Selective behavioral deficits from focal inactivation of primate inferior temporal (IT) cortex: a new quantitative constraint for models of core object recognition

Rishi Rajalingham (rishir@mit.edu), **Hyodong Lee (hyo@mit.edu)**, **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

Primate core visual object recognition is thought to rely on the ventral visual stream, a hierarchy of cortical areas culminating in inferior temporal (IT) cortex. Previous work has shown that the IT population responses accurately predict primate object recognition behavior, suggesting that these IT codes underlie these behaviors. However, direct causal evidence for this decoding hypothesis has been equivocal at best, especially beyond the specific case of face-selective sub-regions of IT. Here, we tested the general causal role of IT in core object recognition by reversibly inactivating individual, millimeter-scale regions of IT via injection of muscimol while monkeys performed several binary object discrimination tasks, interleaved trial-by-trial. Our results show that inactivating different millimeter-scale sub-regions of primate IT resulted in different patterns of task deficits. These results provide new constraints for computational models of the ventral stream and this behavior. To this end, we tested state-of-the-art deep convolutional neural network models by constructing topographic deep artificial neural networks (TDANNs) on which we simulated inactivation experiments. Our results show that TDANNs recapitulated first-order experimental phenomena slightly better than randomly mapped deep artificial neural network models. Taken together, these results establish and test a new class of experimental constraints for computational models of core object recognition.

Keywords: object recognition; deep neural network; primate; inactivation; topography

Introduction

Primate core visual object recognition — the ability to rapidly recognize objects in spite of naturally occurring identity-preserving image variability — is thought to rely on the ventral visual stream, a hierarchy of visual cortical areas (DiCarlo et al., 2012). In particular, decades of research suggest that inferior temporal (IT) cortex, the highest level of the ventral stream hierarchy, is a necessary part of the brain's neural network that underlies core recognition behavior (Logothetis & Sheinberg, 1996; Tanaka, 1996; Rolls, 2000; DiCarlo, Zoccolan, & Rust, 2012). For example, it has been shown that the population of neurons in IT not only matches overall primate behavioral performance (Hung, Kreiman, Poggio, & DiCarlo, 2005; Zhang et al., 2011) but also predicts primate behavioral patterns (Sheinberg & Logothetis, 1997; de Beeck, Wagemans, & Vogels, 2001; Majaj, Hong, Solomon, & DiCarlo, 2015), suggesting that IT is a good neural correlate of primate recognition behavior. These observations are consistent with the causal dependency of core object recognition behavior on IT, but could also reflect epiphenomenal mechanisms (e.g.(Katz, Yates, Pillow, & Huk, 2016; Liu & Pack, 2017)). For clarity, we adopt the terminology of (Jazayeri & Afraz, 2017), whereby causal dependencies link a dependent variable to an experimentally controlled variable, in contrast to correlational dependencies (associations that we measure but do not control). Thus, to infer a causal link between activity in IT and behavior, it is necessary to directly manipulate activity in IT (e.g. via the application of pharmacological agents into IT to silence neurons, etc.) while measuring behavior.

To date, the most successful direct manipulations of IT exclusively targeted millimeter-scale clusters of face-selective neurons in IT (S.-R. Afraz, Kiani, & Esteky, 2006; A. Afraz, Boyden, & DiCarlo, 2015; Moeller, Crapse, Chang, & Tsao, 2017; Sadagopan, Zarco, & Freiwald, 2017), and suggest that these IT sub-regions are necessary for at least some basic- and subordinate-level face recognition behaviors. However, results from direct manipulations of IT in general visual recognition behavior have been equivocal at best. Lesions of IT sometimes suggest the necessity of IT and visual behaviors (Cowey & Gross, 1970; Manning, 1972; Holmes & Gross, 1984; Biederman, Gerhardstein, Cooper, & Nelson, 1997; Buffalo, Ramus, Squire, & Zola, 2000) but the resulting behavioral deficits are often contradictory (with often no lasting visual deficits) (Dean, 1974; Huxlin, Saunders, Marchionini, Pham, & Merigan, 2000) and at best modest (Horel, Pytko-Joiner, Voytko, & Salsbury, 1987; Matsumoto, Eldridge, Saunders, Reoli, & Richmond, 2016). Thus, it is still unclear if IT is necessary for general core object recognition behavior, and furthermore if that assumed causal role is spatially organized.

