Object-enhanced and Object-centered Representations Across Primate Ventral Visual Cortex

Tahereh Toosi (tahereh.toosi@columbia.edu)
Nikolaus Kriegeskorte (nk2765@columbia.edu)
Elias Issa (elias.issa@columbia.edu)
Zuckerman Mind Brain Behavior Institute, Columbia University, 3227 Broadway
New York, New York 10027

Abstract

Which features of an image are represented in each area of the visual cortex? Early visual areas are thought to extract primitive features such as edges but as we go deeper into the ventral visual cortex which supports object recognition there is less consensus on what features are represented in the population of neurons. Here, we extract and evaluate hypothesis-driven interpretable intermediate features ranging between pixels and object categories. We revealed interpretable representations in neuronal populations of the mid- and high-level ventral visual stream that progress from being retinally-based to object-centered to coding object identity at the highest stage. Furthermore, object-centered representations came to dominate the neural response over time suggesting a dynamical process. These interpretable features explained the neuronal response at a comparable or sometimes better level than feedforward deep neural networks with millions of parameters that are currently considered the best models of natural object recognition.

Keywords: object recognition; primate; deep neural networks; representation

Introduction

How do the neural representations evolve from retinal light-sensitive cells to the object selective cells in the Inferior Temporal (IT) cortex? Classical experiments showed that single neurons in the primary visual cortex (V1) respond with more spikes to Gabors of specific orientations, while neurons in the IT cortex respond to specific objects. There is less agreement on what image features are extracted by the neurons in between these areas. Deep Neural Networks (DNNs) learn intermediate features trained to transform the pixel input into semantic category output, and the patterns of activation of their resulting internal features are similar to those observed in neurons of the visual cortex (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). Thus, the match between DNN features and cortical neurons currently offers the best quantitative explanation for higher visual cortical areas. However, DNNs are often criticized for their lack of interpretability and explainability, especially as models of object recognition in the brain. So despite the simultaneous recording of hundreds of neurons from intermediate regions of the ventral stream and successful prediction of this activity with DNNs, the nature of intermediate, hidden layer neural representations remained largely unknown.

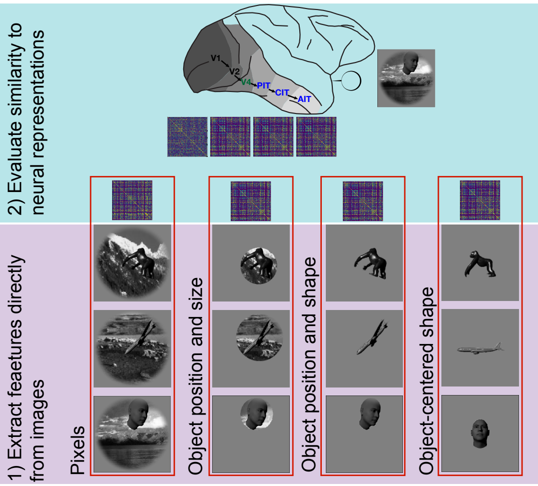
2) Evaluate similarity to neural representations
V1
V2
V4
PIT
CIT
AIT
1) Extract feaetures directly from images
Pixels
Object position and size
Object position and shape
Object-centered shape

Figure 1: Meta-Space Representational Analysis. We take the natural images the monkeys viewed (first column) and by projecting metadata hypotheses into pixel space, create interpretable images of the instantiated metadata (i.e., the meta-space). We then compute the representational similarity (or image-by-image distance) of the projected images instantiated in each meta-space to that of the natural image distances computed from neural features in a brain area.

To approach this question, we aimed to extract interpretable intermediate features between pixels and categories to evaluate their explanatory power for neural recordings from different visual cortices. By applying this analysis to the data recorded from the higher visual cortex of monkeys (Majaj, Hong, Solomon, & DiCarlo, 2015), we discovered that the image representation in the population of neurons across the ventral visual cortex transforms from retinotopically-centered in early areas to object-enhanced in posterior and central IT (not the particular retinal field, but the spatial region around the object) and finally to object-centered in anterior IT (only identity retained and not position).

Methods

Meta-Space Representational Analysis (MSRA) Does the neural activity of a certain region represent an interpretable feature of an image? If yes, how can we identify that feature? For example, a hypothesis could be that a target area represents the object in the image and ignores the background. To

