Disentangling Neuron Representations with Concept Vectors

Laura O'Mahony*[1,2]
lauraa.omahony@ul.ie

Vincent Andrearczyk[1]
vincent.andrearczyk@hevs.ch

Henning Müller[1,3]
henning.muller@hevs.ch

Mara Graziani[1]
mara.graziani@hevs.ch

[1] Haute école spécialisée de Suisse occidentale, Hes-so Valais, Sierre, Switzerland
[2] University of Limerick, Limerick, Ireland
[3] The Sense Research and Innovation Center, Sion, Lausanne, Switzerland

Abstract

Mechanistic interpretability aims to understand how models store representations by breaking down neural networks into interpretable units. However, the occurrence of polysemantic neurons, or neurons that respond to multiple unrelated features, makes interpreting individual neurons challenging. This has led to the search for meaningful vectors, known as concept vectors, in activation space instead of individual neurons. The main contribution of this paper is a method to disentangle polysemantic neurons into concept vectors encapsulating distinct features. Our method can search for fine-grained concepts according to the user's desired level of concept separation. The analysis shows that polysemantic neurons can be disentangled into directions consisting of linear combinations of neurons. Our evaluations show that the concept vectors found encode coherent, human-understandable features.

1. Introduction

Mechanistic interpretability is a fast-emerging research topic that aims at deciphering the internal representations held by a model by reverse engineering into understandable computer programs [1, 20, 21]. Previous work in this field breaks down convolutional neural networks (CNNs) into the features learned by the fundamental units of a layer, which are considered as directions of a geometric space. Many previous works consider neurons as these units [20, 21]. Breaking down the model into such interpretable units allows us to better understand how models store representations in vision tasks [4, 5, 20, 21] and language models [8]. This could even allow us to predict and edit model behaviour such as work by Bau *et al*. that removes units that are important to a class [5] and has also been studied for other architectures such as GANs and GPT models [3, 16, 17].

A frequent issue is the occurrence of *polysemantic neurons*, namely neurons that respond to several unrelated features, or concepts [19–21]. They can be found by looking at the maximally activating dataset examples and finding they consist of multiple groups that are conceptually very different [6, 20]. This makes the interpretation of individual neurons challenging since they cannot be mapped to unique features. This is exemplified in Olah *et al*. [20, 21] by a neuron equally likely to respond to car shields and cat paws at the same time, and with the same intensity. There is evidence that the training of models pushes networks to represent many features within individual neurons [7, 24]. Models have a limited number of neurons meaning a discrete neuron is often not possible for all features. This is related to the idea of *superposition*, which refers to when neural networks represent more features than they have neurons [7]. These empirical observations in existing research indicate that neurons are not always the right fundamental unit encapsulating an individual feature represented by a model. If we define activation space as all possible combinations of neuron activations, we can widen our lens to look for meaningful vectors in activation space instead of single neuron basis vectors.

There is evidence that suggests monosemantic regions in activation space exist [4, 6, 7, 11, 21], but they are not always made obvious by studying individual neurons [6, 26]. A key issue resulting from this observation is the question of how directions in activation space representing distinct features can be found [21].

Previous work has shown the existence of high-level human interpretable concepts such as textures, shapes and parts of objects present as directions in activation space.

*corresponding author lauraa.omahony@ul.ie
Source code available at https://github.com/lomahony/sw-interpretability

