Benchmarking Models of the Ventral Stream

by

Diego Ardila

B.S Biomedical Engineering
Johns Hopkins University, 2011

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN NEUROSCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2015

Signature of Author: Signature redacted
Dept. of Brain and Cognitive Sciences
May 7, 2015

Certified by: Signature redacted
James J. DiCarlo
Professor of Neuroscience
Department Head, Dept. of Brain and Cognitive Sciences
Thesis Supervisor

Accepted by: Signature redacted
Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

Abstract:

This work establishes a benchmark by which to measure models of the ventral stream using crowd-sourced human behavioral measurements. We collected human error patterns on an object recognition task across a variety of images. By comparing the error pattern of these models to the error pattern of humans, we can measure how similar to the human behavior the model's behavior is. Each model we tested was composed of two parts: an encoding phase which translates images to features, and a decoding phase which translates features to a classifier decision. We measured the behavioral consistency of three encoder models: a convolutional neural network, and a particular view of neural activity of either are V4 or IT. We measured three decoder models: logistic regression and 2 different types of support vector machines. We found the most consistent error pattern to come from a combination of IT neurons and a logistic regression, but found that this model performed far worse than humans. After accounting for performance, the only model that was not invalidated was a combination of IT neurons and an SVM.

Introduction

Why are models of the ventral stream interesting?

The human visual system is the largest sensory system, and has long been a target for modeling in computational neuroscience. The inputs are well defined and easily controllable, and there are several well-defined tasks that this system performs which are fairly computationally complex. One of these behaviors has long been of interest to systems neuroscience: object recognition. This task is generally thought to be supported mostly by a subsystem of the visual system called the ventral stream. Lesion studies and neural decoding have suggested that the ventral stream is a key driver of object recognition behavior (1, 2, 3). We operationally define this task as choosing mapping label from a set of possible alternatives to an image. The combination of a concretely defined task and some understanding of which parts of the brain are involved make this a very attractive target for computational modeling.

Why benchmark models?

Models of the ventral stream have a long history, and recently have seen rapid development due to GPU computing, advances in optimization methods, and access to large scale image data sets (7, 8). This rapid advance, which will likely continue, has brought about the need for assessing how “brain-like” these ventral stream models are.

This is important for neuroscience because one top-level goal of neuroscience is to embody understanding in an instantiated model: reproduce the brain in-silico.

Why behaviorally benchmark?

One route is to compare the activity of units within the models to the activity of units in the ventral stream. This is a promising direction which is ongoing and has seen recent developments. There are two practical limitations to this. First, neural data is very expensive and takes a long time to acquire. Part of this is related to the limit of how many neurons can be recorded per monkey, which is a purely technological limitation which will see quick development, but part of this is also due to how many monkeys a single scientist can record from, which is a more complex issue. Second, all methods capable of recording at the high spatial and temporal resolution required for detailed comparisons between a computational model and neural activity are invasive. This means they cannot be used on humans. While there are promising results showing the similarity of object recognition behavior between monkeys and humans, there is no guarantee that all visual behaviors of interest to neuroscience can be achieved by training rhesus macaques.

Given these limitations, we have begun to explore an alternative route: behavioral benchmarking. This method requires building a "full pipeline" model that goes from images to behavior. Any such model can be broken down into at least two subcomponents: an **encoding model** that goes from images to some internal representation, and a **decoding model** which produces behavioral outputs from this internal representation.

What previous behavioral benchmarks exist?

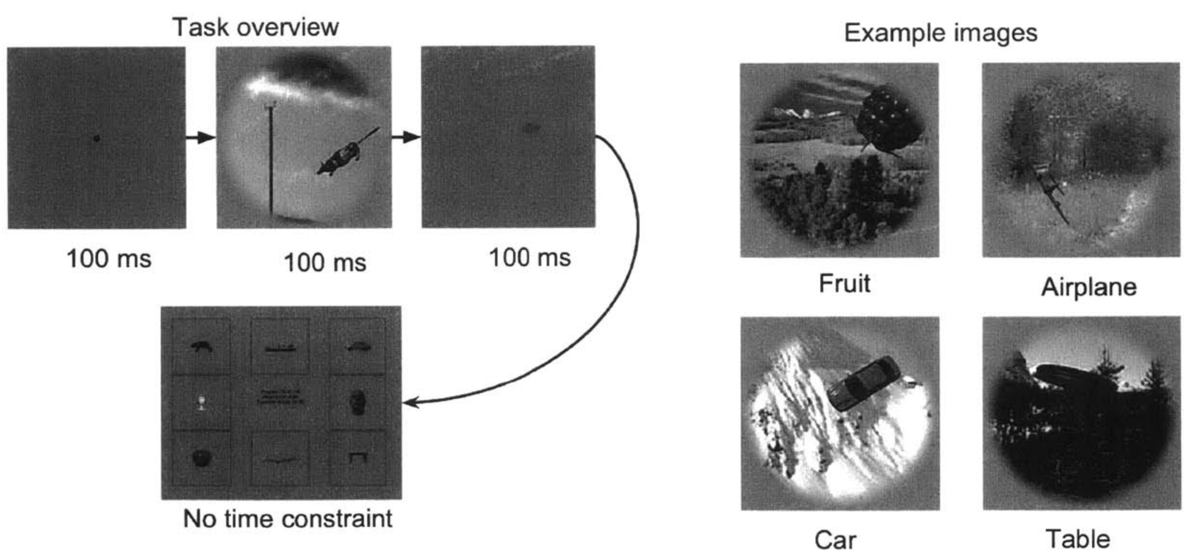
Previous work (1, 3) has benchmarked models on more coarse behavioral measures. This work has resulted in a model of how object recognition behavior is generated by the ventral stream which has a task-level error pattern that is consistent with human behavior. Within object recognition there are many different possible subtasks, such as detecting different objects and within category versus across category tasks. The existing model can predict the ranking of difficulty of these tasks for human observers: the same tasks it finds difficult, humans find difficult.

