# Much Easier Said Than Done: Falsifying the Causal Relevance of Linear Decoding Methods

**Lucas Hayne**[1]* **Abhijit Suresh**[1] **Hunar Jain**[1] **Rahul Kumar**[1] **R. McKell Carter**[2]

[1]Department of Computer Science
[2]Department of Psychology and Neuroscience
University of Colorado Boulder

## Abstract

Linear classifier probes are frequently utilized to better understand how neural networks function. Researchers have approached the problem of determining unit importance in neural networks by probing their learned, internal representations. Linear classifier probes identify highly selective units as the most important for network function. Whether or not a network actually relies on high selectivity units can be tested by removing them from the network using ablation. Surprisingly, when highly selective units are ablated they only produce small performance deficits, and even then only in some cases. In spite of the absence of ablation effects for selective neurons, linear decoding methods can be effectively used to interpret network function, leaving their effectiveness a mystery. To falsify the exclusive role of selectivity in network function and resolve this contradiction, we systematically ablate groups of units in subregions of activation space. Here, we find a weak relationship between neurons identified by probes and those identified by ablation. More specifically, we find that an interaction between selectivity and the average activity of the unit better predicts ablation performance deficits for groups of units in AlexNet, VGG16, MobileNetV2, and ResNet101. Linear decoders are likely somewhat effective because they overlap with those units that are causally important for network function. Interpretability methods could be improved by focusing on causally important units.

## 1 Introduction

Neural networks play critical roles in everything from driving to systems of justice, but trusting neural networks in such roles requires understanding how they function. Widespread adoption of neural networks has raised concerns about trust and bias [1]. For example in medicine, the incorporation of neural networks requires considering how uncertainty in how a decision was made could be argued to undermine medical authority [2]. The requirement for transparency is so widely recognized it has begun to be incorporated in regulation [3]. Improving our understanding of neural network function is in both the interest of the public and AI development communities. Unfortunately, most methods used for interpretation have not been strongly linked to neural network function.

One approach to understanding neural network function is to examine how they represent information internally [4–8]. Internal representations have been shown to be remarkably useful in interpreting neural networks [9–11]. Linear classifier probes [12] are commonly used as a tool for interpreting [5, 13–15] and testing internal representations [16–19]. Although linear probes have utility for understanding network representations, studies removing (ablating) groups of neurons identified by

*lucas.hayne@colorado.edu

I Can't Believe It's Not Better Workshop at NeurIPS 2022.

linear probes have inconsistent outcomes. The best way to confirm the causal importance of groups of neurons is to remove them and see if the network stops performing its function [20], hereafter called ablation. Performance is not consistently degraded when ablating the units most selective for the target class (i.e. units that activate more for the target class than other images) [21, 6, 22]. As an example, Donnelly and Roegiest [23] ablate a neuron highly selective for sentiment and observe negligible deficits in performance. This argument can be extended to both training against class selectivity [24] and for the effect of class selectivity on overall performance [25, 22, 6]. However, there are examples in which researchers observed a weak correlation between class selectivity and ablation deficits [22, 26]. The significant issue remains that it is difficult to predict the effects caused by ablating the selective units identified by linear probes.

Most ablation studies of performance remove a single unit at a time or are based on single measures of importance (activation strength, selectivity, decoding weight, information-theoretic measures, etc.) [6, 22, 27, 21, 28–30], preventing study of the aggregate effects of groups of units. Single unit ablation avoids the combinatorially difficult problem of choosing which units to group, but results in small or no changes in performance. Larger performance deficits can be achieved by targeting aggregate groups of units based on a single metric (e.g., selectivity). However, utilizing single units or single dimensions precludes a systematic exploration of how function and representation vary across activation space. An approach that accounts for these considerations is warranted. In this work, we focus on four heavily studied convolutional neural networks: AlexNet, VGG16, MobileNetV2, and ResNet101. We efficiently group units by their positions in activation space along two dimensions: average class selectivity and average activation magnitude, to highlight the regions occupied by function and representation. This paper documents these contributions:

- We demonstrate a reliable and fast method of grouping units in activation space. This method improves on past clustering work [31] by using a simpler algorithm which clusters units into groups in activation space.
- We show that many groups of units in activation space produce high linear classification probe accuracies with higher accuracies for groups of units that are selective for the target and highly selective for not-the-target.
- We identify a region of activation space in which ablation selectively reduces performance. This region partially overlaps with parts of activation space that produce strong linear classification. This partial overlap falsifies the specificity of linear probes for isolating function.
- We show that this region is consistent across four popular convolutional neural networks.

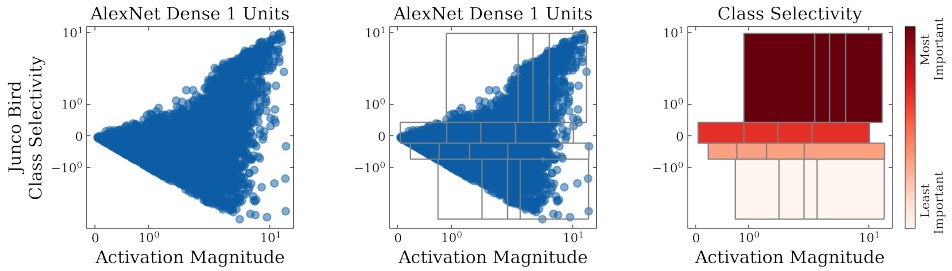
## 2 Methods

For our experiments, we investigated AlexNet [32], VGG16 [33], MobileNetV2 [34], and ResNet101 [35] all pre-trained on ImageNet [36][2]. We modified these pre-trained networks to include lambda masking layers which allowed us to selectively ablate individual units or unit groups in the network. Ablating a unit sets its outgoing activation to zero, preventing information flow from that unit to the next layer. Our process of relating linear probe accuracies and causal network performance for unit groups consists of two steps: 1) Selection of unit groups, 2) Measuring linear probe accuracies and causal performance through ablation, outlined in the following sections.

### 2.1 Selection of unit groups

This section outlines a cheap algorithm for grouping important units in activation space, this allows for all experiments to be run on a laptop, minimizing environmental impact. This algorithm groups units on a reduced activation space constructed using activation characteristics that were implicated in performance based on previous ablation studies. To find groups of units important for class-specific performance in AlexNet, we measured activations for each unit in the network for a set of 2500 images representing 50 classes (e.g., junco bird, brain coral, etc.) drawn from the ImageNet validation set. We extended our findings to three other networks using a subset of 8 image classes. We averaged the activity for each unit for our class of interest and across a random subset of other

[2] AlexNet pre-trained weights from `http://github.com/heuritech/convnets-keras`. Other pre-trained models downloaded from `http://keras.io/api/applications`




(a) (b) (c)

Figure 1: **Clustering of units by selectivity and magnitude. a)** An example of the reduced activation space for a target class. The vertical axis measures a unit's average selectivity for the class. The horizontal axis measures a unit's average magnitude of activation. **b)** We constructed a grid to overlay on the activation space so that each grid cell contained the same number of units. **c)** We predict that the most selective unit groups are important for linear probe and causal network performance.

classes separately for each layer of the network. We then plotted neurons in this two-dimensional activation space with average activation for our class of interest along one dimension and average activation for other classes across the other dimension.

We then rotated this space to form two new dimensions: class selectivity and activation magnitude (Figure 1a). In this space, class selectivity measures the difference in the average activation of each unit between our class of interest and other classes. Activation magnitude, on the other hand measures each unit's average activation in response to all images. Over this space we constructed a 4x4 grid with rectangular cells. The cells vary in area, but each contains an equal number of units. These cells were constructed by first partitioning the space into four horizontal strips at intervals along the selectivity dimension, each with the same number of neurons. These strips were then individually partitioned into four quadrants each along the magnitude dimension so that each of the sixteen resulting cells contained the same number of neurons. Controlling for the number of units in each cell allowed for easy comparisons between cells. Figure 1b shows the constructed grid overlaid on the activation space. Cells were selected one at a time and all the units from that cell were used to measure linear probing accuracy and causal performance deficits, as outlined in the next subsection.

