Isolating sparse, category-computing circuits in deep neural networks

Jeffery W. Andrade (jandrade01@g.harvard.edu)
Harvard University, 33 Kirkland St, Cambridge, MA 02138

Talia Konkle (talia_konkle@harvard.edu)
Harvard University, 33 Kirkland St, Cambridge, MA 02138

Abstract

Humans can recognize many object categories: what are the underlying computational routes from retina to category-level representations that enable this capacity? Deep neural networks are now remarkably competent at visual categorization, and as such can serve as a powerful model system for dissecting the hierarchical processing of visual inputs. In this work, we investigate the computations underlying individual category recognition in CNNs: are all unit-to-unit connections across the layers required to categorize an object, or might more dissociable sparse, modular circuits learned in the network support this task? Extending work on CNN circuit extraction (Hamblin, Konkle, & Alvarez, 2023), we have developed a procedure for identifying functional subcircuits within Alexnet that are important for category discrimination. Our algorithm assigns scores to connections based on their estimated contribution to a category unit's activation pattern and prunes the lowest-scored connections up to a chosen *circuit substitution accuracy* threshold on an extraction imageset. We then evaluate the resulting circuits for function preservation on new images, and analyze the structure of the resulting category circuits. When pruning to an allowed small *circuit substitution accuracy* decrement, we find surprisingly sparse, substantially faithful circuits with an average circuit sparsity of 45.3% and an average *circuit substitution accuracy* of 90.9% that of the unpruned network. These results indicate that category-level representations individually depend upon relatively sparse subnetworks, suggesting a semi-modular neural code with significant, structured sharing of circuitry.

Introduction

Visual input to the retina and early visual system is transformed across hierarchical processing stages into category-selective representations in the higher visual system. A spectrum of theoretical frameworks have been developed for understanding this process, from distributed coding and untangling in large-scale population codes (Haxby, Gobbini, Furey, Ishai, Schouten, & Pietrini, 2001; DiCarlo & Cox 2007), to more category specialization processed in specialized circuits (Kanwisher 2010; Grill-Spector & Weiner, 2014). For example, a fully distributed code implies that every neuron and synapse in the population contributes in some way to the computation of a categorical representation, forming dense computational *circuits*. On the other extreme, strong specialization accounts posit that some parts of the population are relevant to one and only one kind of category. Recently, an intermediate "routing" framework has been proposed that strikes a balance between these accounts: partial separability exists in the population code, where some parts of the population are more relevant for some categories relative to others, with increasingly separable circuit computations over successive hierarchical transformations (Prince, Alvarez, & Konkle, 2024).

Unfortunately, in human brains and larger biological systems, the full visual circuit architecture is not known, as methods do not yet exist for mapping hierarchical circuits across all neurons within and across areas. Fortunately, modern CNNs–whose internal connectome is fully accessible–are very competent at completing visual categorization tasks, allowing us to at least examine how one model system accomplishes image-to-category level representation through circuit computations. Further, as many of these models show emergent representational similarity to the human visual system (Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, & DiCarlo, 2018; Conwell, Prince, Kay, Alvarez, & Konkle, 2024), understanding the "functional neuroanatomy" of a deepnet offers promise that the principles of circuit computation may lead to insights and predictions about visual representation routing in biological systems.

Method

Here, we developed a method to dissect out subcircuits of Alexnet that preserve the function of each of the 1000 output category units, extending work from Hamblin, Konkle, & Alvarez, 2023.

Weight scoring

For each category, we compute an importance score for all of the model weights as follows: For each image in a selected scoring set (chosen from the imagenet training set), the activations of each weight are multiplied by the gradient of the target unit's activation with respect to the weights' activations, yielding a linear estimate of the weight's contribution to the target unit's activation. This is averaged over images to give an estimated importance score per weight. For fully connected layers, these scores are then minmax normalized to be between 0 and 1. Empirically, 4 images per category was the minimum quantity needed for stable weight score rank order.

Circuit extraction

Next, for each output unit, the model weights are pruned to 0 from lowest to highest scored, to arrive at a category-specific circuit of a select sparsity. To determine how much we can prune while "preserving the function" of the unit, we prune only until *circuit substitution accuracy* (CSA) falls below a threshold.

The idea is relatively intuitive: any target output unit has a certain prediction accuracy over images when using the full model architecture; if we substitute in activations for the target unit from the pruned circuit instead, then a good circuit should still classify the image correctly. Using an extraction set of 20 images from each category taken from the imagenet evaluation set, we thus compute the CSA measure as the fraction of times that the network predicted the target category and it was the correct prediction. We then correct this measure for false alarms by subtracting the fraction of times that the target category was predicted when it wasn't present.

One challenge is that the pruned circuit may have a global shift in activation levels; thus we train one free parameter to set the overall bias of the pruned circuit output unit. Our final circuit extraction procedure is as follows: for each target output unit we sweep over increasing circuit sparsities, optimizing the target unit bias to maximize the CSA. This is first done on the fully-connected layers until CSA falls below a chosen threshold, and is then repeated for the convolutional layers, resulting in a single extracted circuit per output unit.

Results and Discussion

Across categories, this method yields sparse circuits that maintain the target unit function across 30 new images from each category from the imagenet evaluation set. We find that we can extract circuits with an average SA of 47.5%, 90.9% that of the unpruned network's average accuracy of 52.3%, and with an average sparsity of 45.3% (and a standard deviation of 16.20%) (fig. 1a).

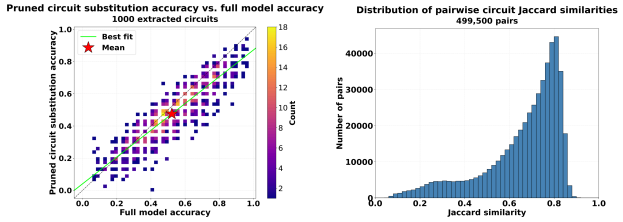

Figure 1: a). Pruned circuit CSA vs. original network accuracy for all output category circuits
b.) Histogram of Jaccard similarity values for all pairs of circuits in the network.

Given the circuits produced by our extraction procedure, what structure and dissociations are present in these circuits? As a first analysis, we consider the amount of circuit overlap between any two categories. We quantified this with the intersection over the union of the circuit masks (Jaccard similarity). A histogram of all such overlaps is shown in fig. 1b, where some categories demonstrate highly dissociable circuits, while most show highly similar computation paths.

These results set the stage for making predictions about which visual categories might interfere more or be processed in parallel (Cohen, Konkle, Rhee, Nakayama, & Alvarez, 2014), and generally understanding the computational emergence of category level visual representations.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1942438. Additionally, this work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute at Harvard University. We gratefully acknowledge the support and contributions from all members of the Harvard Vision Lab.

References

Cohen, M. A., Konkle, T., Rhee, J. Y., Nakayama, K., & Alvarez, G. A. (2014). Processing multiple visual objects is limited by overlap in neural channels. *Proceedings of the National Academy of Sciences*, *111*(24), 8955-8960.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. Nature communications, 15(1), 9383.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, *11*(8), 333-341.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536-548.

Hamblin, C., Konkle, T., & Alvarez, G. (2022). Pruning for Feature-Preserving Circuits in CNNs. *arXiv preprint arXiv:2206.01627*.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-2430.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, *107*(25), 11163-11170.

Prince, J. S., Alvarez, G. A., & Konkle, T. (2024). Contrastive learning explains the emergence and function of visual category-selective regions. *Science Advances*, *10*(39), eadl1776.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like?. BioRxiv, 407007.