Ventral Stream Responses to Inanimate Objects are Equally Aligned with AlexNet (2012) and Modern Deep Neural Networks

Leilany Torres Diaz (leilany_torresdiaz@g.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA.

George A. Alvarez (alvarez@wjh.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA.

Abstract:

The field of deep learning is continuously developing novel neural network architectures, from residual connections in CNNs (He et al., 2016) to vision transformers with self-attention mechanisms (Dosovitskiy et al., 2020). While these advances appear to increase models' visual capabilities, it remains unclear whether modern architectures are becoming more brain-aligned. To address this question, we examined an Inanimate Objects neuro-imaging dataset with reliable neural responses to 72 inanimate objects in early visual cortex (EarlyV), posterior occipito-temporal cortex (pOTC), and anterior occipito-temporal cortex (aOTC). We compared alignment between model feature spaces and voxel-space using classic, unweighted representational similarity analysis. We included top-performing models on the Brain-Score platform (Schrimpf et al., 2018) and a leading visual foundation model (DinoV2, Oquab, 2023). We found that an AlexNet baseline model (Krizhevsky et al., 2012) matches or exceeds these models in alignment with each brain region, with substantial reliable variance in aOTC remaining unexplained by any model. These results suggest that progress in deep neural network development is missing key aspects of high-level visual representation in the human brain.

Keywords: model-brain alignment; representational similarity analysis; deep neural network models; AlexNet; vision-transformers; fMRI

Introduction

To what degree have modern deep neural networks converged on more (or less) brain-like representations than the seminal AlexNet (2012) model? While there have been many large-scale benchmarking efforts that broadly address this question (Conwell et al., 2024; Schrimpf et al., 2018), here we focus on a smaller, targeted neural dataset with neural responses to 72 Inanimate Objects (Konkle & Alvarez, 2022). This dataset (Fig. 1A) is interesting for a number of reasons:

(1) The data were collected with an optimized fMRI protocol, resulting in individual subject representational geometries that are reliable and consistent across subjects, yielding highly reliable group-averaged representational geometries that provide a robust target to predict with different deep neural networks. (2) Because the animate/inanimate divide is essentially principal component #1 in neural response variation (Konkle & Caramazza, 2013), capturing the structure within inanimate-objects only requires models to capture more fine-grained or local structure. (3) Early work showed that there was substantial variance in this neural dataset that remained unexplained by convolutional neural network models, particularly in the anterior ventral stream (Konkle & Alvarez, 2022).

Thus, our goal in the present work is to assess whether there has been any progress in model-brain alignment for this Inanimate Objects Dataset, focusing specifically on emergent alignment between model feature spaces and voxel-spaces in three different neural sectors (early visual cortex, EarlyV; posterior occipito-temporal cortex, pOTC; anterior occipito-temporal cortex, aOTC). To compare model and brain representations, we used classic, unweighted representational similarity analysis (Kriegeskorte, 2008), computing representational dissimilarity matrices (RDMs) directly in the model-activation space, and voxel-space, then comparing matrices using Pearson correlation (among scores in the upper-triangular portion of the RDMs). We focus on unweighted RSA, rather than a procedure that allows linear re-weighting of features (e.g., voxel-wise encoding RSA; Konkle & Alvarez, 2022) to avoid complexities in interpretation that arise with more flexible neural-linking procedures (Prince et al., 2024; Lindsay, 2021).

Methods

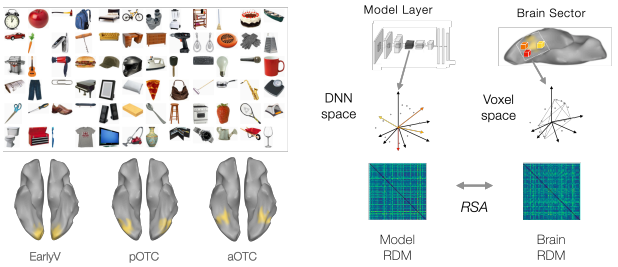

Figure 1. A) The 72 Inanimate objects, and view of EarlyV, pOTC, and aOTC ROIs on the ventral surface of an inflated cortex. B) The representational similarity analysis procedure.

Neural Data: The Inanimate Objects fMRI Dataset (**Fig. 1A**) contains fMRI responses from ten participants viewing 72 inanimate objects presented on white backgrounds. Responses were divided into three sectors: EarlyV, pOTC, and aOTC (see Konkle and Alvarez, 2022 for details).

Model selection. AlexNet (2012) was used as our baseline model. To constrain the selection of target models, we selected models based on current (April, 2025) rankings on the brain-score platform (Schrimpf et al., 2018), including the top 5 overall neural models, and the top 5 models on sub-scores for V1, V4, and IT alignment. We also included a recent visual foundation model (DinoV2; Oquab, 2023) to assess progress for models that achieve current state-of-the-art on computer-vision benchmarks unrelated to brain alignment.

Unweighted Representational Similarity Analysis. We computed model RDMs by passing the 72 objects through each model and calculating dissimilarity (1-

pearsonr) between activations for all item pairs. Neural RDMs were computed similarly using voxel responses. Model-brain alignment was quantified as the Pearson correlation between their respective RDMs.

