efficiently backpropagate attribution scores through the network layer by layer, beginning at a latent neuron until the desired (input) layer is reached [2]. Concretely, the relevance R^l_j of an upper layer neuron j is generally distributed to lower-level neurons i as

$$R^{l-1,l}_{i\leftarrow j} = \frac{z^{l-1,l}_{i\to j}}{z^l_j} R^l_j \tag{1}$$

with $z^{l-1,l}_{i\to j}$ contributing to $z^l_j = \sum_i z^{l-1,l}_{i\to j}$ in the forward pass. Note that multiple refined ways to define $R^{l-1,l}_{i\leftarrow j}$ are proposed in literature for different layer types [16]. These relevance "messages" $R^{l-1,l}_{i\leftarrow j}$ refer to the *edges* of a circuit.

The circuit *nodes* are then characterized by node attributions computed via aggregation of the relevance messages:

$$R^{l-1}_i = \sum_j R^{l-1,l}_{i\leftarrow j}, \tag{2}$$

which reduces to $R^{L-1}_i = R^{L-1,L}_{i\leftarrow k}$ for the first backpropagation step when explaining neuron k in layer L as in Fig. 2.

For simplicity, we investigate in the following only nodes of the next lower-level layer $L-1$. It is to note, that other methods besides LRP can be used for attributions here [10]. We therefore default to Gradient×Activation as an efficient and universal attribution method implementing a simple LRP variant in ReLU-DNNs [26]. Thus, the circuit computation for neuron k in layer L simplifies for PURE to

$$R^{L-1}_i = A^{L-1}_i \frac{\partial A^L_k}{\partial A^{L-1}_i}, \tag{3}$$

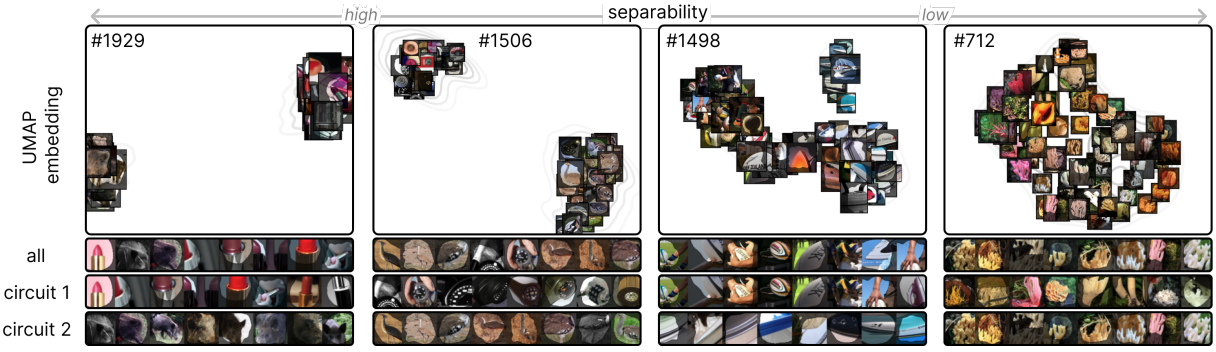

Figure 3. Applying PURE to neurons with varying degree of polysemanticity: We show UMAP embeddings with the maximally activating image patches, and the resulting reference sets before and after purification when identifying two circuits via k-means.

for circuit nodes i in lower-level layer $L-1$, with activation A_k^L of neuron k in layer L, and partial derivative $\partial/\partial A_i^{L-1}$ w.r.t. lower-level layer activations A_i^{L-1}.

Step 2) Assigning Circuits: We represent a neuron by its most activating input samples. Then, for each of the top-n_{ref} activating input samples of a polysemantic neuron k in layer L, we compute the lower-level attributions R_j^{L-1} for all n neurons of layer $L-1$ as given by Eq. (3) in a partial backward pass. If neuron j uses different circuits among the reference samples, we expect to see distinct clusters in the attribution vectors $\mathbf{R}^{L-1} \in \mathbb{R}^n$. To validate this assumption, we visualize a 2D UMAP [15] embedding of $\mathbf{R}^{L-1}$ in Fig. 2. To find the distinct clusters, we use k-means clustering which results in centroids representing new virtual neurons for each circuit. For a new test sample, we can then identify the active circuit by computing $\mathbf{R}^{L-1}$ and assigning it to the closest cluster centroid, *i.e.*, virtual neuron.

4. Experiments

We address the following research questions:

1. **(Q1)** Can we find and purify polysemantic neurons?
2. **(Q2)** How effective is PURE in disentangling representations compared to other approaches?

Experimental Setting We investigate the neurons in the penultimate layer of ResNet-34/50/101 models [12] pretrained [29] on the ImageNet [25] dataset, and evaluate interpretability using the foundational models of CLIP [22] and DINOv2 [21]. We perform the analysis on the $n_{\text{ref}} = 100$ maximally activating input samples (based on max-pooling) for each neuron on the ImageNet test set. To generate feature visualizations, we crop samples such that only the important part of a neuron's semantics remains [1], as illustrated in Fig. 2 (*right*) and detailed in Appendix A.1.

