High-dimensional alignment of neural networks and visual cortex

Tailai Shen (tshen15@jh.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Colin Conwell (cconwel2@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Michael F. Bonner (mfbonner@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Abstract

Research into the representational similarity between deep neural networks (DNNs) and the human visual cortex aims to deepen our understanding of both systems. Here we explored the alignment between DNNs and the ventral visual stream by extending conventional representational similarity statistics to a spectrum of similarities across thousands of latent dimensions. The spectrum is generated by computing the correlations between aligned latent dimensions in model and brain representations. Using this approach, we found that DNN layers and regions of visual cortex have shared high-dimensional representations, spanning thousands of dimensions. The dimensionality of these shared representations exhibits an overall decrease from early to late visual regions. However, by separately reducing the channel and spatial dimensions of DNNs, we found that there is a complex relationship between dimensionality and the visual hierarchy. Specifically, in early visual regions, the alignment with DNNs relies heavily on high *spatial* dimensionality, whereas in late visual regions, it relies heavily on high *channel* dimensionality. Together, these results demonstrate the potential insights that can be gained by characterizing the full spectrum of high-dimensional alignment between computational models and visual cortex.

Keywords: vision; deep learning; deep neural network; latent dimension; visual representation; representational alignment

Introduction

Deep neural networks (DNNs) serve as a computational framework for modeling the human brain (Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Richards et al., 2019). In particular, deep convolutional neural networks have been found to be predictive of neural responses in the visual cortex (Yamins et al., 2014; Güçlü & Gerven, 2015; Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017). Various metrics are available to study the alignment between regions of visual cortex and DNN representations. These include representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008), voxel-encoding RSA (veRSA) (Kaniuth & Hebart, 2022; Konkle & Alvarez, 2022), and average voxel-wise encoding score (Elmoznino & Bonner, 2022), which are often reduced to a single summary statistic for a region of interest.

However, previous studies have uncovered power-law eigenspectra in the latent geometry of model representations (Ghosh, Mondal, Agrawal, & Richards, 2022; Kong, Margalit, Gardner, & Norcia, 2022) and brain representations (Stringer, Pachitariu, Steinmetz, Carandini, & Harris, 2019; Gauthaman, Ménard, & Bonner, 2023), highlighting the significance of high latent dimensions in understanding these representations—something a single statistic cannot capture. Here, we introduced a novel metric that quantifies the similarities between visual cortex and DNN representations across many latent dimensions. This involves projecting both representations into a common space and extracting the shared information across a spectrum of dimensions. We applied this metric across different levels of the visual hierarchy and model layers to conduct a detailed analysis of representational alignment. Additionally, we manipulated the number of channel and spatial dimensions in the model representations to examine their impact on alignment.

Methods

To obtain model representations, we extracted layer activations from a DNN with a ResNet50 architecture trained on ImageNet (Russakovsky et al., 2015; Conwell, Prince, Kay, Alvarez, & Konkle, 2023). We obtained brain representations for areas V1-V4 and the high-level ventral stream from the Natural Scenes Dataset (NSD) (Allen et al., 2021). This dataset includes fMRI responses from eight human subjects who each viewed 10,000 natural images. The analyses on different visual regions and model layers used all images seen by each subject. For the analysis of reduction in channel and spatial dimensions, we used 5,000 images due to computational limitations. We performed separate reductions of channel dimensions and spatial dimensions via sparse random projections.

We computed a spectrum of representational alignment using the cross-decomposition method described in previous work (Gauthaman et al., 2023). Briefly, this method applies singular value decomposition to the cross-covariance matrix of the model and brain activations, which identifies a common space of latent dimensions that maximize the covariance between the two systems. Singular vectors are learned on a training set of images and then used to project a held-out set of test images into the common space. We then correlated the paired latent dimensions of the model and brain in the test set to generate a spectrum of correlations.

Results & Discussion

The cross-decomposition spectra derived from comparing brain and model activations show that these systems have shared information across many latent dimensions (Fig. 1B), and they further demonstrate that the dimensionality of these shared representations is highest in early regions of the cortical hierarchy (V1-V4) and decreases in later regions (high-level ventral stream). These findings show that the DNN and visual cortex can be aligned along many orthogonal latent dimensions with reliable shared variance, allowing us to examine the full spectrum of shared representations between these systems. We next performed follow-up analyses to characterize key properties of these shared dimensions.

