Predicting IT and V4 Neural Responses With Performance-Optimized Neural Networks

Daniel Yamins*, Ha Hong*, Darren Seibert, and James J. DiCarlo (*equal contribution)
McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, MIT, Cambridge MA

Though substantial evidence shows that higher ventral cortex supports the capacity for rapid view-invariant object recognition, processing mechanisms in these areas remain poorly understood. Here, we hypothesized that an effective way to produce models of individual IT and V4 neuronal responses would be to match the higher ventral stream's population-level ``behavior". If IT developed to support invariant object recognition behavior, networks that perform invariant object recognition better tasks should also more effectively capture responses of high-level visual neurons.

To test this we used high-throughput computational techniques to evaluate thousands of candidate architectures within a class of biologically-plausible feedforward neural network models. We measured categorization performance and IT neural explained-variance for each model, finding a strong correlation between a model's performance and IT predictivity. This result has a clear implication: to find highly IT-like models, performance optimization may be an effective strategy. We next identified a highly optimized feedforward network model that matches IT performance on a range of challenging recognition tasks. Critically, even though we did not explicitly constrain this model to match neural data, its output layer turned out to be highly predictive of IT neural responses –- a 90% improvement over comparison models and comparable to the state-of-the-art models of lower-level areas such as V1. Moreover, the penultimate model layer is highly predictive of V4 neural responses, showing that performance optimization imposes biologically consistent constraints on intermediate feature representations as well.

A common assumption in visual neuroscience is that understanding the qualitative structure of tuning curves in lower cortical areas is a necessary precursor to explaining higher visual cortex. Our results indicate that it is useful to complement this ``bottom-up" approach with a ``top-down" perspective in which behavioral metrics are a sharp and computationally tractable constraint shaping individual neural tuning curves in both higher and intermediate cortical areas.