### 2.2 Measuring linear probe accuracies and causal performance deficits

To measure linear probe decoding accuracies, we followed the methodology presented in the original paper [12], fitting linear probes to each cell's unit group to predict class identities. For comparison, the importance of each cell's unit group for class-specific task performance was measured using ablation. First, before ablation, the network's baseline performance was recorded. During inference on each image, classes were ranked according to the magnitude of their corresponding softmax outputs. For instance, a rank of three for a particular image denotes that the network predicted that the correct class was the third most likely class given the image. After recording baseline ranks, every unit in a given cell was ablated using the lambda mask mentioned previously (see Section 2). A performance impact score was calculated for each image by measuring how many ranks the correct class fell compared to baseline, which we refer to as rank deficits. Selective ablation rank deficit scores were calculated by averaging the performance impact scores across the class of interest. A network's top-K classification accuracy depends exclusively on the rank of the correct class, making the rank deficit score a direct measure of the impact of ablation on network performance.

## 3 Results

We show grids colored by linear probe accuracies for each image class in our set (rows in Figure 2a) and for each layer in AlexNet (columns in Figure 2a). Correspondingly, we show performance deficit grids colored by rank deficit for each image class and each layer (Figure 2b). The last row depicts the average grid scores for each layer averaged across all 50 classes. The final column depicts the average grid scores for each class averaged across all seven layers of AlexNet. In all cases, the colorbar displays the range in scores for each class. In addition to AlexNet, we show results for

three other networks: VGG16, MobileNetV2, and ResNet101. Figure 3 depicts the ablation and linear probe grid scores averaged across all layers for each network we tested.

### 3.1 Linear probe accuracy

Linear probe accuracy is somewhat localized in activation space (Figure 2a). In general, across all eight image classes, unit groups with maximum accuracy scores are located in the most selective regions of activation space. High linear-probe scores come from cells that are both highly selective for the target class and highly selective on average for all other classes over the target class. Importantly, deleting these neurons has no effect on the networks ability to identify the class, as shown in Figure 2b. Overall, like ablation impacts, decoding accuracies tend to increase with depth.

### 3.2 Rank deficit score

Rank deficits are localized in activation space. In general, across all image classes, unit groups with a maximum rank deficit are located in the most selective and most active regions of activation space (Figure 2b). In general the most selective and active unit group produces rank deficits far greater than neighboring unit groups. This pattern shows a combination of selectivity and magnitude contribute to unit importance. This relationship appears discontinuous rather than graded. Even though high impact cells consistently emerge in the most selective and active regions of activation space, the layer with the highest impact cell varies across classes. For instance, the junco bird class has the highest impact cell in the penultimate layer of the network (Dense 2), while the bookcase class shows the highest rank impacts in Conv 3. Additionally, rank deficit scores vary considerably between layers and between classes. Ablating the most selective and most active cell in Conv 3 produces a rank deficit score of nearly 300 for the bookcase class. But, junco class rank deficits never exceed 60.

