On the universality of neural codes in vision

Anonymous Author(s)
Affiliation
Address
email

Abstract

A high level of similarity between neural codes of natural images has been reported for both biological and artificial brains. These observations beg the question whether this similarity of representations stems from a more fundamental similarity between neural coding strategies. In this paper, we show that neural networks trained on different image classification datasets learn similar weight summary statistics. Our results reveal the existence of a universal neural code for natural images.

1 Introduction

Deep neural networks reliably achieve high performance on visual tasks such as image classification, with remarkable robustness with respect to the exact details of the architecture, initialization, and training procedure. The success of transfer learning also demonstrates that networks trained on one task can perform well on another (related) task. This raises the question: do all these networks share a universal encoding of images, irrespective of their architecture and training dataset? And if so, is this encoding shared in human and animal visual cortex?

Hidden representations learned by networks trained from different initializations have been found to be similar at all layers [Raghu et al., 2017], and this similarity increases when the network width increases [Kornblith et al., 2019]. Similar observations in the context of human neural encodings have been made by studying fMRI response patterns in visual cortex [Haxby et al., 2011], as well as between neural network representations and IT spiking responses [Yamins et al., 2014]. Here, we ask whether this similarity at the level of network representations/activations arises from a more fundamental similarity between their learned weights.

Network weights are less easily prone to analysis than hidden representations, and have thus been less studied. The first layer can be directly visualized, and learns Gabor-like filters [Krizhevsky et al., 2012] in a wide range of settings. Attempts to generalize comparisons to deeper layers based on matching individual neurons between networks however lead to mixed results [Entezari et al., 2022, Benzing et al., 2022, Ainsworth et al., 2022]. The recent work of Guth et al. [2023] instead considers global weight statistics and shows that they do not depend on the initialization nor the network width.

In this paper, we extend these results by showing that networks trained on different image classification datasets share a set of universal weight statistics even deep within the network. In Section 2, we first review the approach of Guth et al. [2023] to compare weights in hidden layers between different networks. We then show in Section 3 that this approach applied to networks trained on different datasets reveals the universality of the learned weights.

2 Comparing weights of deep networks

How does one meaningfully compare weights of two trained deep networks? In this section, we briefly review the approach introduced by [Guth et al., 2023].

Separating space and channels. Weights of convolutional layers take the form of a four-dimensional tensor with two spatial dimensions and two channel dimensions. To simplify the analysis of the weights, it is helpful to consider a family of architectures for which operations across space and channels are separated. We consider learned scattering networks [Guth et al., 2022], in which the spatial convolutions are not learned and fixed to wavelets, yet achieve classification accuracies on par with ResNets of similar depths on ImageNet. We thus focus on the learned weights which apply along channels only, through a 1×1 (or pointwise) convolution. Our approach is however general and can be applied to any CNN with appropriate modifications [Guth et al., 2023].

Aligning hidden layers. Visualizations and comparisons of learned weights in deep networks are generally limited to the first layer. A major difficulty in comparing weights in deeper layers is that they are adapted to hidden representations which themselves vary across networks, contrary to the input of the first layer which is fixed.

Comparing two layers can be done as follows. Consider two hidden representations $\phi(x)$ and $\phi'(x)$ learned by two different networks: $\phi(x)$ and $\phi'(x)$ are in general not comparable (they might even have different numbers of dimensions). Representational similarity analysis [Kriegeskorte et al., 2008] instead compares their similarity structures (or kernels) $\langle \phi(x), \phi(y) \rangle$ and $\langle \phi'(x), \phi'(y) \rangle$, which have empirically been found to be close in various settings [Raghu et al., 2017, Kornblith et al., 2019]. This implies that the variability in the representation between ϕ and ϕ' must preserve this similarity structure, and is thus limited to an orthogonal transform. In other words, when $\langle \phi(x), \phi(y) \rangle \approx \langle \phi'(x), \phi'(y) \rangle$, there exists an orthogonal alignment matrix A such that $\phi'(x) \approx A\phi(x)$ [Guth et al., 2023].

Now consider two neurons w and w' in the next layer of the two different networks. What does it mean for w and w' to be equivalent? It seems natural to ask that the two neurons compute similar outputs:

$$\langle w, \phi(x) \rangle \approx \langle w', \phi'(x) \rangle.$$

Because $\phi(x) \neq \phi'(x)$, this condition is not equivalent to $w \approx w'$. Rather, we have

$$\langle w', \phi'(x) \rangle \approx \langle w', A\phi(x) \rangle = \langle A^{\mathrm{T}} w', \phi(x) \rangle,$$

so that the two neurons compute similar outputs when $w \approx A^{\mathrm{T}} w'$, or equivalently, when $w' \approx Aw$. Just like the alignment A maps representations in the first network to representations in the second network, it maps next-layer neurons in the first network to equivalent neurons in the second network. Comparing hidden neurons from different networks thus requires aligning their hidden representations and taking this alignment into account in the comparison.