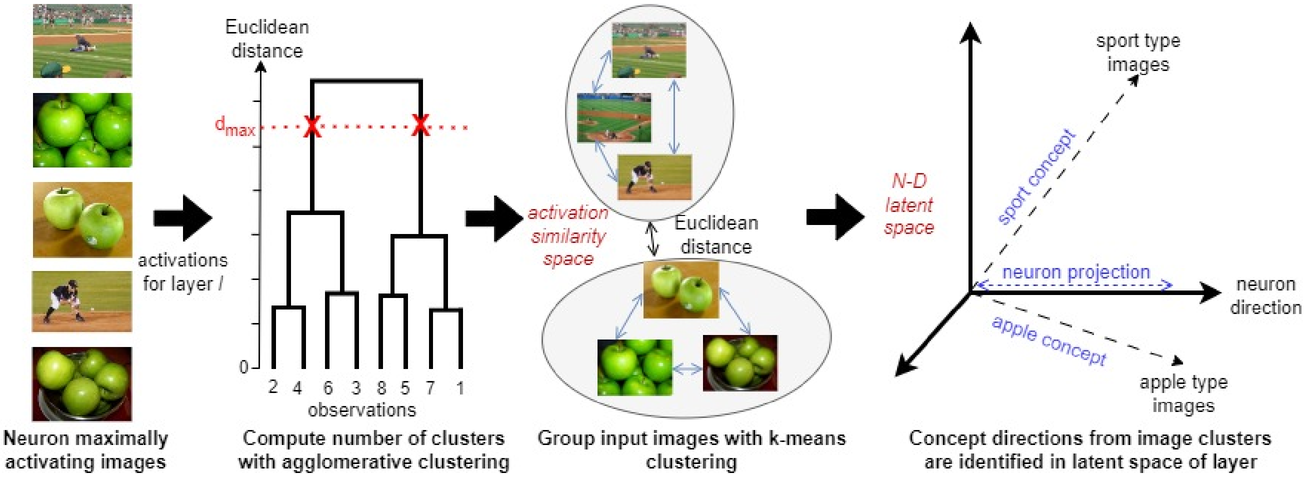
Euclidean distance
d_{max}
0
2 4 6 3 8 5 7 1
observations
activations for layer l
activation similarity space
Euclidean distance
N-D latent space
sport type images
sport concept
neuron projection
neuron direction
apple concept
apple type images
Neuron maximally activating images
Compute number of clusters with agglomerative clustering
Group input images with k-means clustering
Concept directions from image clusters are identified in latent space of layer

Figure 1. Step 1. A set of images that maximally activate a neuron in a model layer is taken. Step 2. The Euclidean distance between images in activation space is used as the similarity space on which the clustering is performed. This returns the appropriate number of clusters for a given distance threshold. Step 3. K-means clustering computes the cluster membership. Step 4. From the images in each cluster, a concept vector is calculated, which points toward the non-neuron aligned direction in activation space.

Some early work by Alain *et al*. [2] took the features of the layers in a model and fit a linear classifier to each layer to predict the class labels. The work by Kim *et al*. [13] on *Concept Activation Vectors (CAVs)* defines a concept as a vector in the direction of the activation values of a set of examples of that concept. The authors find a concept by training a linear classifier to distinguish between examples of that concept and random counterexamples. The concept vector is then taken as the vector orthogonal to the boundary. A limitation of this method is that it requires a handcrafted set of examples of a concept to find the concept direction in latent space. A small number of unsupervised methods used to find concepts have been developed [10, 14, 22], this research direction is known as *concept discovery*.

Concept discovery involves the search for unit vectors in the latent space of a model that encode learned representations of high-level concepts. However, none of the existing methods seek to disentangle polysemantic representations. The concept vectors are linear combinations of units, and as such, they are likely to inherit polysemanticity from polysemantic neurons [10]. Furthermore, none of these methods incorporate the notion of a privileged basis proposed by Elhage *et al*. [7]. A *privileged basis* is where some representations are encouraged to align with basis directions, meaning directions in space corresponding to individual neurons. Even though neuron directions are usually meaningful candidates for representing a feature [4, 7, 20, 21], they likely do not show the whole story due to the countervailing force of superposition [7]. The main contribution of this paper is a method to find and disentangle monosemantic directions starting from polysemantic neurons. Moreover, our method can search for concepts that are fine-grained according to the user's desired level of concept separation. Our analysis shows that polysemantic neurons can be disentangled into directions consisting of linear combinations of neurons.

2. Methods

We consider a CNN predicting a classification output (p-dimensional output vector) from an input image. We note that the method can be generalised to other models, but use a CNN for our analysis. We assume the model was already trained, and that we have access to the intermediate representations of an arbitrary layer inside the model. Fig. 1 summarises the steps discussed in more detail below.

We take a given intermediate layer l. We calculate the embeddings for the entire dataset and apply global average pooling to aggregate the spatial information of the convolutional feature maps. We select a neuron n, and apply the following steps iteratively. In step 1, we take these activations $\{\phi^l(x_i)\}_{i=1}^N$ where $\phi^l(x_i) \in \mathbb{R}^d$, and for the neuron n, we take the top N activating images $\{x_i\}_{i=1}^N$.

The second step involves measuring the similarity of the pooled activations (of the top activating images) at the intermediate layer l. We use the Euclidean distance as a distance metric which has been shown by previous work to be highly predictive of perceptual similarity [28]. We then apply a clustering technique to group these measurements of similarity into sets of close examples. For this, we use agglomerative clustering, a bottom-up type of hierarchical clustering [18], since it does not require us to pre-specify the number of clusters to be generated, as is required by the k-means approach. With clustering settings described in Appendix A.1, we apply agglomerative clustering with a distance threshold d_{max} that specifies the maximum linkage threshold at which clusters will be merged. Its result can be visualised in a dendrogram, or tree-based representation of elements, as is depicted in Fig. 1, step 2. The distance threshold is a hyperparameter that is tuned to an appropriate range for the model layer. It can be tweaked to fine-grain concepts into big or small buckets. Please refer to Appendix A.1 for a demonstration of this.