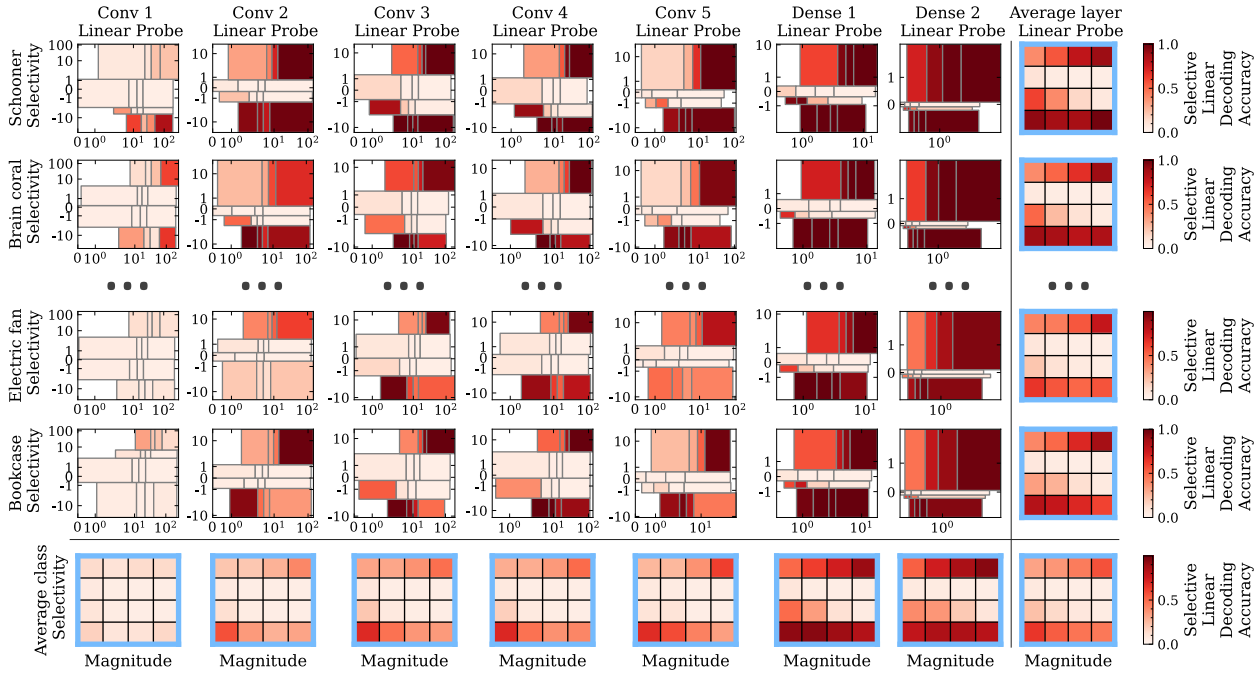
### 3.3 Only a subset of selective unit groups contribute to causal function

We found the units that are the most active while also being selective are the strongest drivers of performance across all four networks. This pattern is most prominent in AlexNet and VGG16. In ResNet101 and MobileNetV2, the most selective and active cell continues to serve as the strongest driver of performance, but unlike VGG16 and AlexNet, other cells become increasingly important. Particularly, high magnitude cells are consistently important in MobileNetV2. MobileNetV2 and VGG16 both utilize batch normalization layers which have been shown to spread out class-specific information across units [22]. Further investigation is needed to establish explicit links between regularizers, like batch normalization, representation, and performance.