What is different about this benchmark?

In this work, we propose going a step further and analyzing behavioral patterns at an image level, instead of at a task level. We use models to make image by image predictions about behavior, then we compare these predictions to actual human behavior.

What is the target behavior for the models?

We collected human behavioral data for 128 images, resulting in an error pattern: a ranking of the difficulty of these images. A consistent model is one that ranks the images in the same order.



In the task, we presented a fixation dot for 100 milliseconds, the image for 100 milliseconds, and then a blank for 100 milliseconds. The subject then picked among 8 categories. The objects were shown at unusual angles on random backgrounds in order to probe invariant object recognition and avoid simple template-matching.

What models were benchmarked?

We chose to use three encoding models a neural one, a computational one, and a control neural model.

1. For a neural encoding model we chose a particular read of neurons in the inferotemporal cortex which was found in a previous study to be the most consistent with human behavior: The average activity from 100-170 ms after image onset, averaged over 30 trials. We refer to this encoding model from now on simply as "IT"
2. For a computational model, we chose the convolutional neural network with the best performance on a difficult object recognition task that was available for use, based on the hypothesis proposed in (PNAS paper) that higher performing computational models are more likely to predict neural activity, and therefore presumably will also produce behavior that is similar to humans. Since this model architecture was based on a publication by a group pat NYU, we refer to this model simply as "NYU"
3. For a control neural model we chose 128 multiunit sites of V4 neurons that were processed in the same way as those in IT (30 trial average, 100-170ms window), and refer to this as "V4"

We chose three different decoding models:

1. A multiway one versus all SVM, which is not probabilistic and was the model previously used which is consistent at the task level, for this reason we refer to it as the "Standard Model"
2. A logistic regression which makes probabilistic predictions. We refer to this simply as the "Logistic Regression Model"
3. A one vs. one SVM built out of many 2 way classifiers which makes probabilistic predictions. In order to produce probabilistic estimates from an SVM, a one vs. one scheme has been found in previous studies to be a good starting point (9). We refer to this as the "SVM probability estimate model"

The main difference between these models is that the second two make probabilistic predictions, and therefore can produce accuracies on a single image that are between zero and one. The "standard model" will either get an image correct or incorrect: the accuracy is either 0 or 1.

Methods

Computational Models

We trained a large convolutional network architecture specified in (8) in order to have a roughly state of the art model to benchmark against. We attempted to reproduce the methodology specified in that paper using the cudaconvnet library. The field is moving very quickly so by now this is already an outdated architecture. One line of future work is to benchmark more recent models the same way.

Decoding Models:

We used scikit-learn to implement the decoding models, using default parameters, except for the regularization parameter, which was selected using a grid search. For the "Standard Model" we used LinearSVC, which is one versus all. For the logistic regression model, we used the LogisticRegression class. For the "SVM probability estimate model" we used the SVC class, which defaults to a one vs. one scheme instead of a one vs all scheme.

Neural data:

The neural data was collected via a multi-electrode array, with 196 channels. The neural data used in this line of work is the same as in (4), where it is more thoroughly described.

Behavioral data

Using an a custom javascript library, we were able to get precisely timed presentations of images. We used 100 ms presentations both in order to match the presentation time of the neural data and in order to avoid eye movements and generate a dynamic range of errors which are the signal used in the behavioral benchmarking method. We showed 128 images to 25 subjects for 2 repetitions of each image. After each image 8 choices were presented, and the subject was asked to classify the image into one of 8 categories: airplane, animal, boat, car, chair, boat, face, fruit, or table. The subjects were rewarded a performance based bonus to incentivize maximum performance. The subjects were recruited via Amazon Mechanical Turk.