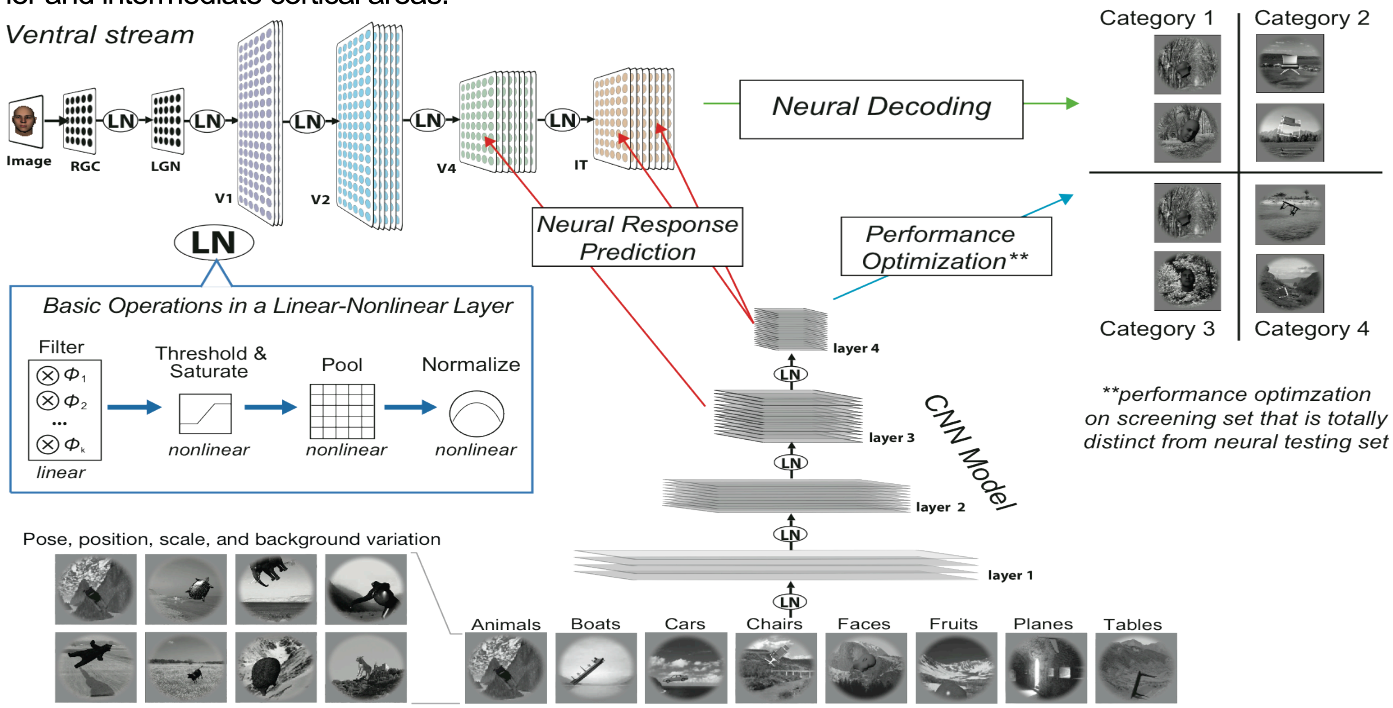


Fig. 1: The ventral visual stream is a group of adjacent cortical areas present in human and primates. The more anterior of these areas, including V4 and IT, play a key role in object recognition. We recorded the responses of 168 IT and 128 V4 multi-unit sites on a set containing 5760 images of a variety of photorealistic 3D objects in eight natural categories, with significant object view variation and complex cluttered backgrounds. We seek models within a large parameter space of biologically plausible Convolutional Neural Network (CNNs), whose intermediate and top layers match V4 and IT neural responses, respectively. Since IT is believed to underlie high-level object recognition (e.g. neural decoders trained on IT response data can match human categorization behavior on challenging categorization tasks), we hypothesized that optimizing CNN network parameters for object recognition performance at the output layer would drive the discovery of neurally-predictive models.

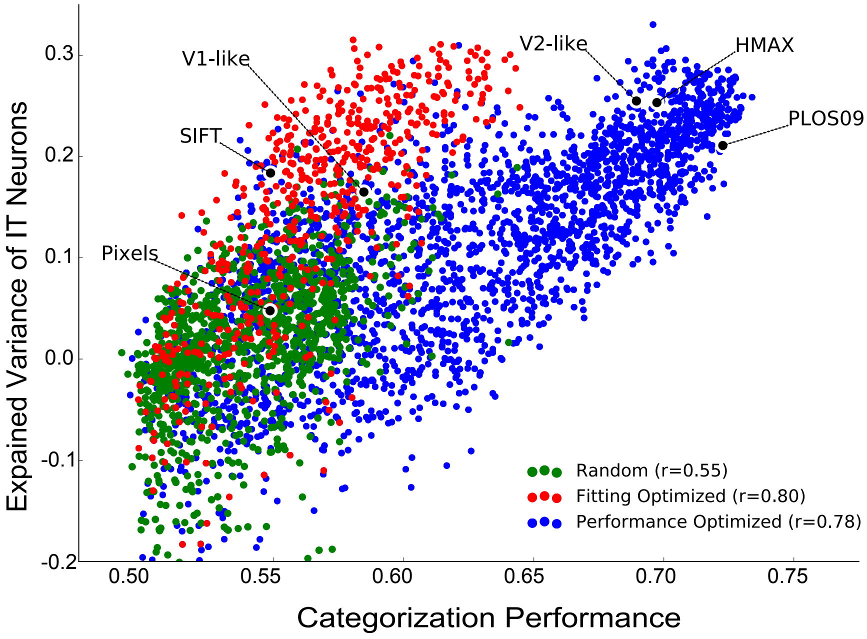
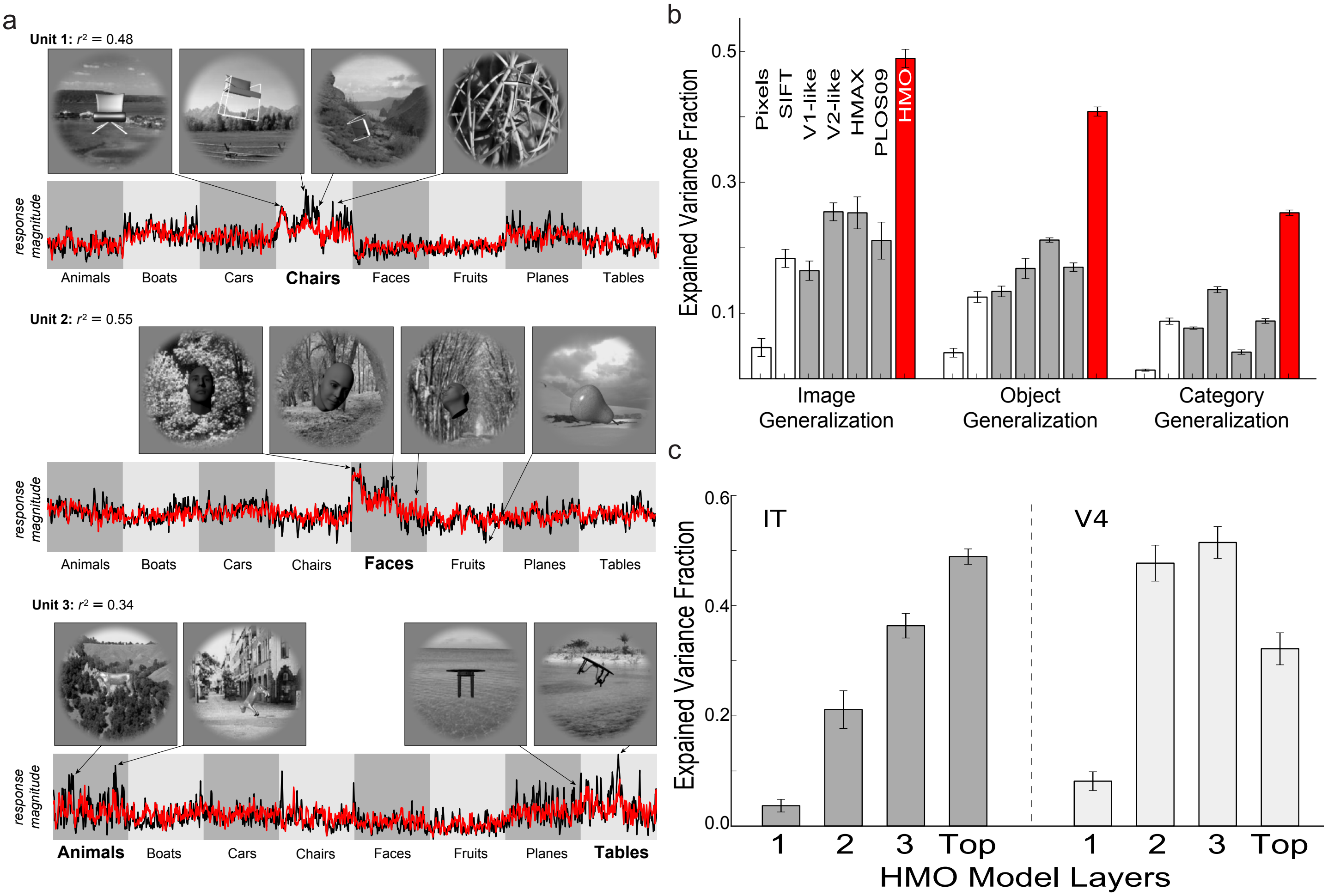