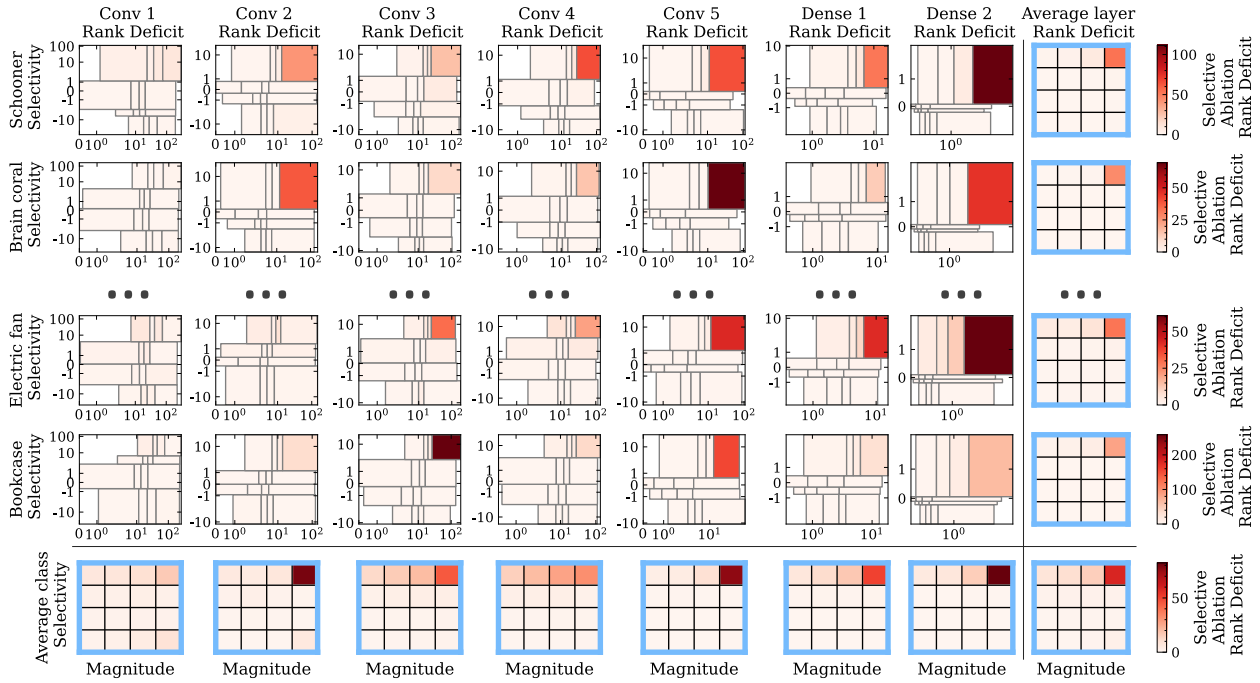
## 4 Discussion

In this work, we used a computationally inexpensive method for selecting important groups of units in neural networks. Ablating these groups revealed a consistent pattern whereby the most important unit groups emerge in highly selective and highly active areas of activation space. Previous ablation studies may have failed to identify this pattern for the following reasons. Morcos et al. [6] and Zhou et al. [22] ablate single units across the whole network based on class selectivity alone, but not activation magnitude. In Figure 2b, selectivity alone would have a weak effect on performance because it averages both strong and weak effects, one potential explanation for the weak relationship between selectivity and performance from [22]. In contrast, Meyes et al. [21] visualize both class selectivity and class-specific magnitudes, but not per layer. Since units from early layers are typically more active and selective, clusters in activation space containing units from all layers lump together highly selective and active units from later layers with less selective and active units from early layers. In this way, network-wide groups are too heterogeneous to reveal patterns that exist primarily at the layer level. This insight is leveraged by Liu et al. [26] to find highly impactful neurons using selectivity (and other promising measures) for small MLPs. Here, we show that only the most selective and active groups of units contribute to causal performance in four large CNNs.

Rather than ablating units, most neural network interpretability studies rely on interpreting function from internal representations. Alain et al. [12] proposed using linear classifier probes to extract the semantic concepts learned by specific units. We use linear probes to show that the most selective




(a) Linear Probe




(b) Rank Deficit

Figure 2: **Linear decoding and ablation scores do not match.** Figure 2a) Linear probe scores for each class (rows) and each layer of AlexNet (columns). For each class and layer combination we show accuracies obtained from decoding the target class using linear classifier probes on units in each cell. Figure 2b) Rank deficit scores for each class (rows) and each layer of AlexNet (columns). For each class and layer combination we show class-specific rank deficits caused by ablating all the units in each cell of the grid placed on activation space. Highly selective regions, both for the target class and for all other classes over the target class, consistently produce the highest decoding scores (2a), whereas highly impactful neurons tend have high class selectivity and magnitude (2b). Average effects of class and layer are shown at the edges to assess consistency of a particular pattern.