The third step takes the resulting number of clusters C obtained from step 2 and performs k-means clustering on the same measurements of similarity. The benefit of using k-means clustering over agglomerative clustering alone is that the k-means centroids allow us to easily remove outliers from each cluster that have low similarity to the rest of the cluster as employed in [9].

The final step finds the directions of the $\hat{C}$ [1] concept vectors corresponding to the $\hat{C}$ clusters, $\{\hat{c_j}\}_{j=1}^{\hat{C}}$, by taking the mean of the remaining embeddings for each cluster, giving us a set of vectors $\{avg(\{\phi^l(x_i)\}_{i\in\hat{c_j}})\}_{j=1}^{\hat{C}}$ which are then normalised to give us disentangled concept vectors $\{v_j\}_{j=1}^{\hat{C}}$, where $v_j \in \mathbb{R}^d$.

3. Results

We consider Inception V3 (IV3) [25] for our experiments since it is a de-facto standard convolutional neural network. Moreover, interpretability research has already given multiple insights for this model [9, 10, 13], and pre-trained weights on the ImageNet ILSVRC2012 [23] dataset are available online. As this exploratory study only aims at a proof of concept, we focused on an undersampled version of ImageNet, retaining 130 random images for each class. This kept computation accessible to our infrastructure, feasible and light. Our results can easily be scaled to the entire dataset and larger dataset sizes. Where not stated, we consider the concatenation layer *Mixed 7b*, a convolutional layer with 2048 feature maps ($d = 2048$) near the end of the IV3 model. We pick this layer as we expect it to encode complex concepts [21, 27]. A similar analysis can be done on other layers and architectures.

We demonstrate the results of the method described in Sec. 2 on a number of both polysemantic and monosemantic neurons. We took $N = 100$ top activating dataset examples and set the distance threshold parameter $d_{max} = 15$. We select neuron 35 as an example of a polysemantic neuron, as it activates highly for images of apples, sports, and also three images are dominated by a net-like pattern. This results in 3 clusters. When the k-means clustering step was applied (i.e. Step 3 in Figure 1) and outliers were removed, the cluster containing the net-like images was removed as there were < 5 images in this cluster. Fig. 2a shows the embeddings of the remaining images plotted using UMAP [15] dimensionality reduction. The plot shows how UMAP separates embeddings for images of apples and sports. Note that UMAP is used only for demonstration of our results to depict the clusters found using k-means since we found it accurately reflects the cluster membership result. We select neuron 16 as an example of a monosemantic neuron that activates for many categories of elliptical shapes as depicted in Fig. 2b. The same procedure applied to this neuron yields only one cluster, yielding one monosemantic concept vector which has a much higher similarity than the neuron direction to the original images. The case of neuron 1 is interesting and demonstrates the application of our method to fine-grain concepts. This neuron activates highly for underwater images as shown in Fig. 2c. Further inspection shows how it activates highly for both general underwater images and, images of scuba divers. Applying our method yields two clusters, one for general underwater scenes such as coral, and another for scuba divers, meaning two concept vectors can be found for this neuron. However, at a higher distance threshold, the clusters of images are merged by the algorithm, and one concept is found. Here the features are not entirely conceptually different. However, these two related directions can be differentiated, and concept discovery can separate them. We believe this hints at how it may be helpful to view concepts as varying continuously in the latent space instead of being encoded discretely by neurons. We suspect this phenomenon is related to the notion of 'feature facets' [19]. The same analysis was scaled to multiple neurons in the same layer. We provide additional examples and a depiction of the dendrogram for neuron 1 to see the result of altering d_{max} in the Appendix A.1, Figs. A.7 and A.8. We note that we found that the majority of the neurons we analysed in layer *Mixed 7b* were found to show some amount of polysemanticity. A possible explanation for this is that the number of features may be very high for a later layer in the model as it encodes complex concepts.

We performed a qualitative and quantitative assessment of the identified concepts. The semanticity of concept vectors was evaluated, first by finding the dataset examples with the largest projections along the vectors, analogous to viewing the maximally activating dataset examples for individual neurons. Fig. 3 demonstrates how the two concept vectors found for polysemantic neuron 35 were confirmed to be monosemantic regions in the latent space, cleanly activated by the originally entangled concepts. A further qualitative assessment involved applying the technique of feature visualisation [21] using the Lucent [12] library. Fig. 3 also shows the result of applying this technique on polysemantic neuron 35, and on the concept vectors found with our method. The polysemantic neuron fails to give a human interpretable representation, whereas the disentangled directions closer resemble the distinct categories of images which excite this neuron. For instance, the concept point-

[1] $\hat{C}$ may be different from C since clusters with less than 5 samples are removed.

