Do deep convolutional neural networks accurately model representations beyond the ventral stream?

Dawn Finzi (dfinzi@stanford.edu)
Departments of Psychology and Computer Science, Stanford University

Daniel L. K. Yamins (yamins@stanford.edu)
Departments of Psychology and Computer Science, Wu Tsai Neurosciences Institute, Stanford University

Kendrick Kay (kay@umn.edu)
Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota

Kalanit Grill-Spector (kalanit@stanford.edu)
Department of Psychology, Wu Tsai Neurosciences Institute, Stanford University

Abstract

While primate visual cortex has typically been divided into two processing streams, recent research suggests that there may be at least three functionally distinct streams, extending along the ventral, lateral, and parietal surfaces of the brain. Here, we leveraged the Natural Scenes Dataset (Allen et al., 2022) to compare and model responses across these proposed streams. We show that cortical responses cluster by stream and reflect the hierarchical organization of cortex. We then tested how accurately deep convolutional neural networks (DCNNs) trained on supervised object categorization and action recognition objectives could predict responses in each stream. Given the differences in responses across streams and the prevailing view that only the ventral stream serves object categorization, we were surprised to find that these models fit ventral and lateral responses equally well, though they were slightly worse at predicting parietal responses. These findings suggest that additional constraints are required for model predictivity to match the functional organization of visual cortex.

Keywords: DCNNs; fMRI; high-level vision; streams

The human visual system is thought to be organized into processing streams: traditionally, this has been a "what" vs. "where", ventral vs. dorsal [1] division (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992), but more recently a third pathway has been proposed, extending along lateral occipitotemporal cortex, with hypothesized functions encompassing multimodal processing (Weiner & Grill-Spector, 2013), action recognition (Wurm & Caramazza, 2021), and social perception (Pitcher & Ungerleider, 2020). DCNNs trained on a supervised object categorization objective have been shown to be excellent predictive models of the ventral stream (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015), but it remains an open question whether they uniquely explain responses in the ventral stream. More recently, DCNNs trained for action recognition (Güçlü & van Gerven, 2017) or trained to predict an agent's self-motion (Mineault, Bakhtiari, Richards, & Pack, 2021) have been shown to predict responses in V3A/B (parietal), MT, and MST (lateral), but neither work included a comparison to DCNNs trained on object categorization. Here we sought to elucidate (1) empirically, the extent to which these three putative streams contain different representations and (2) computationally, whether DCNNs thought to be models of the ventral stream better explain ventral, lateral and/or parietal brain responses.

Methods

Data acquisition and processing in brief. We analyzed a high-resolution fMRI dataset that sampled responses to thousands of natural images in 8 individuals (Natural Scenes Dataset (NSD) (Allen et al., 2022)). We defined 7 regions of interest (ROIs): an early visual cortex (EVC) ROI, as well as intermediate and higher-level ROIs for each of the three proposed streams (Fig 1a). We used a set of 515 images shared across subjects for the representational similarity matrix (RSM) analyses (Fig 1) and 6,234 to 10,000 images per subject for modeling analyses (Fig 2). The noise ceiling (NC) was estimated in each voxel as described in Allen et al. (2022); data were thresholded to only include voxels with NC $\geq 20\%$ variance.

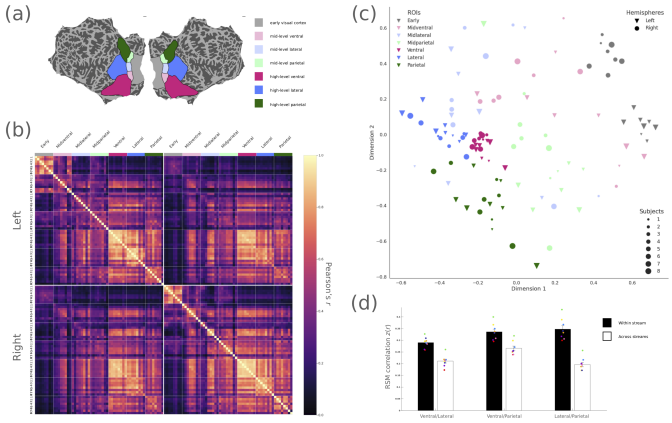

Figure 1: (a) Definition of ROIs on a flat map of the fsaverage cortex. (b) 2nd-order RSM depicting the noise-corrected similarity of distributed representations across subjects, ROIs, and hemispheres. (c) Multidimensional scaling of (b). Colors: ROI; Symbols: hemispheres; symbol size: subject. (d) Correlations among pairs of parcels between and across streams.