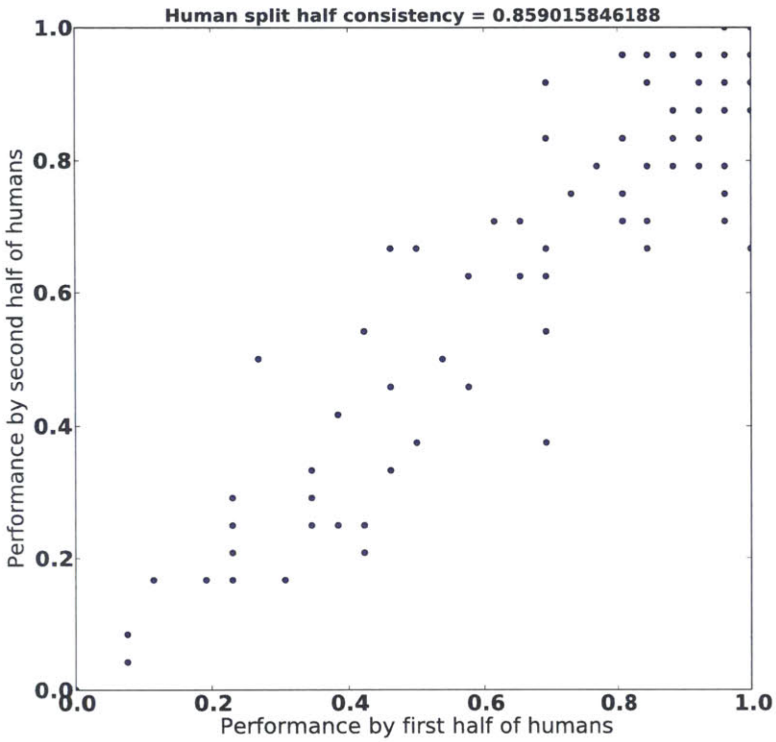
Fine-grained behavioral benchmarking

The data gathered on Mechanical Turk (mturk) was then used to calculate consistency as follows. First, an error pattern is produced by calculating the percentage correct for each image across repeated presentations of each image. This produces a set of error patterns: one per individual. Then, a pooled error pattern is calculated by taking the mean across subjects. The goal is then to calculate how much each human individual looks like the pooled error pattern. This correlation will by how internally consistent the human individuals are and so we estimated this limit by splitting the data in half, and normalizing by it. Specifically, we split each individual's trials in two halves, and correlated the error pattern between the two halves, H1, H2 to calculate internal consistency of that individual: Rh. The pool internal consistency was also calculated by splitting the trials in half, P1, P2 and correlating between the two resulting error patterns: Rp. Then the consistency between each possible choice of human (H1, H2) and pool (P1, P2) halves was calculated, then normalized by the geometric mean of Rh and Rp. This process was repeated for 1000 iterations of different splittings and then the median of all the normalized consistencies was used as the consistency for that individual. This results in a distribution of consistencies to the pool.

Extrapolation:

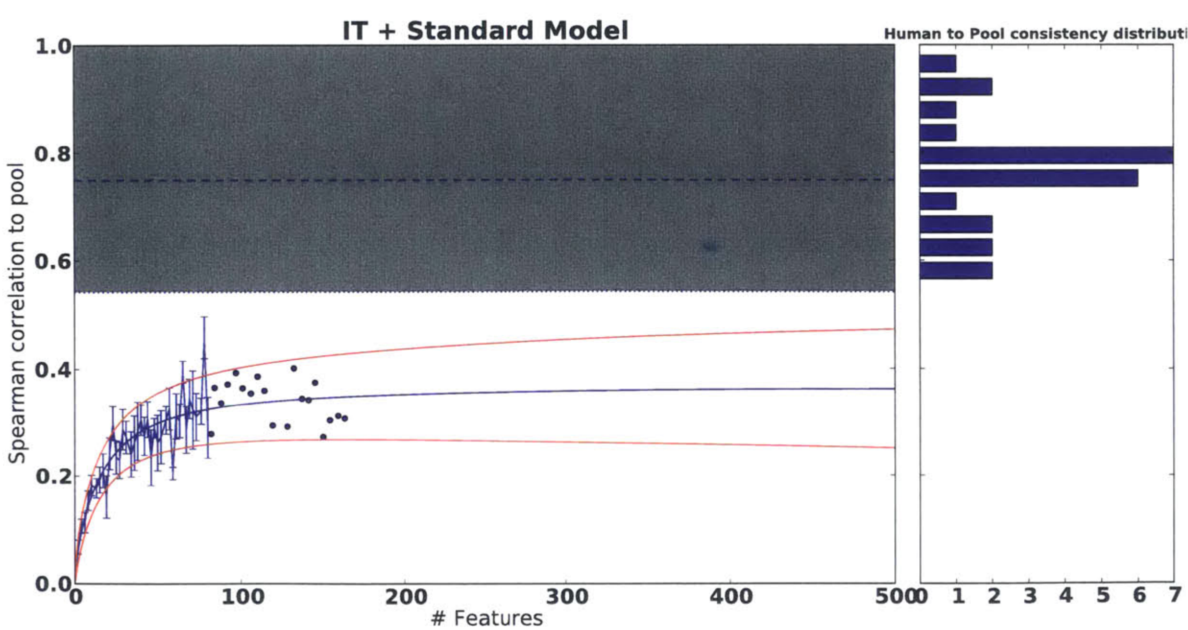
Consistency increases as more sites are added and so we used extrapolation to make some inferences about how different neural models may behave if we had access to more sites. We fit a variety of functional forms to the data, and then selected one of the functional forms based on cross validation. We limited extrapolation to 500 units, as this is roughly the range where a previous study found that an extrapolation of object recognition performance of IT neurons intersects human level performance.