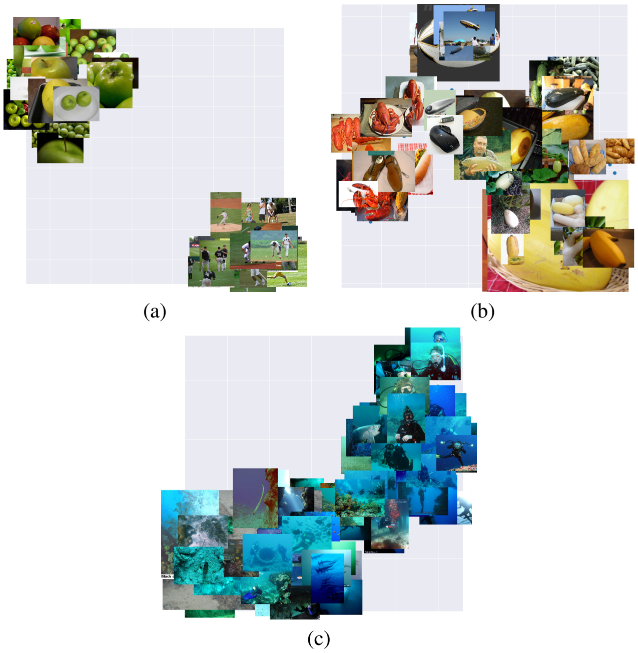

Figure 2. UMAP of the maximally activating images kept after k-means clusters and outlier removal in latent space: (a) separate clusters for polysemantic neuron 35 (b) a single cluster for monosemantic neuron 16.

ing towards representations of apples is maximally activated by round shapes with a stem cavity that is typical of apples, whereas the second concept seems to be maximally activated by large green squares such as football, soccer or baseball pitches.

As has been done in concept discovery works [9, 10], we evaluated our results with human experiments to evaluate the coherency and understandability of concepts. To avoid any cherry-picking of results, concepts from the first 8 neurons were used for all questions on the form, and a random number generator was used to select images from each concept's maximally projecting images[2]. The first four questions evaluated the concepts' coherency by asking participants to identify an intruder out of four other images that maximally activate another concept vector found from the same neuron or concept discovery starting point. The (n = 8, including 3 domain experts) participants selected the intruder image with an overall 100% accuracy showing the images have a coherent theme. The other six questions were designed to evaluate the understandability of concepts. Participants were asked to label two concepts and assess whether they agree with a given label for four sets of images. Agreement with the given labels was observed 97% of the time. Further details are provided in Appendix A.3

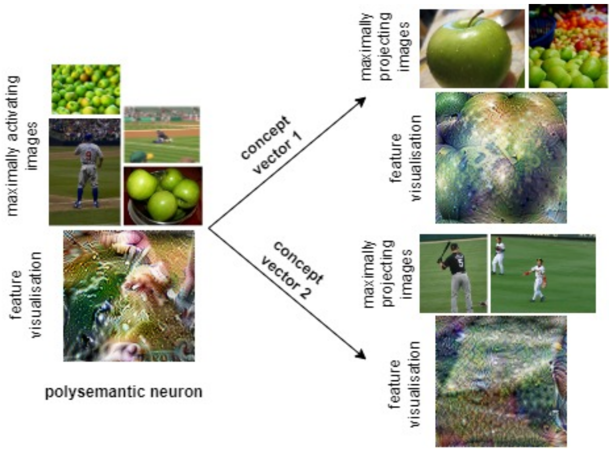

Figure 3. Disentanglement of the representations for neuron 35. The maximally activating inputs and feature visualisation are shown for the polysemantic neuron (left) and the disentangled concept directions (right).

for the only label change suggestion (from a domain expert) that occurred. This confirms that the images have a consistent semantic meaning across multiple individuals.

A number of quantitative metrics were used to analyse the images making up the clusters for computing the concept vectors and also the maximally projecting images along concepts. A natural starting point was to check the distribution of Euclidean distances between maximally activating images as is shown in Fig. 4 (a). Fig. 4 (b) shows how the inter-cluster distance (distance between images of apples and sport-type images) is considerably higher than the intra-cluster distance. Appendix A.2 and particularly Fig. A.9 illustrate the components of the 2048 concept vectors for these two concepts, which differ considerably across other dimensions. When the maximally projecting images along the concept vectors were calculated, our analysis confirmed that the projections and cosine similarities of their corresponding activations with the concept vector are remarkably higher than that with the neuron direction as exhibited in Fig. 5. As shown in additional examples in Appendix A.1, our results are consistent for other monosemantic and polysemantic neurons.

