Inferotemporal Cortex Underlies Primate Generalization Capabilities and Brain-Aligned Models Generalize Better

Marliawaty I Gusti Bagus
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL, Lausanne, 1015 Switzerland

Ernesto Bocini
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL, Lausanne, 1015 Switzerland

Tiago Marques
McGovern Institute for Brain Research
MIT, Cambridge, MA 02139 USA

Sachi Sanghavi
McGovern Institute for Brain Research
MIT, Cambridge, MA 02139 USA

James J. DiCarlo†
McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Quest for Intelligence

Martin Schrimpf†
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL, Lausanne, 1015 Switzerland

† Joint senior authors.

Abstract

Primate inferotemporal cortex (IT) has been linked to the remarkable human ability of visual object recognition. The linear linkage hypothesis posits that a linear readout of IT neural activity predicts human behavior in core object recognition tasks within the domain of naturalistic images. We here ask whether this hypothesis explains human ability to generalize across image distributions. Specifically, we test if the representations encoded in primate IT combined with a fixed linear readout are sufficient to recognize objects across a variety of image styles such as cartoons, paintings, and sketches. We find that a linear decoder trained on primate IT responses to one image style is – without any additional fitting – able to classify IT responses to other image styles. The predicted performance of such a decoder, with a plausible number of neural sites and naturalistic stimulus training, corresponds to human accuracies across test domains. In artificial neural network models, we find that the more similar models are to primate IT, the better they generalize. When explicitly training models for IT alignment, generalization accuracy increases in correspondence with increased IT alignment. Our findings support that representations encoded in primate IT enable generalization to novel image distributions with a fixed linear decoder.

Keywords: Human Vision; Object Recognition; Out-of-Distribution; Generalization; Primate Visual Ventral Stream; Inferotemporal Cortex; Deep Neural Networks; Brain Alignment.

Introduction

Humans are able to effortlessly recognize objects across a wide range of image distributions, including e.g. sketches, cartoons, and paintings (Kubilius, Kar, Schmidt, & DiCarlo, 2018; Geirhos et al., 2018, 2021). What underlies this remarkable human ability to recognize objects in a wide variety of image styles remains unclear. Previous research has provided evidence for a linear linkage hypothesis – positing that a linear readout from primate IT produces core object recognition behaviors consistent with human behavior (Majaj et al., 2015; Hong et al., 2016). However, these studies focused on only a single domain without transfer to other stimulus distributions.

We here aim to distinguish between two key possibilities for the neural mechanisms underlying human ability to recognize objects in diverse image distributions: whether generalization is primarily achieved in the encoder or the decoder. More specifically, we ask whether the representations encoded by the primate visual ventral stream in IT support linear decoding of unseen image styles. A negative answer would suggest that generalization is achieved downstream of IT with a more sophisticated decoder. A positive answer on the other hand would support the linear linkage hypothesis based on representations from an encoder invariant to the image distribution.

Results

We assembled a diverse image dataset composed of thirteen distinct image styles. This set includes synthetic naturalistic images (HvM) (Majaj et al., 2015), from which we derived four additional styles: silhouette, convex-hull, skeleton and outline (Kubilius et al., 2018). We further added images from multiple artistic styles (Kubilius et al., 2018), as well as photographs in both color and grayscale from Microsoft COCO (Lin et al., 2014), and rendered images from ThreeDWorld (TDW) (Gan et al., 2020). Each image style in our dataset contains at least 60 images, with each of the 10 object classes depicted in no fewer than six images per style. This extensive compilation facilitates a robust assessment of visual object recognition across a broad spectrum of image conditions (see Figure 1 for examples).

Primate IT representations generalize across image styles. We obtained primate IT recordings from Utah microelectrode arrays implanted in two macaque monkeys. To test whether primate IT representations generalize across image styles, we trained a linear ridge decoder on HvM images and tested – without additional training – decoder performance on held-out images from HvM (in-domain performance; Figure 1B) as well as the twelve other image styles (out-of-domain "OOD" performance; Figure 1C). We simulated two-alternative-forced-choice (2AFC) classification by selecting the class with the highest probability from each target-distractor pair.

Given recording limitations, we sought to simulate a more complete primate IT neural population and its access to larger numbers of in-domain training images that a decoder could reasonably be trained on during e.g. development. Across numerous analyses, a multiplicative logarithmic function emerged as the most suitable choice for modeling how the decoder's performance might evolve with increased data richness. We estimate that with ~ 800 neural sites and $\sim 1,000$ training images, the linear decoder achieves the same performance on OOD image styles as humans (Figure 1C). These numbers are in line with previous studies analyzing the linear linkage hypothesis for in-domain classification (Majaj et al., 2015; Hong et al., 2016, Figure 1B).

