Neural Footprints: Variance-Based Methods Systematically Miss High-Level Representations

Jon Gauthier[1] and **Tyler Brooke-Wilson**[2]
[1] University of California, San Francisco [2] Yale University
jon@gauthiers.net, tyler.brooke.wilson@yale.edu

Abstract

The brain computes at vastly different levels of representation, encoding both abstract knowledge and fine-grained sensory information about the world. These two ends of the spectrum have distinct statistical properties: while abstract knowledge about the world is compact and low-dimensional, sensory information is highly variable and difficult to compress. This statistical asymmetry creates fundamental issues for several neuroscience analysis techniques, biasing these methods to favor high-dimensional sensory features over low-dimensional abstract representations. In a simulation experiment, we demonstrate that variance-based techniques fail to distinguish between obviously correct and incorrect models of visual perception due to this statistical asymmetry. Reliable neural analysis therefore requires models that are not merely fit to explain variance in neural activations, but that more deeply link representations to the computations they support.

Introduction

Perception requires the brain to simultaneously maintain representations at multiple levels of abstraction. At one end, neurons encode fine-grained sensory details—textures, color patterns, raw auditory inputs—which vary enormously across stimuli. At the other, the brain must represent compact, abstract knowledge about the world: object identities, physical properties like mass and friction, syntactic structure, and causal dynamics. These two levels of representation have fundamentally different informational footprints: sensory detail is high-dimensional and difficult to compress, while abstract representations of the world are relatively low-dimensional and compact. This creates a problem for standard analysis methods that measure a model's plausibility by its capacity to explain variance in brain data, including encoding models and principal component analysis (Kay et al., 2008; Naselaris et al., 2011; Yamins et al., 2014).

We demonstrate this in a simulation experiment in which synthetic neural activity is generated from a known feature set via linear projection, guaranteeing that all relevant variables are recoverable in principle. We show that encoding models and PCA nonetheless fail to detect the presence of high-level features in the neural signal. We argue that *computational sufficiency*—whether a model's representations support the relevant downstream computations—is a necessary complement to neural goodness of fit.

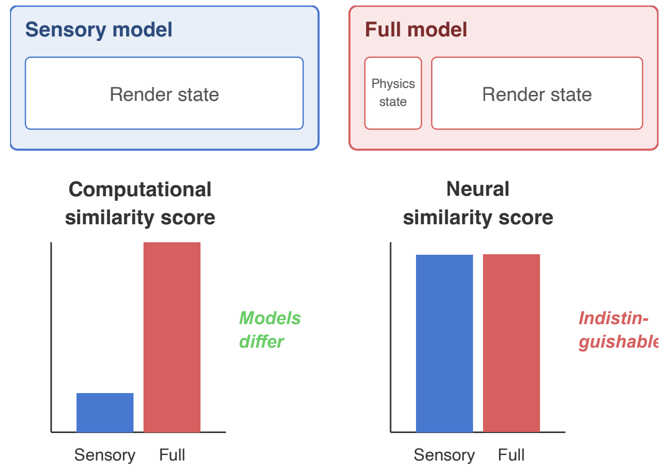

Figure 1: We characterize two models of visual perception (top). A sensory-inputs model represents only low-level visual properties — high-dimensional and difficult to compress. A full model additionally represents physical states such as object identity and motion — low-dimensional and compact. Despite major computational differences, standard neural analyses (brain encoding and PCA) systematically fail to distinguish the two.

Methods

We construct a simple model of visual perception using a rigid-body physics simulation and 3D renderer, in which a single object moves around a possibly occluding wall. We generate synthetic *neural activity* by mapping the full state of the simulation and renderer — including both rendered scene properties and underlying physical variables — to a high-dimensional distributed activation.

We consider two candidate models of the resulting neural activity (Figure 1). A **sensory-only model** proposes that neural activity is driven solely by the surface-level properties of these scenes—a buffer of ∼49,000 floats comprising color, depth, and segmentation maps.

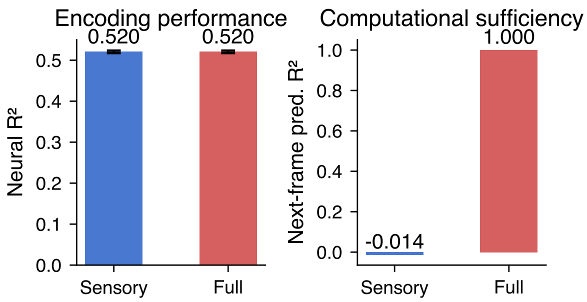

Figure 2: Brain encoding vs. computational sufficiency. The sensory model dominates neural variance but cannot predict future scenes; the full model barely contributes to neural variance, yet perfectly predicts future frames.

