 DOI: 10.3217/978-3-99161-050-2-233

Parametric control of neurons in IT cortex as a stringent generalization test for deep encoding models

Jacob S. Prince[1]*, Binxu Wang[1,2,3], Akshay V. Jagadeesh[3], Thomas Fel[1,2], Emily Lo[1], George A. Alvarez[1], Margaret S. Livingstone[3], Talia Konkle[1,2]

[1]*Harvard University, Cambridge, MA, USA;* [2]*Kempner Institute for the Study of Natural and Artificial Intelligence, Cambridge, MA, USA;* [3]*Harvard Medical School, Cambridge, MA, USA*

*33 Kirkland St., Cambridge, MA, USA. E-mail: jprince@g.harvard.edu

Introduction: Deep neural network (DNN) models have potential to transform our understanding of how sensory inputs are processed in cortex. While many models increasingly achieve high neural predictivity, a critical question is whether their predictive capacity generalizes robustly beyond the original training domain. We propose that a model's ability to control neural responses—rather than merely predict them—serves as a powerful indicator of such generalization. ***Materials, Methods and Results:*** Using neural recordings in macaque inferotemporal (IT) cortex, we compared two DNN-based encoding models: a standard ResNet-50 and an adversarially robust variant. Both achieved comparable predictive performance (R^2) for natural images, yet differed substantially in their capacity for *parametric control*. We used an explainable AI method called "feature accentuation" to synthesize new images that systematically varied along each model's encoding axes. These accentuated stimuli were then presented to the same animal under identical conditions the next day. We found that stimuli from the robust model achieved precise modulation of neural firing: responses reliably and predictably aligned with each feature level. In contrast, baseline ResNet-derived stimuli showed far weaker parametric control. Qualitative analyses further showed that robust model accentuations emphasized cohesive object-like contours, whereas baseline accentuations altered mostly textural patterns. ***Conclusion:*** Parametric control offers a stronger test of whether an encoding model has genuinely identified the features represented in neural populations. This form of generalization will be essential for encoding-model-based BCI systems that must reliably operate under new conditions and stimuli. By identifying models whose representations support accurate neural modulation, this approach paves the way for more robust, flexible BCIs.

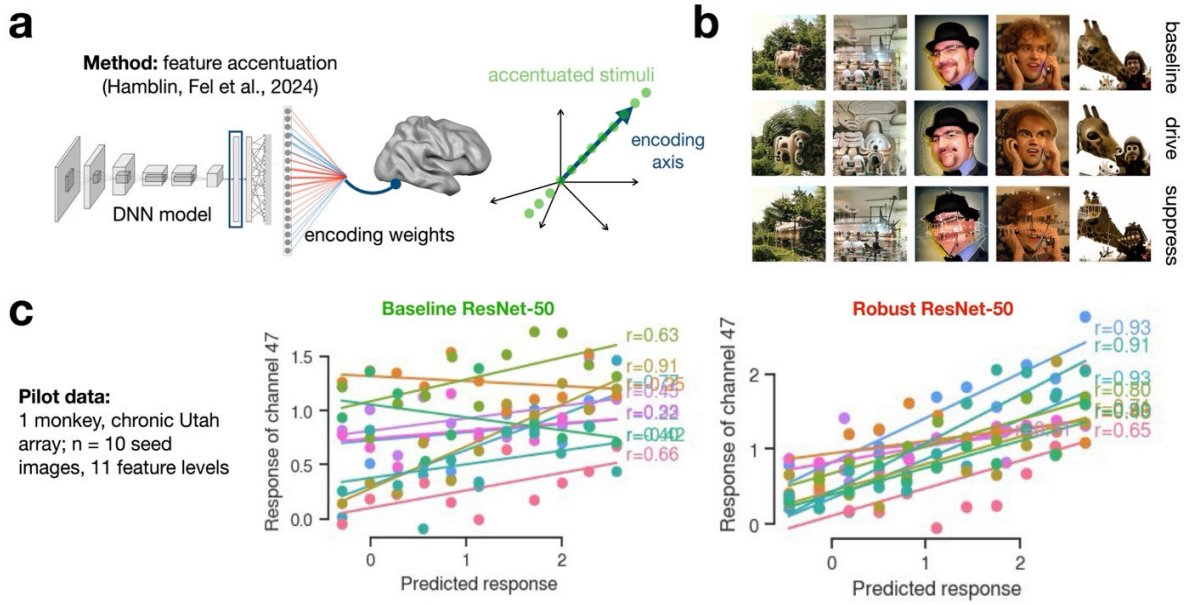

Figure 1: (a) Schematic of the feature accentuation pipeline, which uses a DNN's encoding axes to generate "drive" and "suppress" stimuli at even intervals. (b) Example accentuated images from baseline vs. robust models. (c) Pilot data reveal stronger parametric control by the robust ResNet-50 (right)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

Acknowledgments and Disclosures: This research was supported by NSF CAREER BCS-1942438 (TK), and an NDSEG grant (JSP).

 Published by Verlag der Technischen Universität Graz