Parametric control along the encoding axes of IT neurons uncovers hidden differences in model-brain alignment

Jacob S. Prince (jprince@g.harvard.edu)
Dept. of Psychology, Harvard University, Cambridge, USA

Binxu Wang (binxu_wang@g.harvard.edu)
Kempner Institute, Harvard University, Cambridge, USA

Akshay V. Jagadeesh (Akshay_Jagadeesh@hms.harvard.edu)
Harvard Medical School, Boston, USA

Thomas Fel (tfel@g.harvard.edu)
Kempner Institute, Harvard University, Cambridge, USA

Emily Lo (emily282lo@gmail.com)
Dept. of Psychology, Harvard University, Cambridge, USA

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

Margaret S. Livingstone (margaret_livingstone@hms.harvard.edu)
Harvard Medical School, Boston, USA

Talia Konkle (talia_konkle@harvard.edu)
Dept. of Psychology, Harvard University, Cambridge, USA

Abstract:

As model-brain alignment scores increasingly saturate under current assessment methods, new approaches are needed to test whether there are actually hidden differences in how well models capture biological feature tuning. To this end, we introduce a paradigm for comparing deep encoding models based on their ability to *control* neural responses along their hypothesized encoding axes. Using recordings from macaque inferotemporal cortex, we compared two DNN-based encoding models: a standard ResNet-50 and an adversarially robust variant. These models achieved comparable performance in predicting neural responses over a wide range of natural images. However, we found they differed substantially when subjected to a test of "parametric control." Leveraging an explainable AI technique called feature accentuation, we synthesized image sets that varied systematically in precise intervals along each encoding axis, based on the hierarchical computations of each model. We found that accentuated stimuli from the robust model achieved superior control of neural firing. We then synthesized "controversial" stimuli that further validated the brain alignment of RN50-robust over the baseline model. Our framework offers a new means to arbitrate between models, requiring a more precise characterization of feature tuning in targeted local regions of image space.

Introduction

Popular neural encoding benchmarks such as BrainScore have increasingly saturated, with recent surveys showing that hundreds of models are capable of scoring within Pearson r = 0.1 of each other in prediction of responses in high-level visual cortex, despite major differences in architectures and training objectives (Schrimpf et al. 2018; Conwell et al., 2024). Are these models truly learning equivalent representations, or might they rely on different feature tuning and different underlying computational mechanisms, some of which are more brain-aligned than others? Recent evidence suggests that models can achieve high predictivity despite containing misaligned features (Prince et al., 2024), due to the inherent flexibility of popular encoding procedures such as ridge regression. More rigorous procedures for evaluating these models are needed.

Here we introduce **parametric neural control** as a more stringent test of brain alignment. In this paradigm, given a linear encoding model fit from a DNN layer to a particular recording site (an "encoding axis"), we use **feature accentuation** (Hamblin, Fel, et al., 2024) to systematically manipulate a set of natural images along that axis. This process creates "accentuated" image sweeps, expected to systematically modulate neural responses in incremental steps either above or below the response of the original seed image. If the encoded features truly align with the brain's feature tuning, neural