Results

To investigate these open questions, we here reversibly inactivated individual, arbitrarily sampled millimeter-scale regions of IT via local injection of muscimol while monkeys performed several (6 or 10) binary core object discrimination tasks between five objects, interleaved trial-by-trial (see Figure 1A for behavioral paradigm). To enforce true invariant recognition, stimuli consisted of naturalistic synthetic images of 3D objects rendered under high view-uncertainty. Figure 1B-D shows the behavioral deficits for six tasks, for each of three example inactivation experiments, plotting the relative performance for control and muscimol inactivation conditions. For each experiment, the corresponding anatomical location of the inactiva-

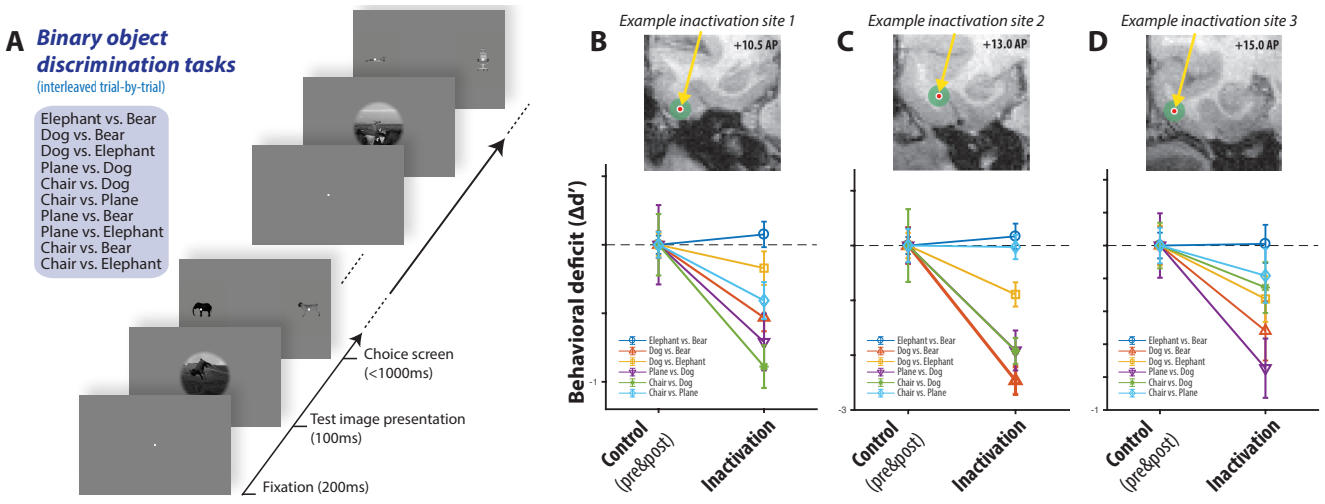

Figure 1: (A) Behavioral paradigm. The list shows all tested pairwise object discrimination tasks between five objects, interleaved trial-by-trial. Each trial was initiated when the monkey acquired and held gaze fixation on a central fixation point for 200ms, after which a test image (8x8 degrees of visual angle in size) appeared at the center of gaze for 100ms. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right; one of these two objects was always the same as the object that generated the test image (i.e. the correct choice), and its location (left or right) was randomly chosen on each trial. The monkey was allowed to freely view the choice images for up to 1000ms, and indicated its final choice by holding fixation over the selected image for 700ms. Animals were rewarded with small juice rewards for successfully completing each trial. After the end of each trial, another fixation point before the next test image appeared. Each trial consisted of a different randomly selected binary task. (B-D) Example inactivation experiments. For three example inactivation experiments in three different IT sites, the resulting behavioral deficits over tasks and the corresponding anatomical locations are shown.

tion site is shown in the inset. For each inactivation experiment, we observed a strong and significant deficit for some tasks but not others.

