Linking image-by-image population dynamics in the macaque inferior temporal cortex to core object recognition behavior

Kohitij Kar (kohitij@mit.edu), **Kailyn Schmidt (kailyn@mit.edu)**, and **James J. DiCarlo (dicarlo@mit.edu)**
McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Abstract

Primates can rapidly identify visual objects; an ability supported by the ventral visual stream. We have recently reported that object information emerges in the inferior temporal (IT) cortex with distinct image-dependent dynamics. However, the current most parsimonious model that accurately links IT neural activity to primates' core object recognition behavior involves learned weighted sums of IT firing rates, specifically integrating the IT evidence over a single, fixed time window (70 to 170 ms post image onset). Here we collected new data to test whether this baseline model could accurately predict image-level primate object confusion patterns and found that it could not fully do so. Therefore we built and tested a more biologically-plausible linking model that implements leaky IT evidence accumulation. This models accurately predicts the monkeys' image-by-image behavioral patterns tested on 45 binary object discrimination tasks. Furthermore, we discovered that the trial-by-trial behavior of this same model partly predicts the animal's trial-by-trial choices on ambiguous images. Taken together, these results argue that IT population dynamics are relevant to core object recognition behavior and we provide a new, improved model of the mechanistic linkage between IT and core object recognition behavior.

Keywords: IT; core object recognition; dynamic decoder; population code; leaky evidence accumulation

Introduction

Previous studies on the neural mechanisms of primate visual object recognition (Hung, Kreiman, Poggio, & DiCarlo, 2005) have demonstrated that the identity and category of an object in an image at the center of gaze is often accurately conveyed in the population activity patterns of the inferior temporal (IT) cortex in macaques. To quantitatively link the IT population activity to the primates' behavioral patterns, Majaj et al. (2015) provided a simple linear decoding model (learned weighted sums of randomly selected average neuronal responses spatially distributed over monkey IT; LaWS of RAD IT) that was sufficient to explain and predict the average performance in each of a set of 64 tested core object recognition tasks. However, that study did not have the behavioral or neural resolution to assess if that linking model could accurately predict behavioral performance for each individual image. In addition, the linking model seemed somewhat biologically simplistic in that it integrated IT responses across a long, fixed temporal window (i.e. 70 ms to 170 ms triggered on image-onset). Using many more repeats of model-selected images, we have