Fig. 2 (right): Object categorization performance vs IT explained variance of model top-level output units. Performance (x-axis) is measured as balanced accuracy for correlation classifiers trained on model output. IT predictivity (y-axis) is the explained variance fraction for individual IT units, averaged over all 168 units. Each point represents a distinct model in a large parameter space of possible network architectures. Randomly drawn models (green points) exhibit a significant correlation between performance and IT predictivity (r=0.55). Models optimized for categorization performance (blue points) show a significantly stronger correlation (r=0.78), even though the optimization was done without reference to neural data. Models directly optimized for IT predictivity (red points) show comparable correlation (r=0.80) --- but while the best models from performance-optimization predict IT neural output as well as the those explicitly optimized for explained variance, the reverse does not hold. We also assessed a variety of published ventral stream models, including a V1-like model, a V2-like model, and HMAX, a model targeted at IT. These results suggest that the CNN model architecture class encodes a relationship between high-level behavioral metrics and more detailed neural mechanisms, but that directed optimization accesses regions within parameter space where this constraint is much stronger. **Fig. 3** (below): We followed up on the observation that higher performance tend to lead to better models of IT, using recent methods to identify a high performing object categorization model (HMO) [1]. We then assessed the ability of this model to predict V4 and IT neural responses. (a) Qualitatively, the model's per-image predictions (red lines) mirrors the characteristic balance of selectivity and tolerance exhibited by actual IT neurons (black lines). (b) The HMO model predicts 48.3% of explainable variance (median over 168 measured IT units) a significant improvement over existing ventral stream models. (c) Each successive model layer predicts IT units increasingly well, suggesting that as more effective object recognition features are constructed at each processing stage, representations become increasingly IT-like. The model's penultimate layer is highly effective at predicting V4 neural responses (51.2% exp. var), while the IT-like top layer explains V4 responses significantly less well (34.3% exp. var.). These results suggest that performance optimization drives top-level output model layers to resemble IT and also imposes biologically consistent constraints on the intermediate feature representations.



[1] http://nips.cc/Conferences/2013/Program/event.php?ID=4078