We quantify the task selective deficits resulting from focal inactivations in Figure 2. Over all inactivation sites ($n = 25$ in two monkeys), we observed a significant decrease in performance of $\mu_\delta = -0.2 \pm 0.02$ in units of d' ($p = 1.23 * 10^{-16}$, one-tailed exact test; see Figure 2A left panel red bar). We observed no such behavioral deficit on otherwise identical experiments but without inactivation, ($mu_\delta = 0.02 \pm 0.03$, $p = 0.78$; one-tailed exact test; see Figure 2A left panel, blue bar). We quantified this task-selectivity by computing a sparsity index (SI) for each inactivations behavioral deficit pattern, i.e. the sparsity over tasks. This index has a value of 0 if all tasks are equally affected, and a value of 1 for a perfectly task-specialized or one-hot deficit pattern. Figure 2A (right panel) shows that inactivation of local regions in IT leads to highly non-uniform deficits ($SI = 0.71 \pm 0.05$; mean$\pm$SE over sites); this degree of task selectivity is greater than expected for a uniform deficit ($p = 2.42 * 10^{-16}$; relative to simulated uniform, see Figure 2A right panel) but significantly less than expected for a one-hot deficit pattern ($p = 5.28 * 10^{-3}$; relative to simulated one-hot, see Figure 2A right panel).

These data establish new quantitative constraints for computational models of core object recognition behavior. That is, the correct model of the ventral stream should produce the same magnitudes and types of deficits when the neurons in that model are locally suppressed. While we and others have previously shown that deep CNN models are good models of the response of ventral stream neurons and behavior (Rajalingham, Schmidt, & DiCarlo, 2015; Jozwik, Kriegeskorte, & Mur, 2016; Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016; Kubilius, Bracci, & de Beeck, 2016; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Guclu & van Gerven, 2015; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016), these models are limited in that they do not specify how different types of neurons in each model visual area (a.k.a. model layer) are spatially organized. Thus, we have built specializations of this family of models that have topographic organization that we refer to as Topographic Deep Artificial Neural Networks (TDANNs, , schematized in Figure 2B). To build a TDANN, we adapted state-of-the-art deep convolutional neural network models. To do so, we first measured the profile of response correlations versus cortical distances for thousands of pairs of neurons recorded from macaque IT cortex of monkeys that were not used in this study; this response profile served as a topographic constraint to the models topographic cost function. In essence, the cost function expresses the idea that neurons with similar patterns of responses (over images) should try to be spatially close, and neurons with different patterns of responses should be spatially far.