Models more similar to IT generalize better. Our results so far suggest that representations encoded in high-level visual area IT are sufficient to generalize to new image styles with a fixed linear decoder. Turning to computational models, we investigated whether this finding would transfer to representations in contemporary artificial neural networks (ANNs).

Specifically, we tested if models with representations that are more similar to primate IT also exhibit improved performance on generalization tasks. We estimate the similarity between model and primate IT representations with a linear predictivity metric (Yamins et al., 2014; Schrimpf et al., 2018, partial-least-squares regression), and model generalization performance with the same fixed linear decoder setup

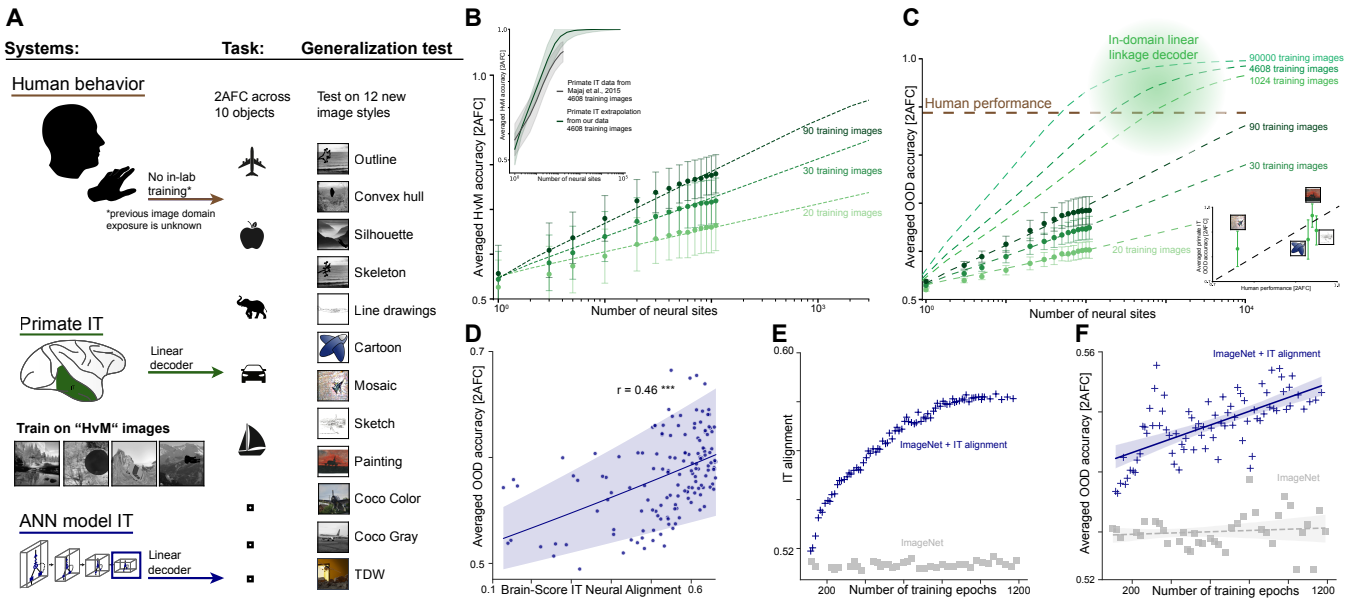

Figure 1: **Primate IT representations are sufficient for generalization across image styles.** **A** Graphical overview: We trained a linear decoder on primate (green) and model (blue) IT representations from one image style (HvM) and tested, without additional training, classification performance on 12 held-out image styles. Human behavioral performance (brown) is a key reference point. **B** Primate IT decoder in-domain performance, varying the number of neural sites and training images. Dots indicate the raw primate IT data; dotted lines are extrapolations along increasing numbers of neural sites (x-axis) and numbers of training images (shades of green). The inset shows the validation of our extrapolations (green line) on a separately recorded HvM dataset (Majaj et al., 2015, real data, gray line). **C** Primate IT decoder out-of-domain (OOD) performance, relative to human performance (brown line). Extrapolations are done as in B. The green circle indicates previously proposed linear linkage decoders (Majaj et al., 2015; Hong et al., 2016). In these primarily results, we excluded image styles with missing behavioral measurements or where the extrapolation over-estimated performance in the interpolation regime. Inset shows performance of a sample decoder vs. human performance on four domains. **D** Model IT alignment and OOD performance are significantly correlated. Dots indicate different models. Exponential curve fit: $y = 0.51e^{0.26x}$ (blue line with 95% confidence intervals, $R^2 = 0.22$). **E** Training a model for IT alignment. Over several epochs, model alignment increases when explicitly training for it (blue), but not otherwise (gray). **F** Generalization performance of the IT-aligned model. OOD accuracy increases across training epochs for the IT-aligned model (blue, $p \ll 0.01$), but not otherwise (gray; $p \gg 0.05$).