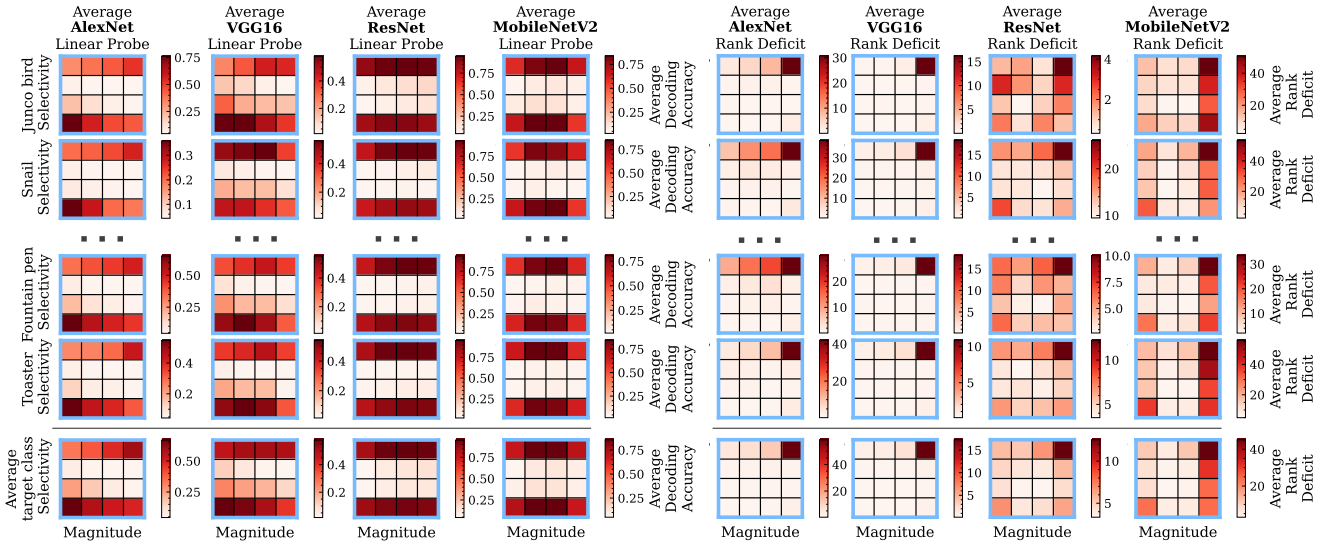



Figure 3: **High ablation impact and linear probe accuracy neurons occupy different regions of activation space across four CNNs.** Linear probe accuracies (left) and ablation deficit scores for each network averaged across all layers. For every network, the most selective and active cell produces the highest selective ablation rank deficits. This pattern is most prominent for AlexNet (50 classes) and VGG16 (8 classes), where few other cells contribute to performance. However, in the case of ResNet101 (8 classes) and MobileNet2 (8 classes), other cells can produce high rank deficits. Linear probe accuracies exhibit a distinctly different pattern whereby highly selective cells (both for and against the target class) produce the highest decoding accuracies.

units in activation space do indeed contain the most decodable semantic information. However, ablating these highly selective units in activation space does not necessarily produce the largest performance deficits. In fact, only a relatively small subset of selective units produce performance deficits. Across four popular CNNs, high ablation impact neuron groups tend to have both high selectivity and magnitude. This finding suggests that interpretability methods which rely on linear probes may be seeking out directions in activation space with only minimal effects on network performance. We improve on past work by identifying a region of activation space crucial for interpreting function.

## 5 Limitations

Limitations in this work stem from three assumptions made in our methods. First, we only investigate class-specific selectivity, while other studies have emphasized that class-specific selectivity can play a different role in network function from low-level feature selectivity [37]. Additionally, other importance measures (i.e. mutual information [26]) are not studied here, but should be the focus of future study. Second, our unit grouping method relies on handpicked hyperparamters that determine the number of cells and produces cells that contain different numbers of units in each layer, making it difficult to compare scores across layers. Lastly, we study the performance changes at the output of the network induced by ablation, while neglecting other causal changes in the rest of the network.