Results

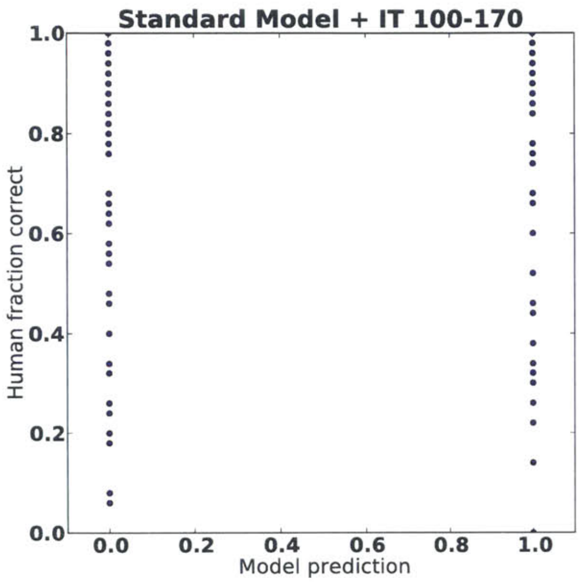


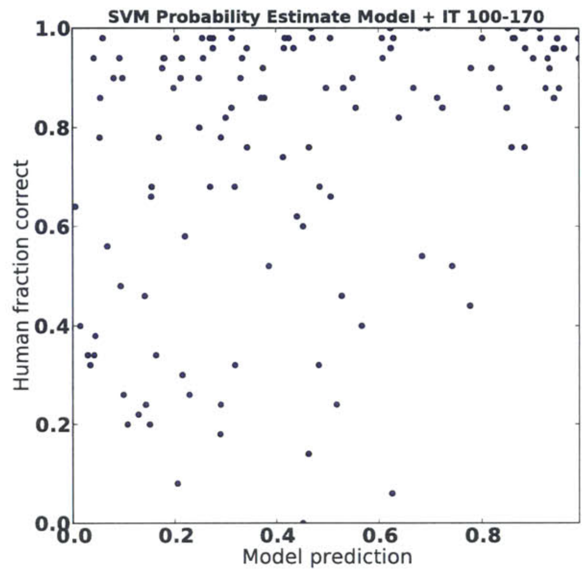
We collected the performance of humans on an 8-way object recognition task for 128 images. The human data is highly internally consistent: the first half of individuals and the second half of individuals have a very similar error pattern, as judged by the spearman correlation of these error patterns.

With this data, we compared the error patterns of the models to the error pattern of the humans, comparing against how well individual humans predicted the pooled human error pattern. The first main result is that the previous benchmark model: IT neurons + our previous "standard model", a non probabilistic SVM, fails to produce error patterns that fall within the distribution of human to pooled human consistencies. The error pat-



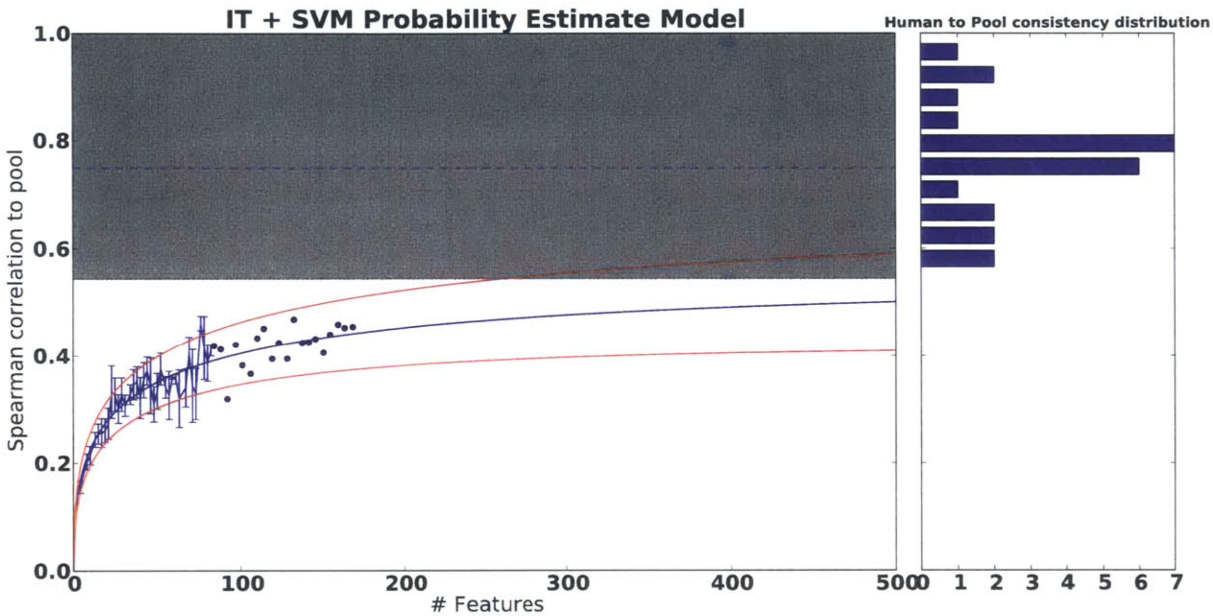
terns from this combination of encoding and decoding model are less consistent with the pooled human error pattern than any of the humans found in the sample, even if we extrapolate out the trend in increasing consistency to 500 units.