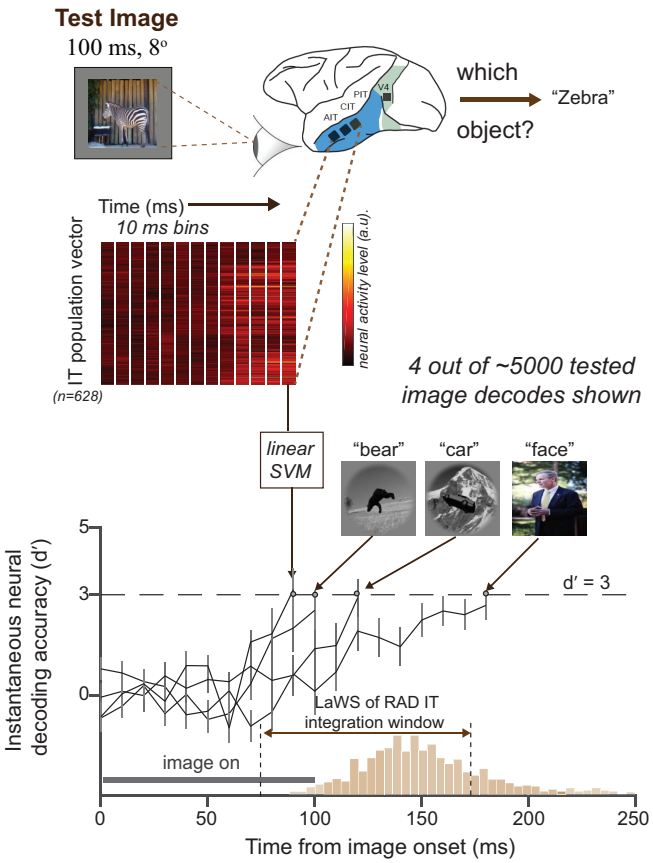

Figure 1: Temporal evolution of instantaneous linearly-decodable object identity evidence in IT on an image-by-image basis. For each tested image (top panel: example image containing zebra), we measured the IT population response vector (n=628) across time (10 ms bins). For each time bin, we built a linear decoder (cross-validated across images) to estimate the instantaneous neural decode (I_1; refer Methods) accuracy (entire time course shown in the bottom panel). Three additional example image dynamics are included in the bottom panel. The histogram shows a distribution of latencies of these instantaneous decoders across the tested images to reach a d' of 3.

recently observed that the temporal evolution of linearly separable object identity information in IT has reliably different dynamics for each image over the LAWS of RAD IT decoding window (see Methods; Figure 1, bottom panel; also refer to (Kar, Kubilius, Issa, Schmidt, & DiCarlo, 2017)). This suggests that the LaWS of RAD IT model could be put to a much stronger test, and possibly be falsified, if we obtained data to test it at image-level resolution. Therefore, in this study we first tested whether the baseline model (LaWS of RAD IT) proposed by Majaj et al. (2015) could explain primate image-level behavior. Upon the failure of this model to accurately predict the image-level behavioral error patterns of the monkeys, we considered a more biologically plausible linking model: a leaky integration of appropriately weighted IT responses. Our results demonstrate that this linking model could not only explain the prior results, but it also provided more accurate behavioral predictions for each image as well as better predictions of the monkeys' choices on individual trials.

Results

In the present work, we have compared the predictions of multiple candidate decoding hypotheses that link the primate ventral stream's neural activity with that of the monkeys' behavior. We have used a trial-averaged image-level behavioral metric (I_1), and a trial-by-trial choice correlation metric within a 2-AFC binary object discrimination paradigm.

Benchmarking the image-level behavioral patterns

To quantify the performance of the monkeys, we used a battery of 45 interleaved binary object discrimination tasks, similar to prior work (Rajalingham, Schmidt, & DiCarlo, 2015; Rajalingham et al., 2018). We collected a large number of behavioral trials to estimate the image-by-image behavioral performance pattern at relatively high SNR (median split half reliability, $\tilde{\rho}$ = 0.84 and 0.83 in monkey M and monkey N respectively). This pattern is referred to as I_1 (see Methods). Monkey M's I_1 pattern was 70.92% consistent with that of Monkey N, and we used the I_1 pattern and this value as a more stringent (relative to prior work) minimum target for any model linking IT population activity to behavioral choice.

Comparison of population decoding hypotheses

Testing LaWS of RAD IT Previous work from Majaj et al. (2015) has shown that a linking model that takes a learned weighted sum of IT population activity integrated between 70 to 170 ms post stimuli onset accurately predicts human behavioral confusion patterns. Thus, we first tested whether this baseline linking model could accurately predict the monkeys' behavioral I_1 pattern. Unlike primate-matched predictions of object confusion patterns reported earlier, we observed a $\sim$4 % divergence between the LaWS of RAD IT to monkey consistency and monkey-to-monkey I_1 consistency (Figure 2A). We reasoned that this inconsistency between monkey I_1 and the predictions of LaWS of RAD IT might be most prominent for images which have a slower rate of increase in the instantaneous linearly-decodable object identity evidence over time. Therefore, we specifically sub-selected the images where the instantaneous neural decode accuracy (refer bottom panel of Figure 1) took longer than 170 ms to reach a threshold d' of 3 (Figure 2B).