A **full model** additionally includes a high-level description of *physics state*: object positions, shapes, colors, velocities, masses, and lighting conditions (33 parameters). These variables are disjoint subsets of the simulation's ground truth state; only the latter variables are sufficient alone to render the scene.

The key asymmetry between these two models is dimensional: the sensory description comprises $\sim$49,000 floats per scene, while the high-level description requires only 33 parameters. Let $X_s \in \mathbb{R}^{N\times 49000}$ and $X_h \in \mathbb{R}^{N\times 33}$ denote the sensory and high-level descriptions across N scenes, and let $X_f = \text{concat}(X_s, X_h)$. Synthetic neural activity is generated as a variance-preserving random linear projection of X_f.

Results

We first establish that the two models are computationally distinct, then show that two popular neural analysis methods fail to distinguish them.

Computational dissociation. By design, the high-level features X_h are sufficient to generate a visual rendering of not only the current frame but of future frames of the physical simulation. Even a high-capacity multilayer perceptron cannot accomplish the same task working from sensory features (Figure 2 right).

Encoding models. We fit linear encoding models predicting neural activation from the sensory model X_s or the full model X_f, and find that the two models are statistically indistinguishable (improvement in explained variance incorporating high-level features is ΔR^2 = 0.000006; under a permutation-based null hypothesis test we find $p = 0.223$; Figure 2 left).

PCA. For each scene, we label the direction of the foregrounded object's motion (left vs. right) and ask whether this can be decoded from the top principal components. Binary decoders of motion direction succeed using the top 10 principal components (accuracy 52.5%, $p = 0.038$; Figure 3). However, further analysis reveals that this decoding success is not clearly driven by high-

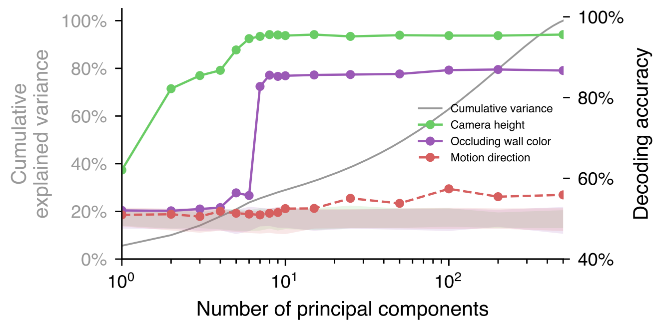

Figure 3: Principal components analysis. The leading principal components of the neural data are dominated by sensory contrasts (height of the camera; color of the occluding wall), while high-level features (direction of motion of the foreground object) are not decodable in the top principal components.

level features. We estimated a regression mapping from sensory features X_s to neural activity, and computed the residuals of this regression (the neural data unexplained by sensory features) on held-out data. The principal components of these residuals do not successfully decode motion direction (49.7%, $p = 0.518$). This indicates that success in motion decoding from this neural data cannot be clearly attributed to neural traces of high-level features.

Discussion

Across two analysis methods common in cognitive neuroscience, we found the same pattern of failure: a model that captures only low-level sensory features is statistically indistinguishable from one that additionally represents the physical variables that generate the scene—even though only the latter can predict what happens next. Variance-based neural analysis methods are structurally insufficient for identifying the kinds of high-level, abstract representations that drive perception and behavior.

This blindspot is not specific to vision. In visual cortex, model agreement on low-level textural features is likely to mask disagreement about physical scene properties (cf. Yamins et al. (2014)). The same logic applies to speech perception, where encoding of acoustic features could obscure differences in how models represent semantic or syntactic structure, and to motor cortex, where high-variance motor commands may swamp the low-dimensional programs that govern motor planning. In each case, the statistically dominant representations are not necessarily the computationally relevant ones.

These results require us to shift our empirical standards: rather than asking only how well a model explains variance in neural data, we should ask whether the representations it posits are sufficient to support the computations the brain actually performs.

Disclosure

Large language models (Claude Opus 4.7 and Sonnet 4.6) were used for the development and evaluation of these analyses. All results were thoroughly validated by the authors prior to writing the manuscript.

References

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. https://doi.org/10.1038/nature06713.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073.

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.