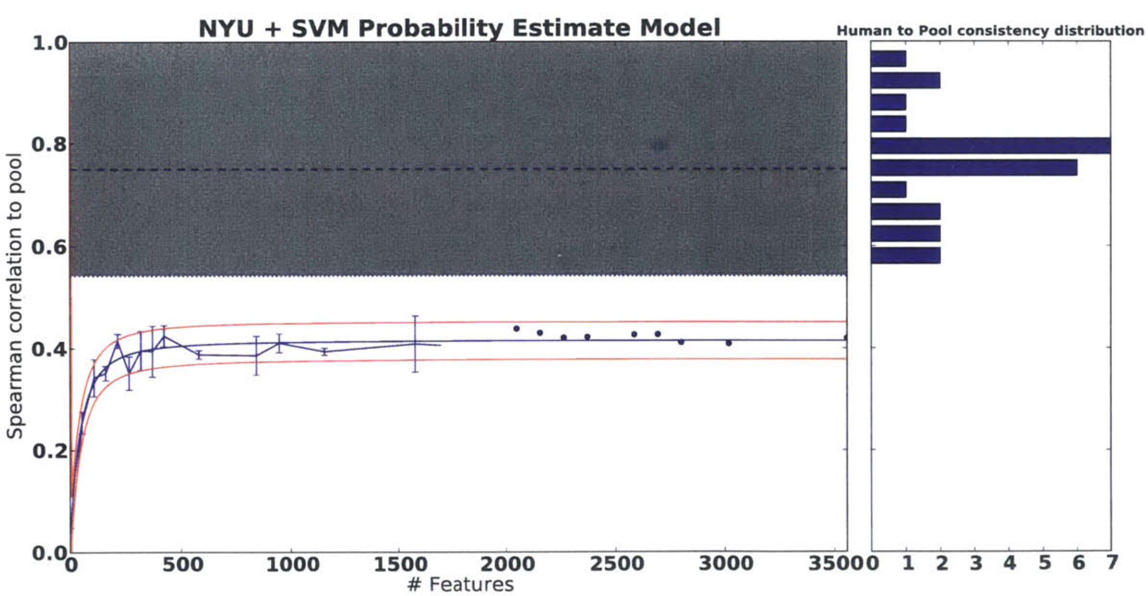


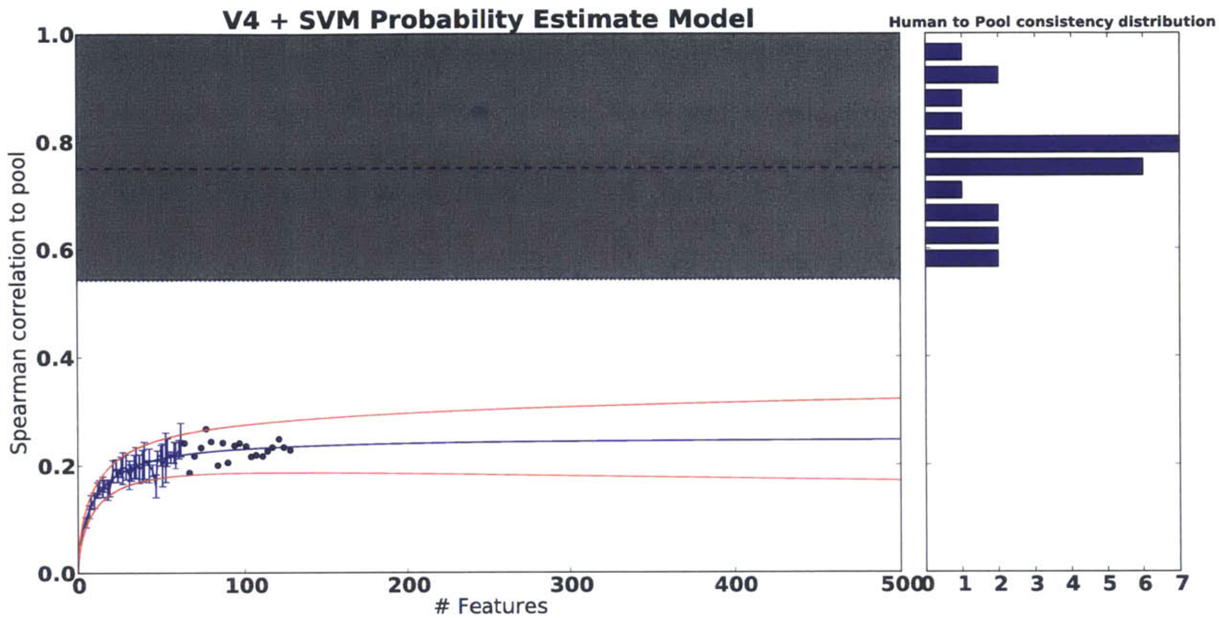
This means that as a model of human behavior, this particular read of IT plus an SVM model are strongly invalidated. If we look closely at the correlation between the error patterns of the model and the human pool, immediately the problem is clear: the predictions of the model can only be either correct or incorrect on a per image basis, which makes the correlation between the model predictions and the observed human behavior suffer.