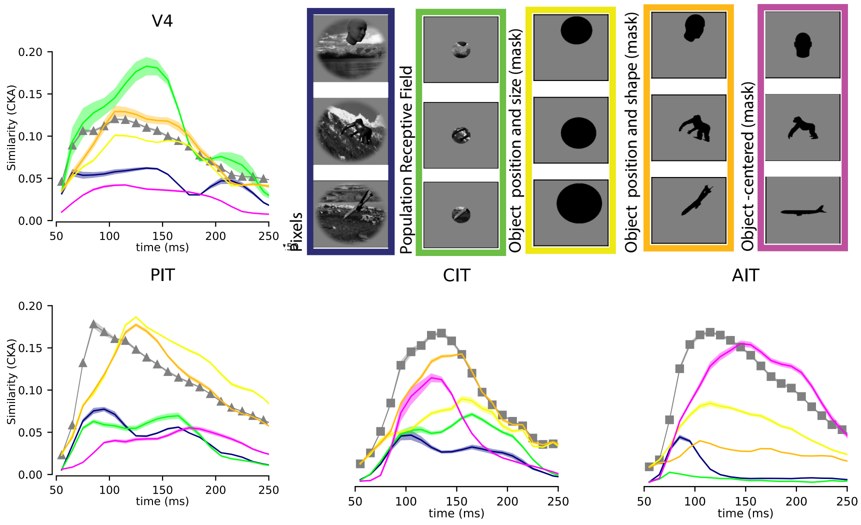

Figure 2: Identification of dynamics of neural activity in higher visual cortex. For example, a binary array of the same size as the image with 0s and 1s as values where a value of 1 indicates the corresponding pixel in the image belongs to the object (fourth column, orange line). For comparison the best layer of ResNet50 for each region was also depicted in gray, RN50 layer 3: gray triangles, RN50 layer4: gray squares. Error bar shows sem over the choice of neurons in each region. Stimulus was on between 0 to 100 ms, followed by a gray screen between 100 to 200ms.

test this hypothesis, classical vision studies construct a stimulus set of objects-only and another stimulus set of scene-only to investigate which brain area responds more vigorously to each class. Thus the number of hypotheses that can be tested is limited in each study. Moreover, this reduction of the natural image to its constituent parts to construct the stimuli assumes a form of superposition and deviates from studying vision in non-reduced, natural settings where nonlinear interactions are known to occur across constituent parts. On the other hand, DNNs can take in the same natural image an animal viewed as the input and produce a series of feature spaces, where each can be mapped to neural data recorded from the brain region of interest. The properties used in designing the network – training regime, architecture, and/or objective function – can yield an inference about the general properties of a brain region but lack a direct specific description at that level of neural features. Our method aims to reveal a set of semantically identifiable features (with a close tie to the underlying generative factors of a natural image) to explain the representations in the areas of the visual cortex. A natural image can be generated by a set of meta-parameters: an object (e.g. a face) undergoes transformations such as rotation and resizing and is then placed in a certain location overlaying a certain background. Our method turns a hypothesis (e.g. upon viewing a natural image, area X encodes object shape while ignoring the background) into a filter (object-only mask) applied to the natural image (Figure 1). We aim to reveal the meta-parameter that when projected into pixel space best explains the activity pattern in a certain visual area. To evaluate the similarity of each meta-parameter representation (i.e., meta-space), we compute the representational similarity between the neural population activity of a region and a meta-space (e.g. shapes), we call this method Meta-space Representation Analysis (MSRA). We show that these novel explicitly interpretable meta-spaces which can be instantiated simply by projecting meta-parameters onto pixel space have high quantitative explanatory power in revealing the features represented by neural activity of different regions.

Results

Identification of interpretable features in neural activity We applied MSRA on neural data recorded previously from V4, PIT, CIT, and AIT from two monkeys passively viewing 5760 natural images from 8 categories (Majaj et al., 2015; Yamins et al., 2014; Hong, Yamins, Majaj, & DiCarlo, 2016). We first derived meta-spaces such as shapes (object silhouette), size and position (area encircling the object), and object prototype silhouette from natural images in the provided dataset. Based on recording information, we also defined a meta-space called population Receptive Field (pRF). For example, the array on V4 covered 2 visual degrees on the fovea, hence we included a pRF of 2 degrees. We then measured the representational similarity by computing linear Centralized Kernel Analysis (CKA (Kornblith, Norouzi, Lee, & Hinton, 2019)) of each of these meta-spaces to the neural activity of each region at each time after the stimulus onset. To provide a reference for comparison, we also included the best account provided by the layers of ResNet50, a feedforward DNN (RN50, gray lines). We observed that pRF explains V4 data better than other meta-spaces and layer 3 of RN50 (Fig. 2). Across IT, from PIT to CIT to AIT, the explanatory power of pRF diminishes while increasing for object-related meta-spaces; object boundary, shape, and prototype provides better accounts for PIT, CIT, and AIT, respectively, especially for later part of the neural responses. Earlier parts of responses, however, are better explained by the DNN (RN50).

Conclusions

By identifying the representational content of each visual area we were able to bridge the modern image-computable mechanistic models of vision to the classical vision studies which were mainly focused on understanding what features excite individual neurons in a visual area.

Acknowledgments

We would like to thank DiCarlo lab at MIT for providing the neural data. T.T. is supported by NIH RF1DA056397, NSF 1707398, and Gatsby Charitable Foundation GAT3708. This work was supported by NIH R34 (NS116739), Klingenstein-Simons fellowship, Sloan Foundation fellowship, and Grossman-Kavli Scholar Award as well as a NVIDIA GPU grant and was performed using the Columbia Zuckerman Axon GPU cluster.

References

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016, April). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.*, *19*(4), 613–622.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, November). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, *10*(11), e1003915.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, May). Similarity of neural network representations revisited.

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

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, June). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(23), 8619–8624.