Comparing neuron distributions. Comparing individual neurons in two different networks amounts to searching for a one-to-one mapping between them. If the two networks had exactly the same neurons, but possibly in a different order, then their representations would differ by a permutation [Entezari et al., 2022, Benzing et al., 2022, Ainsworth et al., 2022]. The use of rotations when aligning representations suggests that more variability might be present.

Rather than comparing individual neurons from two different networks, we search for similarities between the neural populations at a global level: do they have the same statistics? This corresponds to testing whether the neurons in both networks can be modeled as samples from the same distribution. as done in so-called "mean-field" analyses of neural networks.

Weight principal directions. When considering probability distributions, which statistics of the neural populations should we measure and compare? Guth et al. [2023] have shown that the covariance of neuron weights, and in particular its leading eigenvectors, captures most of the encoding properties, as knowledge of the weight covariances can be sufficient to generate new networks with similar performance.

In summary: to compare weights between two networks, for each layer, we compute the alignment matrix A between the input representations, and use it to compare the covariances of the neuron weights of both networks. In Figure 1, we reproduce some of the results of Guth et al. [2023]. We show that two networks with different random initializations learn the same weight eigenvectors when trained on CIFAR-10. This shows that the leading eigenvectors of the weight covariance correspond to a stable low-dimensional "informative" subspace. In the next section, we extend this result to networks trained on *different* datasets.

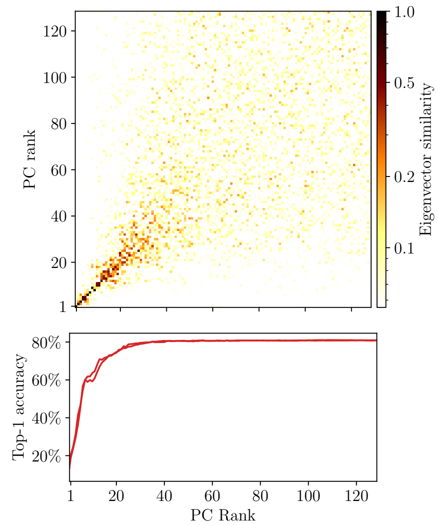

Figure 1: Comparison between the weight covariance eigenvectors of two four-layer learned scattering networks trained on the CIFAR-10 dataset (after alignment). We focus on the second layer (all layers lead to similar results). **Top:** matrix of pairwise cosine similarities between covariance eigenvectors of the two networks. High values along the diagonal at low ranks indicate that the same leading eigenvectors are learned by both networks. **Bottom:** classification accuracy after projection of the neuron weights in the subspace spanned by the top r eigenvectors as a function of maximal rank r. A relatively small fraction of the eigenvectors is sufficient to achieve the maximal performance for both networks. This fraction coincides with the number of eigenvectors that are stable to the random initialization.

3 Universality of weight eigenvectors learned from natural images

In order to evaluate the universality of the learned neural codes, we train the same eight-layer learned scattering architecture on various image classification datasets which vary in the number and diversity of their image classes. We consider subsets of CIFAR10, CIFAR100, and ImageNet (downsampled to 32×32 resolution for direct comparison with the same architecture). In particular, CIFAR5 is the subset of CIFAR10 composed of the first 5 classes, while ImageNet100a and ImageNet100b are two subsets of ImageNet composed of 100 random classes. Naturally, networks trained on classification tasks with more classes learn a higher number of relevant weight eigenvectors. A meaningful learned encoding comparison therefore requires considering networks trained on tasks with similar numbers of classes. This justifies the choices of the pairs of datasets presented in Figure 2.

Using the same image resolution and architecture for all datasets ensures that all networks have the same number of layers and receptive field sizes. This allows comparing each layer independently. The results are shown in the figure. Interestingly, we find universal weight eigenvectors over an appreciable range of layer depth. We observe this universality of learned weight eigenvectors to vanish towards the final classifier. This is expected to happen at some level as the representation has to become task-specific. Further, we find that the more challenging datasets (with more classes, or more diversity such as ImageNet100 as opposed to CIFAR100) lead to a richer encoding with a higher number of universal weight eigenvectors.

