Ventral stream models that solve hard object recognition tasks naturally exhibit neural consistency.

Daniel Yamins, Ha Hong, Ethan Solomon, and James J. DiCarlo
McGovern Institute for Brain Research and Dept. of Brain and Cognitive Sciences, MIT, Cambridge, MA

Humans recognize objects rapidly and accurately, a major computational challenge because low-level pixel data can undergo drastic changes in position, size, pose, lighting, occlusion, etc, while still containing the same high-level content. There is substantial evidence that the brain solves this challenge via a largely feedforward, hierarchical, network called the ventral visual stream. However, fundamental questions remain about the actual neural implementation, an understanding gap reflected in the difficulty computer models have had in equaling human performance. Here we describe models that perform substantially closer to human levels on a hard object recognition task, and in doing so, naturally discover representations consistent with experimentally observed high-level ventral stream neural populations.

We first constructed a large parameter set of hierarchical feedforward computational models, encompassing a variety of mechanisms that have shown promise in describing ventral stream encoding [1, 2]. To search this vast space for high-performing models, we developed a "Principled" High-Throughput (PHT) approach that blends powerful computational techniques with a structured selection procedure. The PHT procedure solves multiple recognition subtasks simultaneously, identifying target subtasks by error pattern analysis. Complementary model components emerge naturally, forming a representational basis that supports non-screened tasks. This process is repeated hierarchically, producing deep networks that are nonlinear combinations of lower-level components.

Models were constructed using this procedure with screening images containing objects on natural backgrounds, and then tested on neurophysiologically-measured images of entirely different objects in differing categories to rule out overfitting. The models showed major improvement in performance compared to existing computational models, even with the significant pose, scale, and position variation that typically hurt algorithm performance. They also exhibited feature representations strikingly similar to those observed in IT cortex, suggesting that the model's component substructures may predict identifiable functional motifs in higher-level ventral areas.

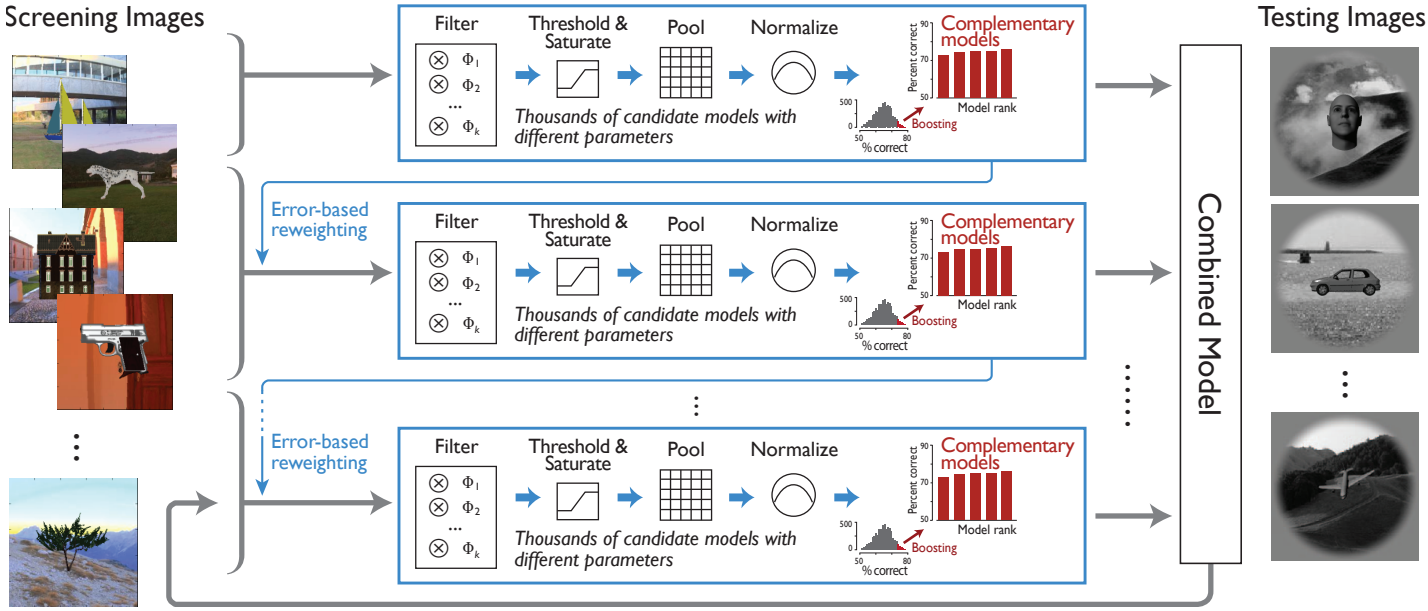

Fig. 1: The Principled High-Throughput (PHT) process. Parameterized models are screened for performance on a series of subtasks identified via error-based example reweighting. This identifies component models that are then combined and can either directly generate output features or serve as input to next hierarchical layer. Models trained on one imageset were then tested with a different imageset on which macaque neurophysiological and human behavioral data was also collected.