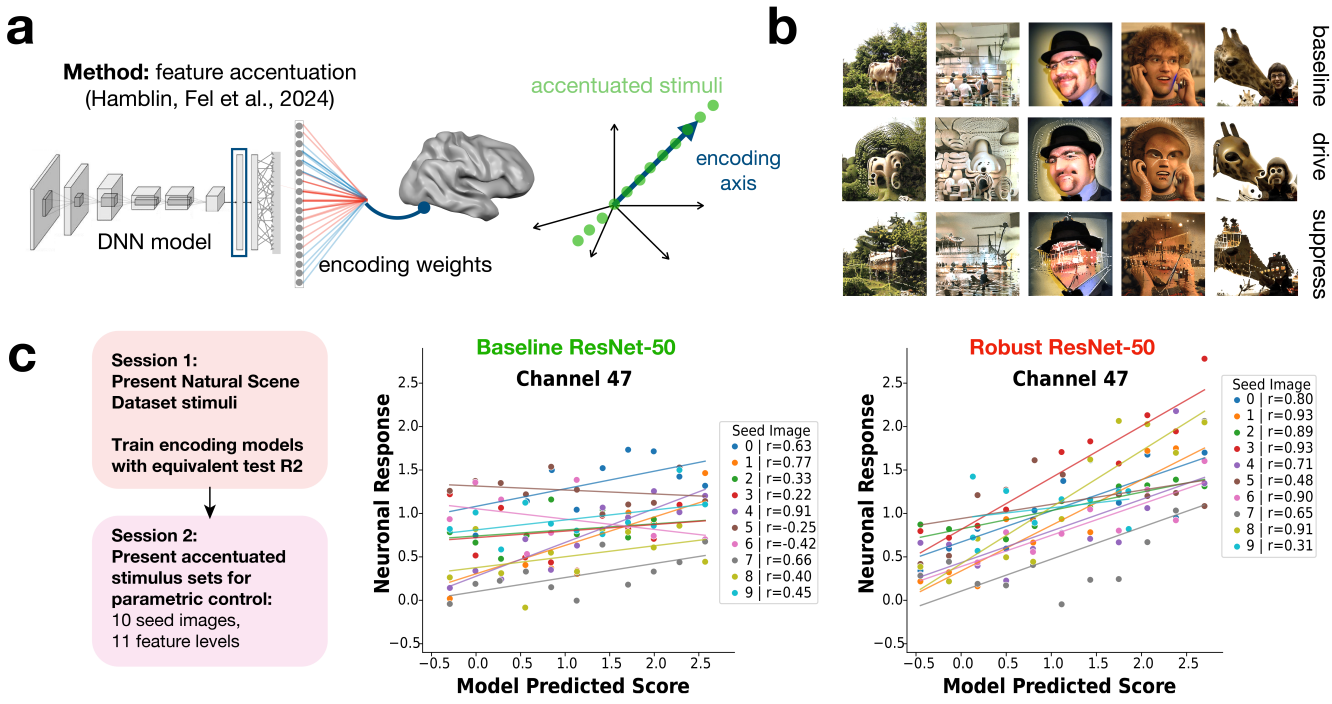

Figure 1: (a) Schematic of the feature accentuation pipeline, which uses a DNN's encoding axes to generate stimuli predicted to control neural responses at even, fine-grained intervals. (b) Example accentuated images from baseline vs. robust models. (c) We observe stronger parametric control arising from the robust ResNet-50 (right) than from the baseline ResNet-50 (left), as indicated by higher correlations between predicted and recorded responses.

responses should show a clear, graded pattern that tracks these manipulations. This closed-loop method directly tests whether high encoding performance on natural images actually translates into meaningful brain alignment, through a stringent generalization test that forces the model to generalize to stimuli that arise from targeted local perturbations along the image manifold.

Results

We recorded responses to images from the Natural Scenes Dataset (NSD; Allen et al., 2022) in face patches in IT cortex of two macaques using floating microelectrode arrays, and conducted a focused case study of encoding models derived from a standard ResNet-50 (RN50) and an adversarially robust variant (RN50-robust). For each neural site, we first fit encoding models using responses to 800 natural images from the COCO dataset. Many neural sites achieved nearly identical encoding R^2 on held-out NSD images. We identified the five most reliable sites for the next analyses, exploring whether the seemingly equivalent encoding models from the RN50 and RN50-robust backbones were actually equal when required to control neural responses along their encoding axes, using stimuli that span targeted local regions of image space.

To test this, we selected 10 seed COCO images, and generated a sweep of 11 accentuated stimuli per image predicted to either enhance, maintain, or suppress the neural sites' responses. Critically, the feature accentuation method accepts a seed image as input, and creates local image perturbations related to the gradients of the model along the encoding axis—in this way, it implements a stronger test of how the model hierarchically computes that particular tuning axis (**Fig 1a**). Image sweeps were created along both the RN50 and RN50-robust tuning axes, and presented to the monkey the subsequent day. We found that RN50-robust-derived stimuli elicited reliable, graded modulation of firing (mean r over seed images = 0.855 +/- 0.055 SD over k = 5 channels), whereas RN50-derived stimuli showed weaker control (mean r = 0.470 +/- 0.116; **Fig 1c)**. Qualitative inspection suggests that RN50-robust encodings emphasized cohesive face contours, while RN50 relied more on local textural features such as hair and fur (**Fig 1b**).