First, when plotting these spectra for all model layers, we observe a hierarchical correspondence between model depth and early vs. late regions of visual cortex (Fig. 2). Specifically, V1-V4 is better explained by early and intermediate model layers, whereas the high-level ventral stream is better explained by deeper layers. This hierarchical correspondence was observed across the entire spectrum of shared dimensions. However, we also noted an interesting trend across

ranks of latent dimensions, whereby differences across model layers are more pronounced at higher-rank dimensions. This suggests that while all layers explain relatively similar variance at low ranks, only the best-matching layer for a region explains reliable variance in higher dimensions.

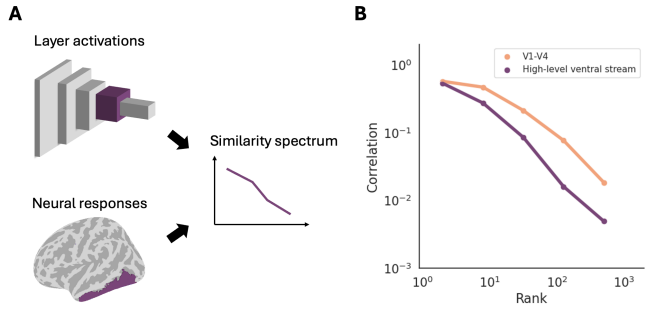

Figure 1: A. We computed the representational similarity spectrum between layer activations from a DNN and fMRI responses from the visual cortex. B. Spectra for comparing V1-V4 and high-level ventral stream with the model layers that best align with these areas.

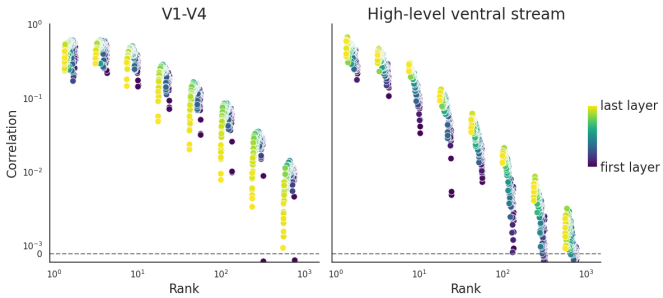

Figure 2: Spectra for comparing all model layers with areas V1-V4 and high-level ventral stream reveal hierarchical correspondence between the model and the visual cortex.

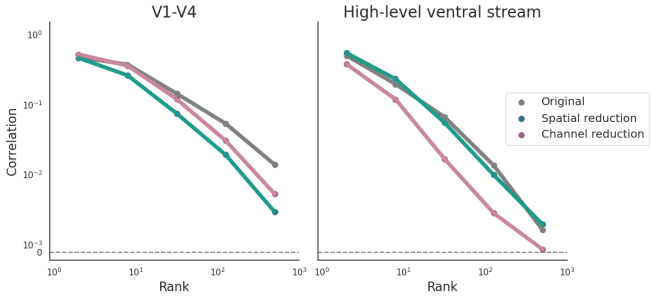

Figure 3: Selective reduction of channel and spatial dimensions in the DNN have differential effects on the spectra for areas V1-V4 and high-level ventral stream.

We next sought to better understand the differences between the spectra for V1-V4 and the high-level ventral stream. We wondered if the higher dimensional spectrum for V1-V4 was largely driven by the dimensionality of the spatial encoding in this region. To probe this, we performed dimensionality reduction in two different ways: in one, we reduced only the spatial dimensions of the DNNs and in another we reduced only the channel dimensions (i.e., features). By doing so, we were able to determine the degree to which the shared representations between the model and brain were driven by high-dimensional spatial encoding versus high-dimensional feature encoding. This analysis revealed a double dissociation (Fig. 3). The shared spectrum for V1-V4 dropped substantially when the spatial dimensionality of the DNN was reduced, but the impact of channel reduction was much smaller. In contrast, the shared spectrum for the high-level ventral stream dropped when the features dimensionality of the DNN was reduced, but was almost completely unaffected by spatial reduction. These findings reveal the need to account for the multifaceted nature of dimensionality in DNNs, whose dimensions can be factorized into spatial and feature components with distinct functional properties.