as before (Figure 1A). In other words, we fit a ridge classifier to in-domain HvM representations in the model layer associated with IT, and evaluated its OOD performance on the 12 held-out image styles. Across 155 models from Brain-Score (Schrimpf et al., 2018, 2020), we observe a significant correlation between models' IT alignment and their OOD accuracy ($r = 0.46, p \ll 0.01$; Figure 1D).

Models trained with IT alignment generalize better. Beyond correlational analyses, we tested if explicitly training models for improved IT alignment would also improve their generalization performance. We fine-tuned a pre-trained CORnet-S model (Kubilius et al., 2019) by jointly minimizing a classification loss on ImageNet and a representational similarity loss (CKA (Kornblith, Norouzi, Lee, & Hinton, 2019)). This loss specifically targets discrepancies between representations in the model's "IT" layer and those observed in primate IT, and thus promotes a closer alignment with primate neural characteristics (Dapello et al., 2022). Training successfully improved model IT alignment (Figure 1E) as well as OOD generalization performance (Figure 1F, correlation between IT alignment and OOD accuracy $r = 0.67, p \ll 0.01$), while training the model with the ImageNet loss alone does not affect IT alignment or OOD accuracy.

Conclusion

Our results establish the IT linear linkage hypothesis for primate generalization ability across image distributions. We find that representations in primate IT combined with a fixed linear decoder trained on only one domain are sufficient for the recognition of objects in other image styles. These findings transfer to computational models: in a correlational analysis with over 150 models as well as direct model optimization, we observe that models generalize better the more IT-like they are.

Acknowledgments

This work was supported by the Netzwerk Engagement e.V. (M.IGB.), the Lothar and Sigrid Rohde Foundation (M.IGB.), the PhRMA Foundation Postdoctoral Fellowship in Informatics (T.M.), the Semiconductor Research Corporation (SRC) and DARPA (J.J.D., M.S.), Office of Naval Research grant MURI-114407 (J.J.D.), the Simons Foundation grant SCGB-542965 (J.J.D.), the MIT-IBM Watson AI Lab grant W1771646 (J.J.D.), the Takeda Fellowship in AI and Health (M.S.), and the Friends of the McGovern Fellowship (M.S.).

References

Dapello, J., Kar, K., Schrimpf, M., Geary, R., Ferguson, M., Cox, D. D., & DiCarlo, J. J. (2022). Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness. *bioRxiv preprint*. doi: 10.1101/2022.07.01.498495

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., ... Yamins, D. L. K. (2020). ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation. In *Neural information processing systems (neurips)*.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Neural Information Processing Systems (NeurIPS)*.

Geirhos, R., Schütt, H. H., Medina Temme, C. R., Bethge, M., Rauber, J., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Neural information processing systems (neurips)* (pp. 7538–7550).

Hong, H., Yamins, D. L., Majaj, N. J., & Dicarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622. doi: 10.1038/nn.4247

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. E. (2019). Similarity of neural network representations revisited. *CoRR*, *abs/1905.00414*.

Kubilius, J., Kar, K., Schmidt, K., & DiCarlo, J. (2018). Can Deep Neural Networks Rival Human Ability to Generalize in Core Object Recognition? In *Cognitive computational neuroscience (ccn).* doi: 10.32470/ccn.2018.1234-0

Kubilius, J., Schrimpf, M., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In *Neural information processing systems (neurips)* (pp. 12785—-12796).

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2014). Microsoft COCO: Common Objects in Context. *arXiv preprint*.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, *35*(39), 13402–13418. doi: 10.1523/JNEUROSCI.5181-14.2015

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*. doi: 10.1101/407007

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*. doi: 10.1016/j.neuron.2020.07.040

Yamins, D. L. K., 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 (PNAS)*, *111*(23), 8619–8624. doi: 10.1073/pnas.1403112111