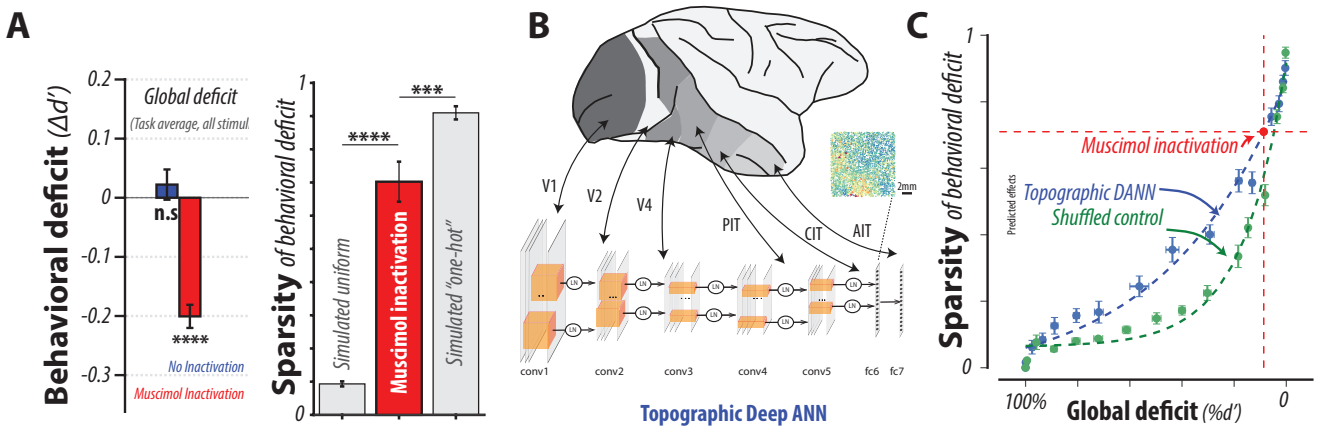

Figure 2: (A) Quantitative measurements of deficit magnitude and sparsity over $n = 25$ inactivation experiments in two monkeys. (B) Schematic of the Topographic Deep Artificial Neural Networks (TDANNs) models. TDANNs are adaptations of the Alexnet (Krizhevsky et al., 2012) architecture with tissue maps attached to the first fully connected (fc6) layer (Figure 2B). Networks were simultaneously optimized for image classification task on the ILSVRC-2012 dataset, and penalized if the model unit response profile did not match the experimentally derived spatial response profile. (C) TDANNs recapitulated some first-order phenomena, namely the relationship between magnitude and sparsity of behavioral deficits (blue), slightly better than randomly mapped deep artificial neural network models (green). The empirical values from (A) are plotted in red.

The TDANNs presented here are adaptations of the Alexnet architecture (Krizhevsky et al., 2012) with tissue maps attached to the first fully connected (fc6) layer (Figure 2B). In order to simulate the topographic maps, each network unit was initially assigned a random position in a two-dimensional surface. Networks were simultaneously optimized for image classification task on the ILSVRC-2012 dataset and for spatially organizing those learned neural response types according to the topographic cost function (above). For reproducibility, 10 TDANNs were trained with different parameter initializations.

With these TDANNs in hand, we could ask how well they predicted the IT focal inactivation results. To do this, we simulated focal inactivation experiments, by zero-ing spatially contiguous subsets of features in the fc6 layer and propagating the simulated responses to a (pre-trained) behavioral readout. We then characterized the differences in behavioral responses between the intact and inactivated models, for a large number of randomly localized inactivations, while also varying the size of the inactivations.

Our results show that TDANNs recapitulated the observed experimental phenomena, namely the magnitude and sparsity of behavioral deficits (see Figure 2C, blue), slightly better than randomly mapped deep artificial neural network models (see Figure 2C, green). We do not claim this model class captures all aspects of the observed deficits, but rather provides a good starting point for that greater goal. Taken together, these results establish and test a new class of experimental constraints for computational models the ventral visual stream and its role in core object recognition.

References

Afraz, A., Boyden, E. S., & DiCarlo, J. J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination [Journal Article]. *Proceedings of the National Academy of Sciences*, *112*(21), 6730-6735.

Afraz, S.-R., Kiani, R., & Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization [Journal Article]. *Nature*, *442*(7103), 692-695.

Biederman, I., Gerhardstein, P. C., Cooper, E. E., & Nelson, C. A. (1997). High level object recognition without an anterior inferior temporal lobe [Journal Article]. *Neuropsychologia*, *35*(3), 271-287.

Buffalo, E. A., Ramus, S. J., Squire, L. R., & Zola, S. M. (2000). Perception and recognition memory in monkeys following lesions of area te and perirhinal cortex. *Learning & Memory*, *7*(6), 375–382.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition [Journal Article]. *PLoS computational biology*, *10*(12), e1003963.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*, 27755.