4.1. From Polysemanticity to Pure Features (Q1)

We begin with the quest to find polysemantic units by studying their feature visualizations, *i.e.*, their most activating image patches. As a quantitative and objective measure for monosemanticity, we evaluate CLIP embeddings, where visually similar feature visualizations presumably result in small embedding distances [13, 30]. Concretely, for each neuron k we compute the distance matrix

$$\mathbf{D}_{ij}^k = \sqrt{\left(\mathbf{e}_i^{\text{CLIP}} - \mathbf{e}_j^{\text{CLIP}}\right)^2} \tag{4}$$

between the CLIP embeddings $\mathbf{e}_i^{\text{CLIP}}$ of all pairs of feature visualizations (cropped reference samples) i and j.

To optimize the process of finding polysemantic units, we perform k-means clustering on the CLIP embeddings with a fixed number of two clusters. Then, inter- and intra-cluster distances ρ_k for neuron k are computed as

$$\rho_k^{\text{intra}} = \frac{\sum_{i,j\neq i}^{n_{\text{ref}}} \mathbf{D}_{ij}^k \mathbf{1}_{c_i=c_j}}{\sum_{i,j\neq i}^{n_{\text{ref}}} \mathbf{1}_{c_i=c_j}}, \quad \rho_k^{\text{inter}} = \frac{\sum_{i,j}^{n_{\text{ref}}} \mathbf{D}_{ij}^k \mathbf{1}_{c_i\neq c_j}}{\sum_{i,j}^{n_{\text{ref}}} \mathbf{1}_{c_i\neq c_j}} \tag{5}$$

with indicator function $\mathbf{1}_{c_i=c_j}$ equaling one if feature visualizations i and j have the same cluster index c, and zero else. A large difference $\rho_k^{\text{inter}} - \rho_k^{\text{intra}}$ indicates clearly separated clusters with different semantics.

In Fig. 3, neurons of varying degree of polysemanticity, as given by $\rho_k^{\text{inter}} - \rho_k^{\text{intra}}$, are shown for ResNet-50. Here, we also depict UMAP embeddings based on PURE attributions given by Eq. (3), and the feature visualizations before and after applying PURE. Note, that for PURE, we here disentangle one neuron into two virtual ones by clustering the lower-level attributions. As visible, polysemantic neurons such as `#1929` and `#1506` exist and can be effectively disentangled. The disentangled semantics can be visually different ("lipstick" and "boar") or more related ("white spots" on dark circular objects or bird wings). On the other hand, we can also find rather monosemantic neurons, *e.g.*, `#1498` and `#712`, that encode for lines and coral structure, respectively. More examples and more detailed results w.r.t. the distribution of polysemanticity are given in Appendix A.2.

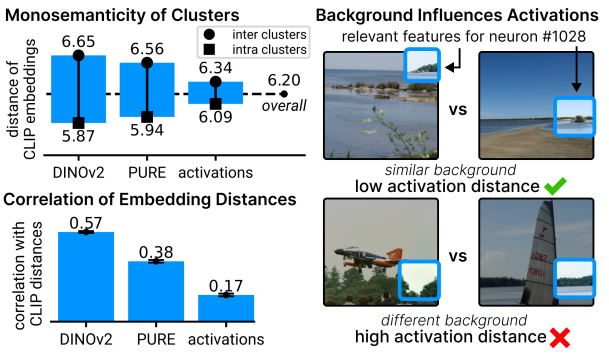

Figure 4. PURE leads to more interpretable representations as measured via CLIP embedding distances on feature visualizations (*top left*), thereby improving upon activation-based clustering and reaching almost DINOv2 scores. (*Bottom left*): Distances of CLIP embeddings between feature visualizations correlate with PURE embeddings significantly more than activations. (*Right*): Activations tend to overestimate distances when unrelated features vary.

4.2. Evaluating Feature Purification (Q2)

PURE aims to turn polysemantic neurons into purer virtual neurons that are easier to interpret, and works by computing and identifying the relevant circuits based on lower-level attributions R_i^{L-1} as by Eq. (3). Alternatively, O'Mahony *et al*. [20] propose activation-based disentanglement, where the activations A_i^L resulting from reference samples are clustered. When applied to a (polysemantic) neuron, both methods should result in more meaningful reference sample subsets for each disentangled feature, as, *e.g.*, in Fig. 3.