## 6 Conclusion

This work contributes directly to the interpretability of neural networks. We demonstrate a reliable and fast method of grouping units in activation space. We identify a region of activation space in which ablation selectively reduces causal performance, consistent across four popular CNNs. Surprisingly, we found that important ablation impact units only occupy a relatively small subset of the activation space covered by selectively important units identified by linear probes. As a result, interpretability tools that rely on decoding semantic information likely only weakly relate to performance. Future investigations would benefit from extending this research to more networks and tasks to uncover theoretical relationships between neural network representation and function.

## Acknowledgments and Disclosure of Funding

This work was supported by the University of Colorado Boulder Computer Science Department.

## References

[1] Tom Simonite. When it comes to gorillas, google photos remains blind. *Wired*, Jan 2018. ISSN 1059-1028. URL https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

[2] Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, Mar 2020. ISSN 0306-6800.

[3] EU 2016/679. General data protection regulation, 2016.

[4] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, page 233–242. PMLR, Aug 2017.

[5] 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.

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

[7] M. Raghu, T. Unterthiner, and S. Kornblith. Do vision transformers see like convolutional neural networks? *Advances in*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html.

[8] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[9] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[10] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, John E Hopcroft, et al. Convergent learning: Do different neural networks learn the same representations? In *FE@ NIPS*, pages 196–212, 2015.

[11] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34, 2021. ISSN 1049-5258. URL https://papers.nips.cc/paper/2021/file/01ded4259d101feb739b06c399e9cd9c-Paper.pdf.

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

[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.

[14] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2016.

[15] Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926, 2018.

[16] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[17] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1058–1067, 2017.

[18] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[19] Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Cote, and R Devon Hjelm. Unsupervised state representation learning in atari. *Advances in neural information processing systems*, 32, 2019.

[20] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

[21] Richard Meyes, Constantin Waubert de Puiseau, Andres Posada-Moreno, and Tobias Meisen. Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations. *arXiv preprint arXiv:2004.01254*, 2020.

[22] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Revisiting the importance of individual units in cnns via ablation. *arXiv preprint arXiv:1806.02891*, 2018.

[23] Jonathan Donnelly and Adam Roegiest. On interpretability and feature representations: an analysis of the sentiment neuron. In *European Conference on Information Retrieval*, pages 795–802. Springer, 2019.

[24] Matthew L Leavitt and Ari Morcos. Selectivity considered harmful: evaluating the causal impact of class selectivity in dnns. *arXiv preprint arXiv:2003.01262*, 2020.

[25] Yuta Kanda, Kota S Sasaki, Izumi Ohzawa, and Hiroshi Tamura. Deleting object selective units in a fully-connected layer of deep convolutional networks improves classification performance. *arXiv preprint arXiv:2001.07811*, 2020.

[26] Kairen Liu, Rana Ali Amjad, and Bernhard C Geiger. Understanding individual neuron importance using information theory. *arXiv preprint arXiv:1804.06679*, 2018.

[27] Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317, 2019.

[28] Ming Tu, Visar Berisha, Yu Cao, and Jae-sun Seo. Reducing the model order of deep neural networks using information theory. In *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 93–98. IEEE, 2016.

[29] Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Group fisher pruning for practical network compression. In *International Conference on Machine Learning*, pages 7021–7032. PMLR, 2021.

[30] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

[31] Shlomi Hod, Stephen Casper, Daniel Filan, Cody Wild, Andrew Critch, and Stuart Russell. Detecting modularity in deep neural networks. *arXiv preprint arXiv:2110.08058*, 2021.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[35] 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.

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[37] Jumpei Ukita. Causal importance of low-level feature selectivity for generalization in image recognition. *Neural Networks*, 125:185–193, 2020.