Cowey, A., & Gross, C. (1970). Effects of foveal prestriate and inferotemporal lesions on visual discrimination by rhesus monkeys [Journal Article]. *Experimental Brain Research*, *11*(2), 128-144.

Dean, P. (1974). The effect of inferotemporal lesions on memory for visual stimuli in rhesus monkeys. *Brain research*, *77*(3), 451–469.

de Beeck, H. O., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature neuroscience*, *4*(12), 1244.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? [Journal Article]. *Neuron*, *73*(3), 415-434.

Guclu, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream [Journal Article]. *Journal of Neuroscience*, *35*(27), 10005-10014.

Holmes, E. J., & Gross, C. G. (1984). Effects of inferior temporal lesions on discrimination of stimuli differing in orientation [Journal Article]. *The Journal of Neuroscience*, *4*(12), 3063-3068.

Horel, J. A., Pytko-Joiner, D. E., Voytko, M. L., & Salsbury, K. (1987). The performance of visual tasks while segments of the inferotemporal cortex are suppressed by cold [Journal Article]. *Behavioural brain research*, *23*(1), 29-42.

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

Huxlin, K. R., Saunders, R. C., Marchionini, D., Pham, H.-A., & Merigan, W. H. (2000). Perceptual deficits after lesions of inferotemporal cortex in macaques [Journal Article]. *Cerebral Cortex*, *10*(7), 671-683.

Jazayeri, M., & Afraz, A. (2017). Navigating the neural space in search of the neural code. *Neuron*, *93*(5), 1003–1014.

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in it and perception with non-negative least squares. *Neuropsychologia*, *83*, 201–226.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation [Journal Article]. *PLoS computational biology*, *10*(11), e1003915.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition [Journal Article]. *Scientific reports*, *6*, 32672.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks [Conference Proceedings]. In *Advances in neural information processing systems* (p. 1097-1105).

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity [Journal Article]. *PLoS computational biology*, *12*(4), e1004896.

Liu, L. D., & Pack, C. C. (2017). The contribution of area mt to visual motion perception depends on training. *Neuron*, *95*(2), 436–446.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition [Journal Article]. *Annual review of neuroscience*, *19*(1), 577-621.

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 Article]. *The Journal of Neuroscience*, *35*(39), 13402-13418.

Manning, F. J. (1972). Serial reversal learning by monkeys with inferotemporal or foveal prestriate lesions [Journal Article]. *Physiology and behavior*, *8*(2), 177-181.

Matsumoto, N., Eldridge, M. A., Saunders, R. C., Reoli, R., & Richmond, B. J. (2016). Mild perceptual categorization deficits follow bilateral removal of anterior inferior temporal cortex in rhesus monkeys [Journal Article]. *Journal of Neuroscience*, *36*(1), 43-53.

Moeller, S., Crapse, T., Chang, L., & Tsao, D. Y. (2017). The effect of face patch microstimulation on perception of faces and objects [Journal Article]. *Nature Neuroscience*, *20*(5), 743-752.

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

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition [Journal Article]. *Neuron*, *27*(2), 205-218.

Sadagopan, S., Zarco, W., & Freiwald, W. A. (2017). A causal relationship between face-patch activity and face-detection behavior [Journal Article]. *eLife*, *6*, e18558.

Sheinberg, D. L., & Logothetis, N. K. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences*, *94*(7), 3408–3413.

Tanaka, K. (1996). Inferotemporal cortex and object vision [Journal Article]. *Annual review of neuroscience*, *19*(1), 109-139.

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 [Journal Article]. *Proceedings of the National Academy of Sciences*, 201403112.

Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., & Desimone, R. (2011). Object decoding with attention in inferior temporal cortex [Journal Article]. *Proceedings of the National Academy of Sciences*, *108*(21), 8850-8855.