Comparing representations. For each individual and ROI, we computed the similarity (Pearson's *r*) between distributed responses across the ROI to all pairs of shared images, resulting in a RSM from which we extract the flattened lower triangle as a representation vector. Representation vectors were correlated across all subject and ROI combinations (corrected by the trial-to-trial reliability) to generate a 2nd-order RSM (Fig 1b), which characterizes the similarity of representations across subjects and ROIs.

Models and fitting procedure. We tested 8 candidate DCNNs - 6 models trained on the 1000-way ImageNet object categorization task: AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGG-16 (Simonyan & Zisserman, 2014), Cornet-S (Kubilius et al., 2018), ResNet-18, ResNet-50, and ResNet-101 (He, Zhang, Ren, & Sun, 2016); one untrained AlexNet; and one action recognition network: SlowFast (Feichtenhofer, Fan, Malik, & He, 2019), a dual-pathway network with a 3D ResNet-50 backbone trained on the Kinetics-400 video dataset (Kay et al., 2017). We leveraged the power of the NSD dataset to predict voxel-level responses by regressing model features from the best-fitting layer onto individual voxel responses using ridge regression. The NSD images were preprocessed using the image transforms used on the validation set during model training and model features for these stimuli were extracted from each layer of the models. As in

[1] We will refer to this stream as the parietal stream to avoid confusion with the lateral stream.

Schrimpf et al. (2020), we first projected these features into a lower dimensional space using a subsample of the ImageNet validation images and retained the first 1000 PCs. Performance was evaluated on a left-out test set (80/20 split) for each subject separately. To evaluate the upper-bound model performance given the shared variance across subjects, we calculated subject-to-subject predictivity (leave-one-out subject cross-validation) using ridge regression.

Results

Empirical testing of representation structure. Comparing representations across subject and ROI combinations (Fig 1b), we found that representations were similar across subjects within each ROI, particularly in EVC and high-level ROIs, illustrating that representations in these ROIs are consistent across individuals. This data-driven analysis provides evidence that there are indeed representational differences between these ROIs and recovers known features of the visual system. For example, EVC representations were highly correlated across subjects within, but not across, hemispheres. In contrast, representations in high-level ROIs were highly correlated both within and across hemispheres, as expected from the larger receptive field sizes in these regions that extend to the ipsilateral visual field. Visualizing this structure in 2D (Fig 1c) illustrates a rough hierarchical progression from EVC ROIs in the top-right (gray), to mid-level ROIs (light colors, middle), to high-level ROIs in the lower-left. Additionally, there is a large-scale separation by stream for high-level ROIs, rather than subject or hemisphere, with lateral high-level ROIs (blue) separated and more superior from a tight ventral cluster (magenta), which is in turn, largely distinct from the parietal ROIs (green, though these show greater between subject variability). To further test whether each stream showed a distinct representational structure, we parcellated cortex into 1000 equally spaced ROIs and then calculated the correlation between each pair of parcels. Each comparison was grouped based on whether both parcels were located within the same stream or whether they were located in two different streams, revealing significantly higher correlations within than across streams for this three-stream organization (main effect of within vs. across: p=4.19x10^{-7}; Fig 1d; à la Haak and Beckmann (2018)). The difference in parcel correlations within vs. across streams did not simply reflect anatomical proximity, as the neighboring lateral and parietal streams showed the greatest differentiation.