Testing different temporal pooling schemes Based on the failure of the LaWS of RAD IT model, we reasoned that downstream neurons might not be limited to integrating IT responses across a large, fixed temporal window. Instead, they might be integrating weighted sums of IT responses (which constitute pieces of evidence supporting different object choices) at shorter time intervals and over the entire duration until the choice screen was presented. Therefore, we tested a more biologically plausible leaky-integration model of appropriately weighted IT responses (see Methods). We observed that predictions of this model was not distinguishable from the monkeys' behavior (Figure 2A). In addition, even for the slow evolving images, they were highly consistent with the monkeys' I_1 pattern (Figure2C).

Testing different spatial pooling schemes In addition to testing how the dynamics in a randomly selected IT neural population might be optimally combined by a downstream neuron, we also specifically tested whether this leaky-integrator linking model is specific to any sub-region of IT (posterior, central and anterior) or might benefit with an additional read-out from area V4. We observed that, behavioral consistency was maximal when we considered the entire IT population, instead of V4, PIT, CIT and AIT separately, or a combination of V4 and IT (90 neurons sub-sampled randomly from each spatial pooling scheme). Of note, despite the highest consistency with monkey behavior observed with a random sub-sampling of IT population, overall accuracy of object decodes increased as a function of spatial hierarchy, i.e. V4$<$PIT$<$CIT$<$AIT (90 total neural sites per spatial pool).

Comparison of choice probabilities

The behavioral I_1 consistency tested so far is a trial-averaged metric. Given that we had access to each monkey's performance on every trial, we also asked whether the new and improved linking model could accurately predict the monkey's performance on a trial-by-trial basis. We reasoned that this would allow us to further test for potential inaccuracies in the two linking models. Our results show that while the predictions of LaWS of RAD IT is not significantly different from the chance-level (dashed red line in Figure 3C), the leaky-integrator model performs significantly better than chance (permutation test; $p<0.05$).

Conclusion

Our results demonstrate that the current baseline IT-to-behavior linking model, a learned weighted sum of IT responses, integrated across a specific temporal window, is insufficient to predict primate image-by-image behavioral error

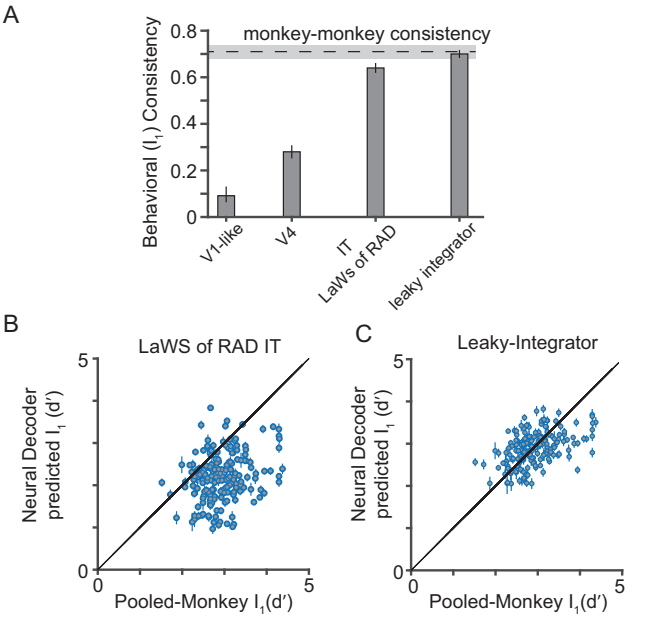

Figure 2: Comparison of different decoding hypotheses on an image-by-image monkey behavioral performance benchmark (I_1). The dashed line corresponds to the correlation of monkey M's I_1 with that of monkey N. The shaded region shows the bootstrap standard deviation of this correlation estimate. A) We show a comparison of behavioral I_1 correlation between a V1-like model based on VGGNet, V4, and two IT-based decoding hypotheses and the pooled-monkeys. B) Scatter plot of behavioral I_1: pooled-monkeys vs LaWS of RAD IT C)Scatter plot of behavioral I_1: pooled-monkeys vs the leaky-integrator linking model. For B) and C) only images where the instantaneous decodes took longer than 170 ms to reach a d' of 3 were considered.

patterns. Instead, we introduce a new linking model: a leaky-integration of appropriately weighted IT responses, until the availability of the choices. Our results demonstrate that this linking model captures the prior behavioral prediction results (Majaj et al., 2015), provides more accurate behavioral predictions over individual images, and provides better predictions of the monkeys' trial-by-trial behavior on ambiguous images.