To systematically evaluate the interpretability of newly disentangled representations, we compute inter- and intra-cluster distances as defined in Eq. (5) using CLIP for the resulting sets of feature visualizations (cropped reference samples). Ideally, intra-cluster distances are low and inter-cluster distances are high, indicating well separated feature visualizations and more meaningful representations. The resulting CLIP embedding distances are reported in Fig. 4 (*top left*) for PURE and activation-based disentanglement for ResNet-101. As another baseline, we perform clustering on DINOv2 embeddings for the cropped reference samples. Here, DINOv2 represents an ideal visual separation, which is, however, computationally expensive as it requires both the computation of the cropped reference samples and a DINOv2 forward pass. The results show, that PURE leads to more disentangled representations than activation-based clustering, and is performance-wise close to DINOv2. Note that so far, clusters are computed using k-means with $k = 2$, but the same trends hold for different $k \in \{3, 4, 5\}$ and other ResNet architectures, as discussed in Appendix A.4.

In a second experiment, we dive deeper into why PURE attributions are more meaningful than latent activations. We thus investigate whether when two feature visualizations are similar according to CLIP, they are also similar according to PURE attributions or activations. Concretely, for feature visualization pairs, we compute CLIP embedding distances via Eq. (4) and distances between PURE attributions as well as activations, and finally measure the correlation between the resulting distances of different methods. Please note, again, CLIP (and DINOv2) embeddings refer to the *cropped* reference samples, whereas PURE and activations are computed on the *full* reference samples. As shown in Fig. 4 (*bottom left*), PURE has higher alignment to CLIP compared to activations. We observe that activations lead to deviating distance scores in some cases, especially when the relevant semantics are very localized in reference samples, as shown in Fig. 4 (*right*) for neuron `#1028` encoding for "vegetation on horizon". Notably, activations take into account *all* present features in the *full* reference sample, which influences distances when unrelated features (*e.g.*, airplanes or boats) vary between samples. Whereas, conditional attributions as used by PURE are more specific to the actual task of a neuron. Correlation results for the other ResNet architectures and more examples when PURE or activation-based clustering diverges from CLIP are given in Appendix A.4.

5. Limitations and Future Work

So far, we assumed that embeddings of foundational models can be seen as ideal indicators for human interpretability and disentanglement. In future work, evaluation in controlled settings or using human feedback will be valuable.

Regarding PURE, a large ablation study will be interesting, *e.g.*, for clustering on full circuits (instead of only lower-level layer attributions), and using different feature attribution methods or other clustering approaches than k-means. Notably, PURE requires a partial backward pass to disentangle a neuron, which is computationally slightly more demanding than activation-based disentangling.

It will be interesting to further study the advantages of purified units for XAI tools such as concept-based explanations, concept discovery and probing, or model correction.

6. Conclusion

We introduce PURE, a novel method for turning polysemantic neurons into multiple purer "virtual" neurons by identifying the active characteristic circuit of each pure feature. The purification of latent units allows to better understand latent representations, which is especially interesting for the growing and promising field of concept-based XAI. Using foundational models for evaluation, PURE results in significantly more purified features than activation-based approaches, which are less neuron-specific. We believe that our work will raise interest in investigating the benefits of cleaner representations for, *e.g.*, concept discovery, concept probing or model correction.

Acknowledgements

This work was supported by the Federal Ministry of Education and Research (BMBF) as grant BIFOLD (01IS18025A, 01IS180371I); the German Research Foundation (DFG) as research unit DeSBi (KI-FOR 5363); the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant TEMA (101093003); the European Union's Horizon 2020 research and innovation programme (EU Horizon 2020) as grant iToBoS (965221); and the state of Berlin within the innovation support programme ProFIT (IBB) as grant BerDiBa (10174498).

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 1, 2, 3

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10 (7):e0130140, 2015. 2

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 1, 2

[4] Kirill Bykov, Mayukh Deb, Dennis Grinwald, Klaus Robert Muller, and Marina MC Höhne. DORA: Exploring outlier representations in deep neural networks. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 1

[5] Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina Höhne. Labeling neural representations with inverse recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[6] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. https://distill.pub/2020/circuits. 2

[7] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[8] Maximilian Dreyer, Reduan Achtibat, Wojciech Samek, and Sebastian Lapuschkin. Understanding the (extra-)ordinary: Validating deep model decisions with prototypical concept-based explanations. *arXiv preprint arXiv:2311.16681*, 2023. 1

[9] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html. 1, 2

[10] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3

[13] Neha Kalibhat, Shweta Bhardwaj, Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3

[14] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 2

[15] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018. 3

[16] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019. 2

[17] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016. 2

[18] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 1

[19] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020. 1, 2

[20] Laura O'Mahony, Vincent Andrearczyk, Henning Müller, and Mara Graziani. Disentangling neuron representations with concept vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3770–3775, 2023. 2, 4

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[23] Achyuta Rajaram, Neil Chowdhury, Antonio Torralba, Jacob Andreas, and Sarah Schwettmann. Automatic discovery of visual circuits. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. 2

[24] Tilman Raukur, An Chang Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483, 2022. 2

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[26] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017. 2

[27] Johanna Vielhaben, Stefan Bluecher, and Nils Strodthoff. Multi-dimensional concept discovery (mcd): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research*, 2023. 2

[28] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[29] Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. 3

[30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 3