If we instead use a decoding model that can produce probabilistic predictions then we can estimate the probability of getting a given image correct by simply assuming that the model will pick each class in accordance with how probable it believes that the class is. That is, if probabilistic prediction for class 1 and class 2 are both 50%, then each class is chosen half the time. This means that if class 1 is the correct answer, then the model will be correct 50% of the time. This model extrapolates much better, and the error bounds on the extrapolation are within the human distribution by the time the extrapolation reaches 500 units, and therefore this model is not strongly invalidated.

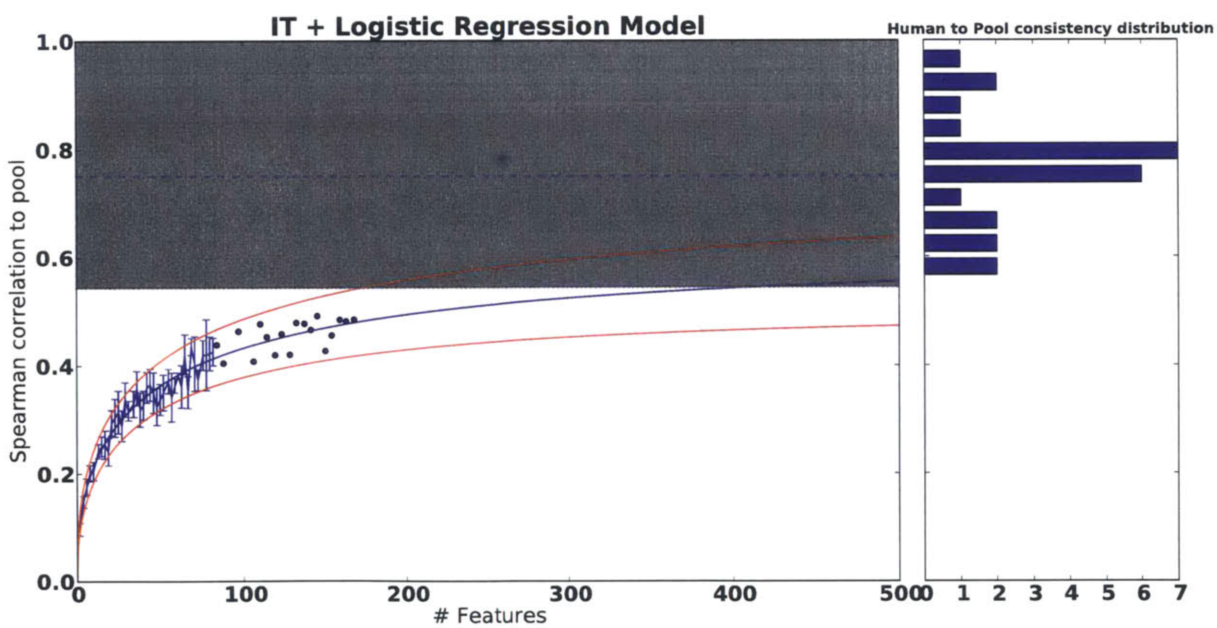


However, this decoding model does not rescue the other two encoding models tested from being strongly invalidated:

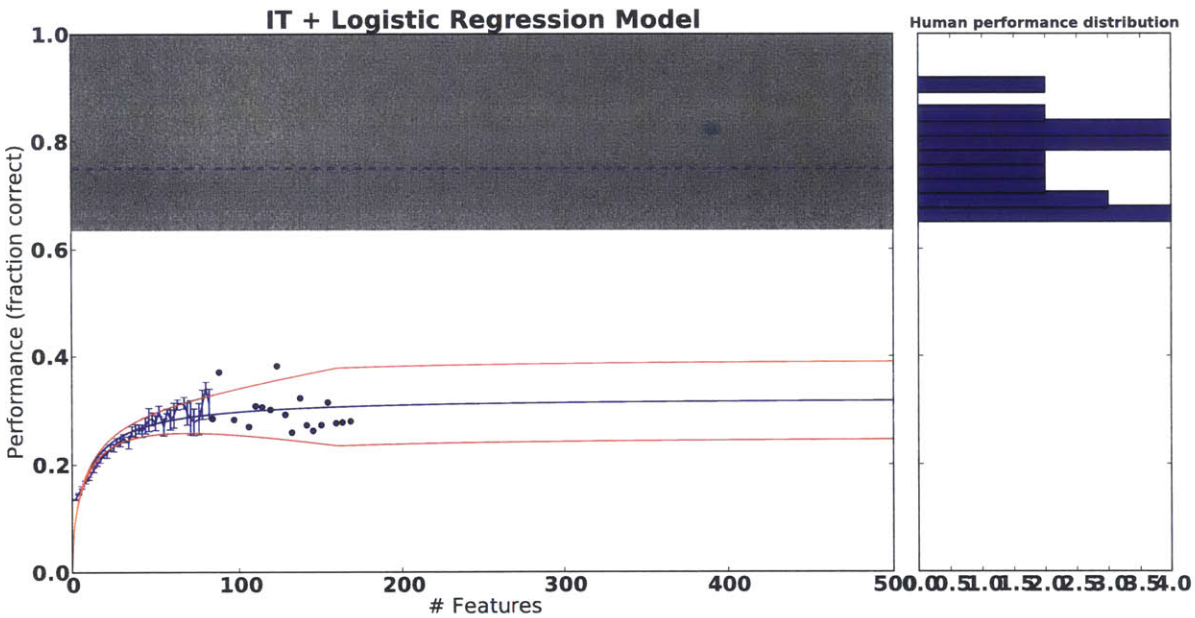


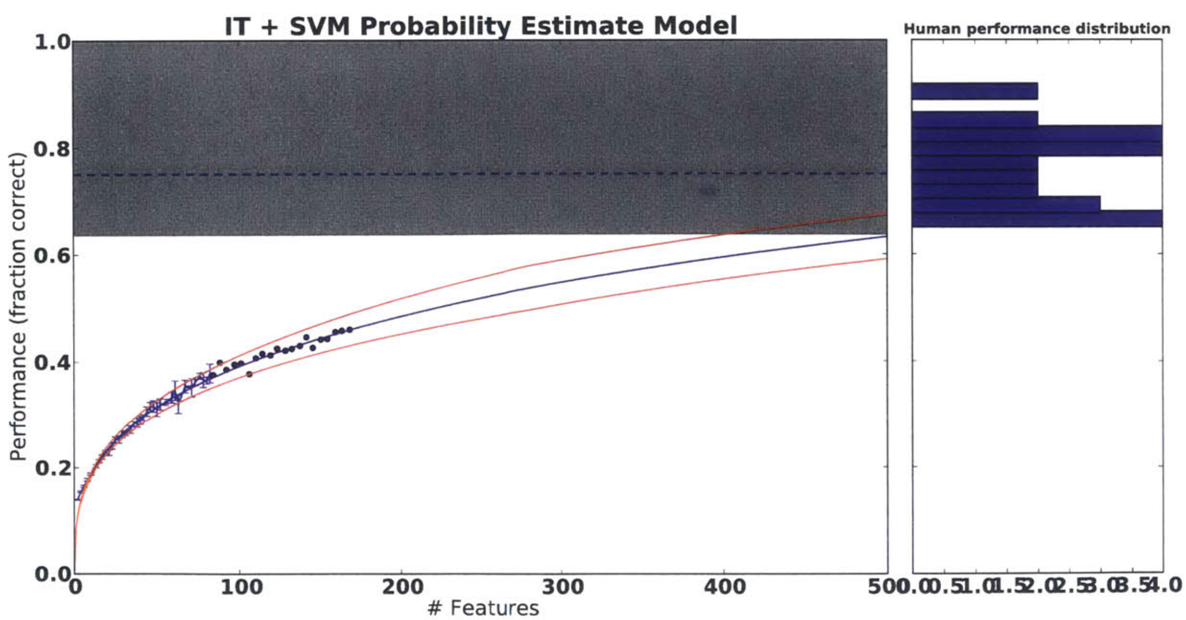


The logistic regression model appears to be slightly more consistent and is also not invalidated on consistency grounds:

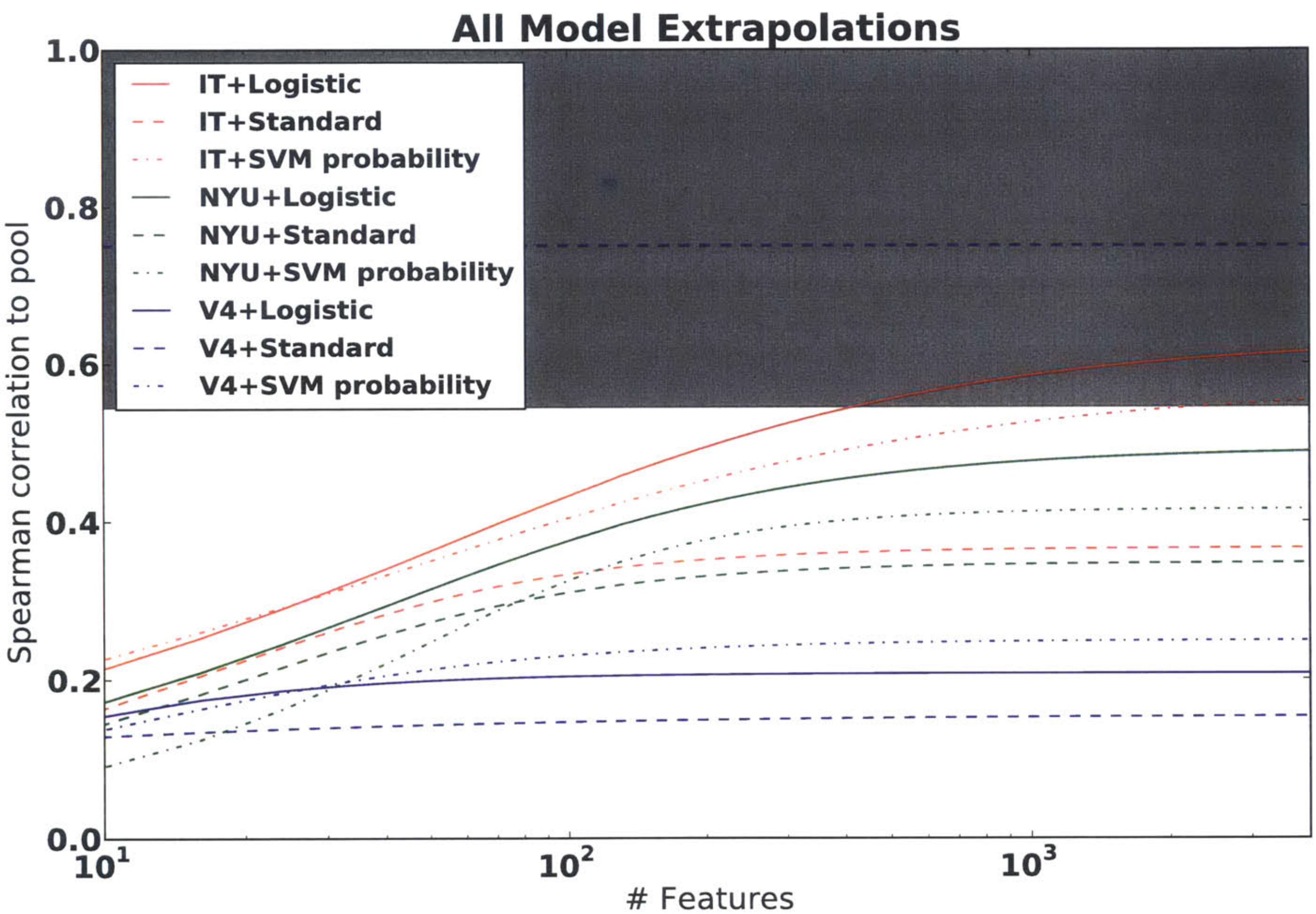


However, its performance (measured as the percentage of responses that are correct) appears to plateau at far below human levels. This means that IT+ a logistic regression is invalidated on performance grounds. This is, however, not an issue for the SVM probability estimate model, which does not plateau in performance as function of the number of features.





In summary, only two models ever enter the human distribution: IT+logistic, and IT+SVM probability estimates. Out of these two, the logistic regression appears to be invalidated on performance grounds.



Conclusions/Future directions

The main conclusion from this work is that in order to accurately predict human error patterns on an image by image basis, probabilistic estimates of class membership out-perform simple maximum operators over class margins. In addition, the particular read of IT we have used still seems like a promising encoding model, while the particular computational neural network model we used has been invalidated. This highlights a key future direction of this work: using the same benchmark in order to assess different computational models. The development of deep neural network models has become a very fast-moving field. This benchmark, among others, may be a useful way to assess whether the development of these models is still “on track” to become more brain-like, as opposed to purely performance driven. In addition, the stringency of this benchmark may be strengthened by taking a larger sample of images with more trials per human, perhaps allowing even the best existing models (IT 100-170 + Logistic regression) to be invalidated.

Bibliography

1. Najib Majaj, Ha Hong, Ethan Solomon, James DiCarlo, “A unified neuronal population code fully explains human object recognition,” Computational and Systems Neuroscience (COSYNE).

2. Cowey, A., and C. G. Gross. "Effects of foveal prestriate and inferotemporal lesions on visual discrimination by rhesus monkeys." Experimental Brain Research 11.2 (1970): 128-144.

3. Hung, Chou P., et al. "Fast readout of object identity from macaque inferior temporal cortex." Science 310.5749 (2005): 863-866.

4. Daniel Yamins*, Ha Hong*, Ethan Solomon, Darren Seibert, and James DiCarlo, “Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex,” *Proceedings of the National Academy of Sciences*, 111, 2014.

5. James DiCarlo, Ha Hong, and Daniel Yamins, “Modern population approaches for discovering neural representations and for discriminating among algorithms that might produce those representations.,” Vision Sciences Society Annual Meeting (VSS).

6. Daniel Yamins*, Ha Hong*, Charles Cadieu, and James DiCarlo, “Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream,” *Advances in Neural Information Processing Systems (NIPS)*, 26, 2013.

7. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

8. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." Computer Vision–ECCV 2014. Springer International Publishing, 2014. 818-833.

9. Wu, Ting-Fan, Chih-Jen Lin, and Ruby C. Weng. "Probability estimates for multi-class classification by pairwise coupling." The Journal of Machine Learning Research 5 (2004): 975-1005.