Methods

Subjects

We have used two adult male rhesus monkeys (Macaca mulatta), referred to as monkey M and monkey N in the text, for simultaneous electrophysiology and behavioral data collection.

Visual stimuli

We used a combination of "naturalistic" images (3D models of objects rendered as 2D images with varied pose, position, size etc on an uncorrelated background; refer Majaj et al., 2015) as well natural images (photographs downloaded from

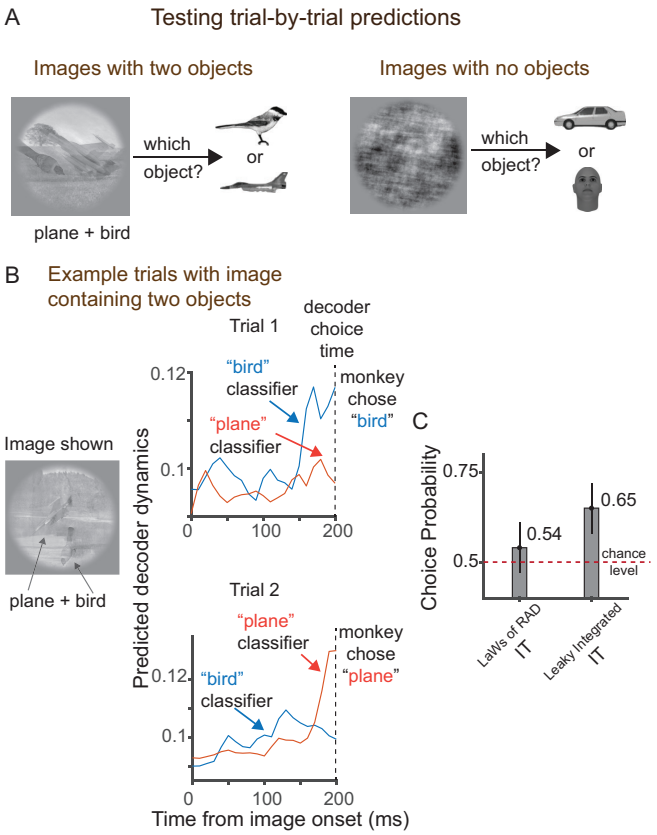

Figure 3: Choice correlation comparison between LaWS of RAD IT and the leaky integration decoding model. A) Examples of two specific task types included to compute choice probabilities. Left panel: example of images with two objects. The choice screen included both the objects. Right panel: example phase scrambled image tested with two randomly chosen object choices. B) Demonstration of successful decoding from neurons on specific trials. We show two example trials (monkey N), tested with the image shown on the left. Trial 1: monkey chose "bird". The bird-vs-rest classifier and the plane-vs-rest values at the decoder choice time successfully predicts the animal's choice. Trial 2 demonstrates a scenario when for the same image, the monkey chose "plane". C) Comparison of choice probabilities estimated across different ambiguous images with LaWS of RAD IT and leaky-integration model. Error bar denotes bootstrapped s.e.m across images. Data from the two monkeys were treated separately to generate trial-by-trial predictions, then pooled together for estimating the mean (shown in the figure)

http://cocodataset.org.)

Binary object discrimination task

Monkeys fixated a white square dot (0.2°) for 300 ms to initiate a trial. The trial started with the presentation of a test image for 100 ms, followed by a blank gray screen for 100 ms, after which the choice screen containing a target and a distractor object was shown. The monkey indicated his final choice by holding fixation over the selected image.