4. Conclusions and Future Work

Finding meaningful directions in activation space that are pointing to unique patterns, or concepts is a non-trivial problem encountered in our journey of understanding neural networks. Our results suggest that exploring directions, instead of neurons may lead us toward finding coherent fundamental units. We believe this work helps move toward bridging the gap between understanding the fundamental units of models as is an important goal of mechanistic interpretability, and concept discovery. We evaluated the co-

[2] The evaluation form can be accessed at https://forms.gle/62H6iUiXHdLYs9nr9

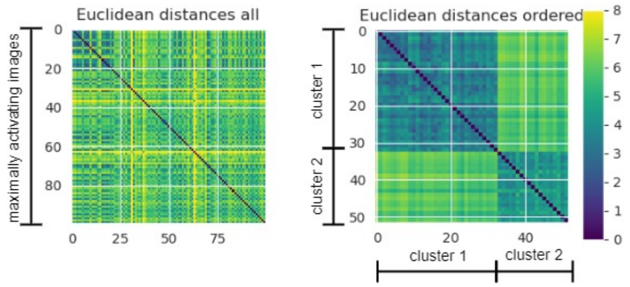

Figure 4. Euclidean distance between activation pairs, for the top 100 maximally activating images as in step 1 (on the left) and for the 52 remaining images after step 3, ordered by cluster membership (on the right).

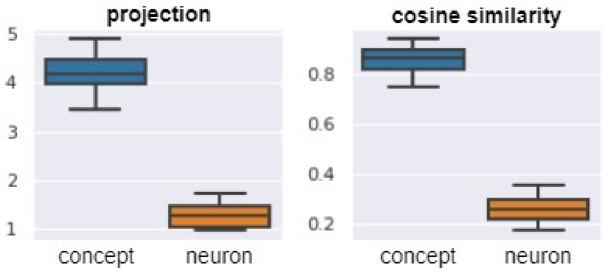

Figure 5. The left plot shows the projection of the elements of a cluster along the corresponding concept vector and the projection along the neuron direction for neuron 35. The right plot shows the cosine similarities between elements of a cluster with the concept vector and the neuron direction.

herency and understandability of the raw images whose embeddings have the maximum projections along a concept vector. We found that the latent space representations have much higher similarities with the concept vectors discovered than with the neuron directions. This work goes in the direction of building interpretability in a human-controlled way, as is important for the field of AI safety, and for applications of image models such as medical lesion analysis. We note a limitation of this work is its reliance on the data used to generate clusters. Furthermore, all experiments were performed on image data as image data is easier to visualise than other data forms. Generalising the method to other data types such as language and tabular data is a direction we wish to pursue in future work, as is looking at other starting candidates for concepts besides neurons.

4.0.1 Acknowledgements

This work was supported by AI4Media of the European Union's Horizon 2020 research and innovation program under grant agreement No 951911. LOM performed this work during an internship facilitated by the AI4Media Junior Fellows Exchange Program. LOM also acknowledges the support of Science Foundation Ireland under Grant Number 18/CRT/6049. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 1

[2] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 2

[3] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 351–369. Springer, 2020. 1

[4] 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

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020. 1

[6] Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, et al. Interpreting neural networks through the polytope lens. *arXiv preprint arXiv:2211.12312*, 2022. 1

[7] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. 1, 2

[8] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. 1

[9] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 4

[10] Mara Graziani, An-phi Nguyen, Laura O'Mahony, Henning Müller, and Vincent Andrearczyk. Concept discovery and dataset exploration with singular value decomposition. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 2, 3, 4

[11] Adam S Jermyn, Nicholas Schiefer, and Evan Hubinger. Engineering monosemanticity in toy models. *arXiv preprint arXiv:2211.09169*, 2022. 1

[12] Lim Swee Kiat. Lucent, lucid library adapted for pytorch, 2021. 3

[13] 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, 3

[14] Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022. 2

[15] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3

[16] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual knowledge in gpt. *arXiv preprint arXiv:2202.05262*, 2022. 1

[17] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022. 1

[18] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012. 2

[19] 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. 1, 3

[20] 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

[21] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 1, 2, 3

[22] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017. 2

[23] 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(3):211–252, 2015. 3

[24] Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022. 1

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[27] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 3

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2