Cross-validated best-layer identification. To identify each model's "maximally-aligned layer" with neural responses, we used a cross-validation procedure (Konkle & Alvarez, 2022). Neural data were split into halves to compute mean RDMs separately for each group. Pearson correlation was calculated between each model layer RDM and the Group1 neural RDM, selecting the layer with highest correlation. This layer's correlation with the independent Group2 RDM was then taken as the model's maximum correlation with the given brain region. We repeated this procedure across all possible subject split-halves, taking the average as our cross-validated max-r measure.

Results

Across the board, AlexNet (2012) shows equivalent or greater alignment with visual responses to inanimate objects compared to all models in our analysis. Figure 2 shows results by brain region, with AlexNet in light grey (left). Other models are grouped by brain-score ranking (blue: top-5 overall neural score; orange: top-5 IT score; green: top-5 V4 score; purple: top-5 V1 score). DinoV2 appears in dark grey (right). No tested model shows stronger alignment than AlexNet in any brain region. In most cases, AlexNet demonstrates significantly greater brain alignment (paired t-tests over

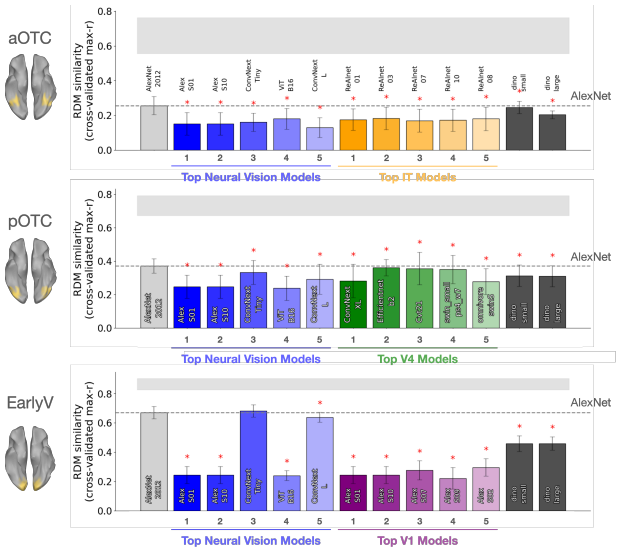

Figure 2. Cross-validated best-layer correlation between model RDMs and neural RDMs for each brain region. In each brain region, AlexNet (light grey, left) shows equivalent or greater alignment than all tested models including top brain-score models (colored bars), and DinoV2 (dark grey, right). The shaded gray regions at the top represent the brain-vs-brain noise ceiling (+/- 95% CI). Error bars show 95% CI over split-halves.

all split-halves with a correction for non-independence, following Bouckaert & Frank, 2004).

Discussion

We found that the baseline AlexNet model shows equal or significantly stronger alignment with ventral stream responses to a set of 72 inanimate objects than leading brain-score models or recent vision foundation models. Thus, models with advanced features like residual connections (ConvNeXt, Liu et al., 2022), multi-head attention (ViT), or large-scale training (DinoV2) showed equal or worse brain-alignment than AlexNet.

These results are somewhat surprising because this set of models includes the most brain-aligned models on the brain-score benchmark, including the leading models for IT cortex predictivity (Schrimpf et al., 2020). One possible explanation for this finding is that the current work used classical RSA without feature re-weighting, whereas many of the brain-score scores involve a linear re-weighting of features. Here we focused on unweighted RSA because it avoids some of the complexities in interpretation that arise when re-weighting is allowed. Nevertheless, an important avenue for future work is to examine whether AlexNet remains the most brain-aligned model for the Inanimate Objects dataset when different model-to-brain linking procedures are used, such as voxel-wise encoding RSA (veRSA; Konkle & Alvarez, 2022), or sparse-positive encoding RSA (spRSA; Prince et al, 2024).

These results are also somewhat inconsistent with prior large-scale benchmarking showing that AlexNet does not have stronger alignment with OTC responses to the Natural Scenes Dataset (Conwell et al., 2024). Unlike NSD (Allen et al., 2022), which spans animate and inanimate categories, our dataset focuses on exclusively inanimate objects, suggesting that AlexNet may better capture fine-grained structure within the inanimate domain compared to state-of-the-art architectures optimized for ImageNet performance.

Our broader goal was to examine whether enhancements in deep neural networks lead to more brain-aligned object responses, focusing on the Inanimate Objects Dataset. Our results indicate that deep neural network advances since AlexNet have not improved emergent brain-alignment for this specific dataset, particularly in anterior ventral stream regions (aOTC). We propose that specifically biologically-inspired model variations may be required to go beyond the early successes of task-optimized convolutional neural networks (Yamins, et al. 2014) in capturing high-level object representation in the human brain. More generally, these results suggest that smaller datasets that target restricted domains may serve as valuable targets for NeuroAI research focused on model-brain alignment.

Acknowledgements

This work was supported by the Kempner Institute for Biological and Artificial Intelligence at Harvard University (supporting L.T.D.) and an NSF PAC COMP-COG grant (1946308; to G.A.A.)

References

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

Bouckaert, R. R. & Frank, E. Evaluating the replicability of significance tests for comparing learning algorithms. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. 3–12 (Springer, 2004).

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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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).

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, *13*(1), 491.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, *33*(25), 10235-10242.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology, 55*, 167–179.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017-2031.

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976-11986).

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Prince, J. S., Conwell, C., Alvarez, G. A., & Konkle, T. (2024, March). A case for sparse positive alignment of neural systems. In *ICLR 2024 Workshop on Representational Alignment*.

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.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413-423.

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.