Together, this work demonstrates the high-dimensional nature of the shared representations between DNNs and visual cortex, and it highlights a methodological approach for characterizing and probing the full spectrum of shared dimensions between these systems. We note that many of the high-rank dimensions that can be revealed with this approach are relatively low-variance and would, thus, have little contribution to conventional variance-weighted metrics, like RSA and voxel-wise encoding accuracy. Approaches for examining the similarities and differences between DNNs and brains in high dimensions may open new opportunities for evaluating and understanding computational models in neuroscience.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2021, December). A massive 7t fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. Retrieved from https://doi.org/10.1038/s41593-021-00962-x doi: 10.1038/s41593-021-00962-x

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023, July). *What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?* bioRxiv. Retrieved 2023-09-07, from https://www.biorxiv.org/content/10.1101/2022.03.28.485868v2 doi: 10.1101/2022.03.28.485868

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017, May). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. Retrieved 2024-04-18, from https://www.sciencedirect.com/science/article/pii/S1053811916305481 doi: 10.1016/j.neuroimage.2016.10.001

Elmoznino, E., & Bonner, M. F. (2022, July). High-performing neural network models of visual cortex benefit from high latent dimensionality. Retrieved from https://doi.org/10.1101/2022.07.13.499969 doi: 10.1101/2022.07.13.499969

Gauthaman, R. M., Ménard, B., & Bonner, M. (2023). Revealing the high-dimensional latent structure in visual cortical representations. In *2023 conference on cognitive computational neuroscience*. Cognitive Computational Neuroscience. Retrieved from http://dx.doi.org/10.32470/CCN.2023.1652-0 doi: 10.32470/ccn.2023.1652-0

Ghosh, A., Mondal, A. K., Agrawal, K. K., & Richards, B. (2022). *Investigating power laws in deep representation learning.* arXiv. Retrieved from https://arxiv.org/abs/2202.05808 doi: 10.48550/ARXIV.2202.05808

Güçlü, U., & Gerven, M. A. J. v. (2015, July). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. Retrieved 2024-04-18, from https://www.jneurosci.org/content/35/27/10005 doi: 10.1523/JNEUROSCI.5023-14.2015

Kaniuth, P., & Hebart, M. N. (2022, August). Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, *257*, 119294. Retrieved 2024-04-18, from https://www.sciencedirect.com/science/article/pii/S105381192200413X doi: 10.1016/j.neuroimage.2022.119294

Kong, N. C. L., Margalit, E., Gardner, J. L., & Norcia, A. M. (2022, January). Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, *18*(1), e1009739. Retrieved from https://doi.org/10.1371/journal.pcbi.1009739 doi: 10.1371/journal.pcbi.1009739

Konkle, T., & Alvarez, G. A. (2022, January). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491. Retrieved 2024-04-18, from https://www.nature.com/articles/s41467-022-28091-4 doi: 10.1038/s41467-022-28091-4

Kriegeskorte, N. (2015, November). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446. Retrieved from https://doi.org/10.1146/annurev-vision-082114-035447 doi: 10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. Retrieved 2023-09-10, from https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... Kording, K. P. (2019, October). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. Retrieved from https://doi.org/10.1038/s41593-019-0520-2 doi: 10.1038/s41593-019-0520-2

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015, April). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. Retrieved from https://doi.org/10.1007/s11263-015-0816-y doi: 10.1007/s11263-015-0816-y

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019, July). High-dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), 361–365. Retrieved 2023-08-31, from https://www.nature.com/articles/s41586-019-1346-5 doi: 10.1038/s41586-019-1346-5

Yamins, D. L. K., & DiCarlo, J. J. (2016, February). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. Retrieved from https://doi.org/10.1038/nn.4244 doi: 10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, May). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. Retrieved from https://doi.org/10.1073/pnas.1403112111 doi: 10.1073/pnas.1403112111