Computational modeling of visual streams. Next, we examined how well DCNNs predict voxel responses in each subject. As cann be seen for an example subject and network (Fig 2a), the best fitting layer for each voxel is consistent with the hierarchical organization of visual cortex, with early DCNN layers providing the best fit for voxels within EVC and later layers best fitting downstream voxels. Additionally, predictivity was similar across models (mean corrected R^2s between 0.32 and 0.69), reaching the subject-to-subject NC in EVC, which illustrates the power of DCNNs as predictive models of visual activity[2]. Overall, deep ResNets (ResNet-50 and ResNet-101) outperformed other candidate models, though this difference was small. Surprisingly, DCNNs trained on object categorization were equally good predictors of lateral ROIs as they were of ventral ROIs (Fig 2d), despite differences in cortical responses across these ROIs (Fig 1d). However, these models were significantly worse at predicting parietal responses (mean corrected R^2 across subjects and object categorization DCNNs for ventral: 0.56±0.04 SD, for parietal: 0.45±0.05; p=1.39x10^{-5}). Further, the action recognition trained network was a worse predictor of all ROIs than a comparable object categorization DCNN with the same architectural backbone (ResNet-50), with no differences in this deficit between SlowFast and ResNet-50 across streams. Both findings ran contrary to our predictions that DCNNs trained on object categorization would predict ventral responses better than lateral or parietal responses, and that lateral responses would instead be better predicted by an action recognition network (as suggested by Güçlü and van Gerven (2017)).

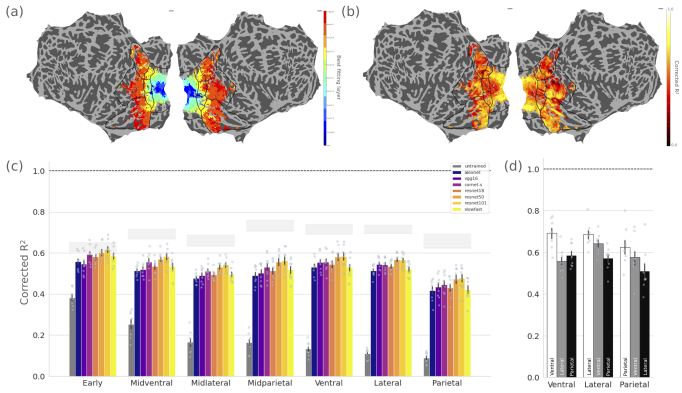

Figure 2: Best fitting DCNN layer (a) and noise-corrected R^2 (b) for each voxel for example subject 1, ResNet-18. (c) Comparison of model fits across candidate models & ROIs. R^2 values are normalized by the NC of each voxel such that 1.0 corresponds to the intrinsic data NC. Each dot represents a subject. Shaded gray error bars: range of subject-to-subject NC. (d) Comparison of ROIs as models of each other. White: within-ROI (same as shaded gray bar in (c)); Gray and Black bars: ROI X's prediction of ROI Y's responses.

Conclusions

Given our findings that representations in visual cortex differ across the three streams, we were surprised that DCNNs trained on object categorization are not only good at predicting responses in the ventral stream but also in the other streams, particularly lateral. These results suggest that additional constraints are needed - either on DCNNs as models of the brain or on the model-to-brain mapping procedure - for models to predict the across-stream differentiation that exists in cortex.

Acknowledgments

This work was supported by a Stanford Graduate Fellowship, MBCT, and NIH grants RO1EY02231801 & RO1EY02391501.

[2]This performance is likely a combination of architecture and training; the untrained AlexNet performed significantly worse in all ROIs, even as its corrected R^2 was close to 0.40 in EVC, highlighting the importance of architectural priors like convolution in these early processing stages.

Collection of the NSD dataset was supported by NSF grants IIS-1822683 & IIS-1822929.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6202–6211).

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20–25.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Güçlü, U., & van Gerven, M. A. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*, 329–336.

Haak, K. V., & Beckmann, C. F. (2018). Objective analysis of the topological organization of the human cortical visual connectome suggests three visual pathways. *Cortex*, *98*, 73–83.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.

Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, *34*.

Pitcher, D., & Ungerleider, L. G. (2020). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. analysis of visual behavior. *Ingle DJ, Goodale MA, Mansfield RJW*.

Weiner, K. S., & Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. *Psychological research*, *77*(1), 74–97.

Wurm, M., & Caramazza, A. (2021). Action and object representation in the ventral" what" stream.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.