Image-level behavioral metric: I_1

We have used the same one-vs-all image-level behavioral performance metric (I_1) to quantify the performance of monkeys and neural based decoding models for the binary object discrimination tasks. This metric estimates the overall discriminability of each image containing a specific target object from all other objects (pooling across all 9 possible distractor choices). For example, given an image of object 'i', and all distractor objects ($j \neq i$) we first compute the average hit rate, $HitRate^i_{image}$. We then compute the false alarm rate for the object 'i' ($FalseAlarm^i$) The unbiased behavioral performance, per image, was then computed using a sensitivity index d',

$$d'_{image} = z(HitRate^i_{image}) - z(FalseAlarm^i), \qquad (1)$$

where z is the inverse of the cumulative Gaussian distribution. The values of d' were bounded between -5 and 5.

Large-scale electrophysiology

For each monkey, we implanted three chronic Utah arrays in IT, on one hemisphere, and a combination of one V4 array and two IT arrays on the other hemisphere. Recording sites that did not yield a significant visual drive or high response reliability were discarded from the analyses. In total, we had 628 valid IT sites and 166 valid V4 sites (n = 2 monkeys).

Neural Population Decoders

To model how downstream neurons might "read" ventral stream dynamics to infer object identities, we constructed multiple candidate linking models (listed below) that convert neural responses into a prediction of behavioral choice.

LaWS of RAD IT Similar to Majaj et al. (2015), to construct the learned weighted sums of randomly selected average (70-170ms post image onset) neuronal responses spatially distributed over monkey IT (aka LaWS of RAD IT linking model), we used a support vector machine algorithm with linear kernels. We used L2 regularization (strength of regularization, optimized for each train-set) and a stochastic gradient descent solver to estimate 10 (one for each object) one-vs-all classifiers. After training each of these classifiers with a set of 100 training image-responses per object, we generated a class score (sc) per classifier for all held out test images. We then converted the class scores into probabilities (P^i_{image}) by passing them through a softmax (normalized exponential) function. Binary task performances were computed as the percent correct score ($Pr^{i,j}$) for each pair (i, j) of object choices given an image.

$$Pr^{i,j}_{image} = \frac{P^i_{image}}{P^i_{image} + P^j_{image}} \qquad (2)$$

We then estimated a neural I_1 score (derived from $Pr^{i,j}$), following the same procedure as the behavioral metric.

Leaky-Integrated read-outs To construct the leaky integrator model, we first estimated a set of learned weighted sums (as above) of average IT responses (time window of averaging was chosen to maximize cross-validated behavioral performance) to construct 10, one-vs-all classifiers, each belonging to one of the tested objects. We then implemented a leaky integration rule of updating each of the classifier outputs continuously across time according to the following equation,

$$\tau \frac{dC(t)}{dt} = -C(t) + I(t) \qquad (3)$$

, where C is the classifier output at each time point 't', and I is the weighted input from IT at that time point 't'. The value of τ was estimated to be 40ms when optimized for behavioral consistency (cross-validated across images). The decision was estimated by comparing the classifier values (similar to LaWS of RAD IT) at the time at which the choice screen was shown (i.e. 100 ms post image offset; as demonstrated in Figure 3B).

Acknowledgments

This research was supported by US National Eye Institute grants R01-EY014970 (J.J.D.), Office of Naval Research MURI-114407 (J.J.D) and grants from the Simons Foundation (SCGB [325500, 542965], [JJD]). We also thank Arash Afraz and Elias Issa for technical help.

References

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Kar, K., Kubilius, J., Issa, E., Schmidt, K., & DiCarlo, J. (2017). Evidence that feedback is required for object identity inferences computed by the ventral stream. *Computational and Systems Neuroscience Annual Meeting 2017*.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *bioRxiv*, 240614.

Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, *35*(35), 12127–12136.