The existence of this universality has a number of fascinating implications. It suggests that the training procedure of artificial networks could be significantly simplified as networks could be preset with these generic features. It also opens the door to quantitative comparisons between datasets through a definition of encoding dimensionality or complexity. If such results also apply to biological brains, it suggests that different individuals may encode the visual world in a very similar manner.

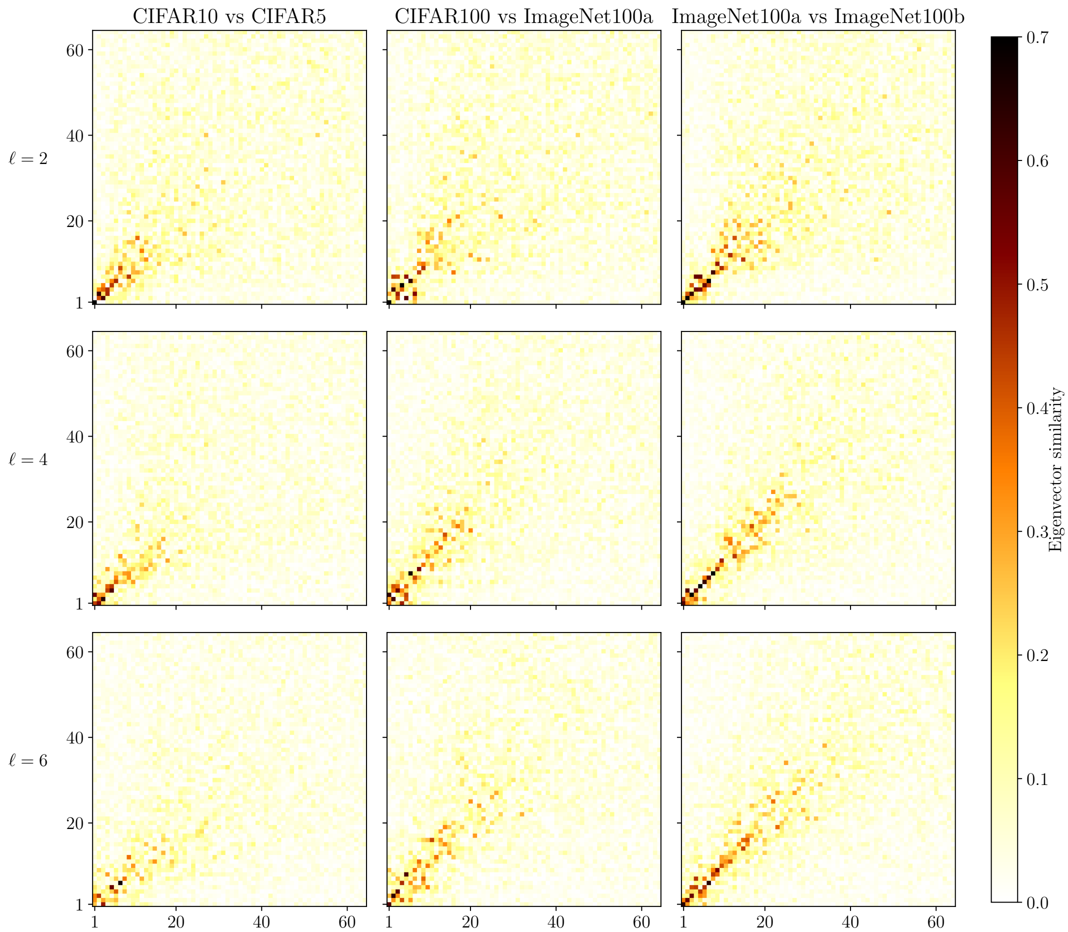

Figure 2: Universality of the leading covariance eigenvectors. We train several eight-layer learned scattering networks on various image classification datasets. We compute pairwise cosine similarities between weight covariance eigenvectors at several layers (**in rows**) for several dataset pairs (**in columns**). The color scheme has been cut off at 0.7 to better represent the dynamic range of the correlations. In the three cases, the low-rank eigenvectors are similar across datasets, even for deeper layers. The number of such eigenvectors increases with the number of classes.

References

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Frederik Benzing, Simon Schug, Robert Meier, Johannes Von Oswald, Yassir Akram, Nicolas Zucchet, Laurence Aitchison, and Angelika Steger. Random initialisations performing above chance and how to find them. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Florentin Guth, John Zarka, and Stéphane Mallat. Phase Collapse in Neural Networks. In *International Conference on Learning Representations*, 2022.

Florentin Guth, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. A rainbow in deep network black boxes. *arXiv preprint arXiv:2305.18512*, 2023.

James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

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.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

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.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.