To further reveal these differences, we synthesized "controversial stimuli" (Golan et al., 2020), which one model predicted would elicit high firing while the other predicted suppression. In empirical testing, only RN50-robust's predictions positively correlated with recorded responses, providing further validation of the features that are unique to that model. Together, these results hint that adversarial training may encourage more IT-aligned high-level feature tuning by pressuring toward more robust, globally coherent visual feature representations (see also Feather et al., 2023).

To assess the generality of these findings, we replicated both of these experiments in a second monkey and found consistent results. The same dissociation between global-predictivity and local-control emerged: while both models achieved similar encoding R^2 scores over a diverse set of natural images, RN50-robust-derived stimuli exhibited superior neural control within local regions of image space compared to RN50-derived stimuli (main control experiment: mean r = 0.711 +/- 0.123 over 5 channels for RN50-robust; mean r = 0.308 +/- 0.189 for baseline RN50). These results reveal that while models may have a similar ability to predict neural responses when assessed over a wide span of natural images, they can be clearly dissociated in their commitments to the feature tuning using accentuated stimulus sets.

Discussion

This parametric control paradigm provides a stringent test of model-brain alignment by directly testing the generalization of each model using stimulus sets that span targeted local regions of image space, in the precise directions that are predicted to exert the greatest influence over neural firing. We find that models with similar ability to predict neural responses over a wide span of natural images can be clearly dissociated using model-derived stimuli that make these feature preferences explicit.

Our approach shares some similarities with prior encoding-based neural control studies (Bashivan et al., 2019; Tuckute et al. 2024). Unlike approaches using diffusion models or generative priors (Ponce et al. 2019; Luo et al. 2024; Cerdas et al., 2024), our method relies only on the encoding model itself, performing sweeps of targeted perturbations along specific feature dimensions, without introducing unrelated priors or biases. The key innovation here is that these image perturbations leverage the full computational graph of the encoding model, testing not just what features matter but how they are computed hierarchically. Importantly, even in research scenarios where closed-loop experiments are not possible, feature accentuation can provide a useful new way to visualize and compare the features that different encoding models exploit.

Ongoing work is extending these experiments to larger model groups to understand which design principles are most important for achieving precise neural control. Further, manipulating aspects of the mapping scheme—such as the feature basis (e.g. PCA, sparse random projections, sparse autoencoders) and regularization of encoding weights (e.g. Ridge, Lasso, sparse-positive)—will help reveal how these technical choices influence control scores.

Overall, these findings emphasize the importance of looking beyond traditional encoding metrics to evaluate mechanistic alignment between DNNs and brains (see also Feather et al., 2025), highlighting precise feature tuning alignment as a crucial dimension of model evaluation.

Acknowledgements

This research was supported by NSF CAREER BCS-1942438 (TK), and an NDSEG Fellowship (JSP).

References

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

Cerdas, D. G., Sartzetaki, C., Petersen, M., Roig, G., Mettes, P., & Groen, I. (2024). BrainACTIV: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation. *bioRxiv*, 2024-10.

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.

Feather, J., Khosla, M., Murty, N., & Nayebi, A. (2025). Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*.

Feather, J., Leclerc, G., Mądry, A., & McDermott, J. H. (2023). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, *26*(11), 2017-2034.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330-29337.

Gu, Z., Jamison, K. W., Khosla, M., Allen, E. J., Wu, Y., St-Yves, G., . . . Kuceyeski, A. (2022). Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247 , 118812.

Hamblin, C., Fel, T., Saha, S., Konkle, T., & Alvarez, G. (2024). Feature accentuation: Revealing 'what' features respond to in natural images. *arXiv* preprint arXiv:2402.10039.

Luo, A., Henderson, M., Wehbe, L., & Tarr, M. (2024). Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999-1009.

Prince, J. S., Conwell, C., Alvarez, G. A., & Konkle, T. (2024). 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.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., . . . Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behavior,* 1–18.