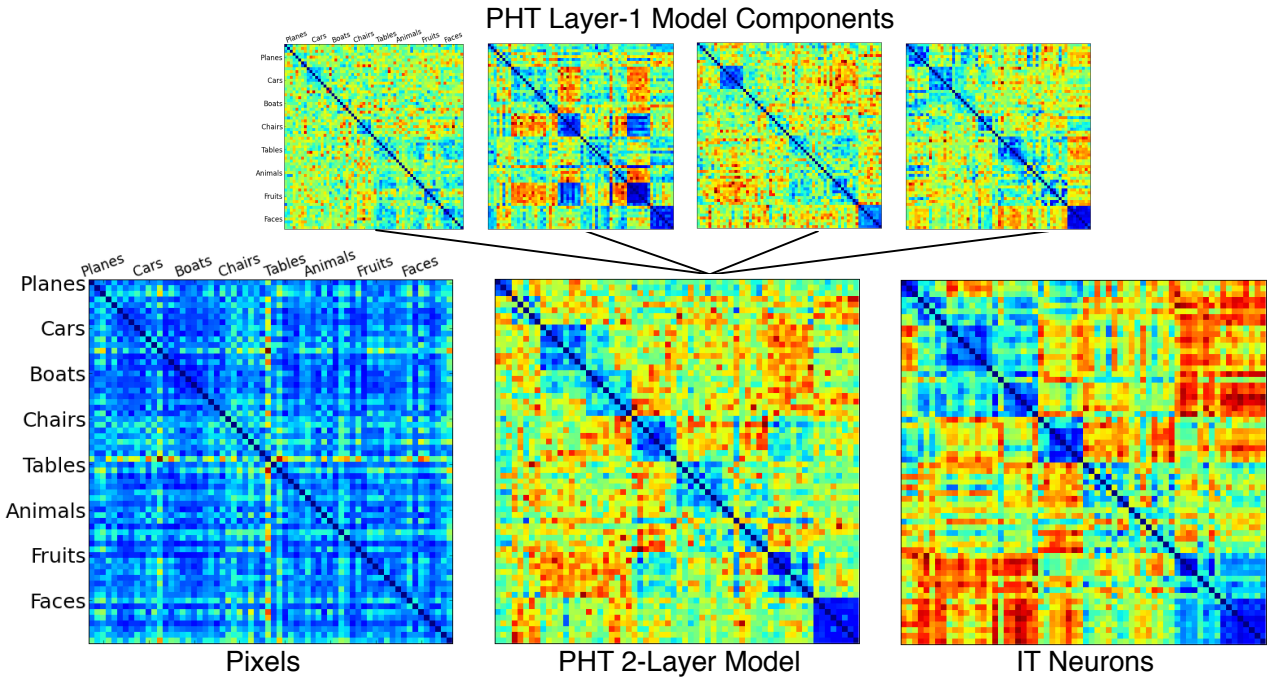

Fig. 2. Correlation matrices characterizing feature-level representations. Each matrix element represents the pairwise correlation between the object-average features of the model or neurons, for each of 64 pairs of objects in the test imageset. Top: four representative complementary components identified in the first layer of the PHT model construction process. Bottom Left: Pixels are a poor basis for object decoding in high variation tasks, as reflected in lack of block-diagonality in the matrix. Center: combined PHT 2-layer model composed of complementary layer-1 components, showing increased diagonality, corresponding to higher classification performance. Right: Neurons from electrophysiological recording of macaque inferior temporal (IT) cortex, a high-level ventral stream area, showing similarity to model even on off-diagonal blocks.

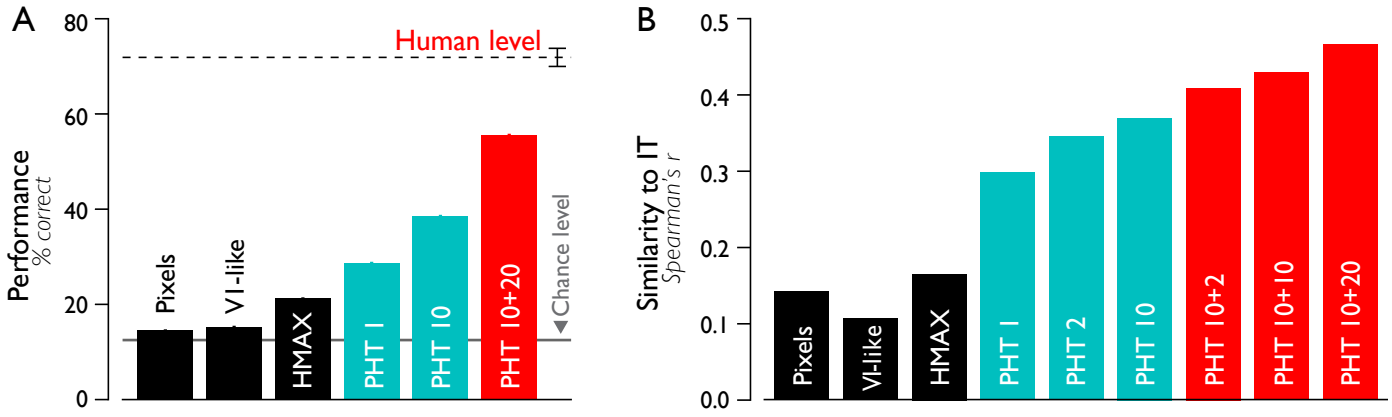

Fig. 3: A) Performance comparisons on a high-variation task 8-way object recognition task from the testing imageset. Gray line indicates chance performance (12.5%). Error bars compute training split uncertainty for models and subject variability for human psychophysics. Results show significantly increased transferable performance on hard object recognition task for PHT models. B) Similarity between feature representation matrix of models and IT neuron representation, measured as Spearman's *r* between matrices shown in Figure 2. Results show that as model complexity grows with added numbers of components and layers, similarity to IT representation increases.

[1] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox. A High-Throughput Screening Approach to Discovering Good Forms of Biologically Inspired Visual Representation. PLoS Computational Biology, 5(11), 2009.
[2] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci U S A, 104(15):6424–9, 2007.