Assessing equivariance in visual neural representations

Akshay V Jagadeesh*, **Will Xiao***, and **Margaret S Livingstone**
Harvard Medical School, Boston, MA, USA
* these authors contributed equally

Abstract

Visual perception must balance two competing goals: invariance and sensitivity. One can recognize a bird, despite significant variation in its pose, color, or texture, yet can also describe those identity-orthogonal features. How do our brains achieve this balance? We test a theory that the brain learns an equivariant representation of objects, in which identity-preserving transformations, specifically 3D rotation, are encoded by a common, predictable transformation of the neural response. Using a stimulus set of 3D objects rendered from spherically-sampled viewpoints, we develop a metric to assess rotational equivariance by and test this on neural activity from primate inferior temporal (IT) cortex as well as features from Imagenet-trained deep neural networks. Although category and identity information are evident in IT cortical responses, evidence for rotation equivariance is weak. We find an optimal subspace of IT cortex that possesses more equivariance than would be expected by chance, but no more than in a deep neural network model or the pixel space of the images. Our results indicate that IT cortex lacks rotation-equivariant representations and suggest the need to explore other cortical systems downstream of IT that may serve as the basis for equivariant visual object perception.

Keywords: vision; object recognition; invariance; IT cortex; equivariance; rotation; deep learning.

Introduction

Many cognitive tasks require the ability to disentangle relevant information from covarying factors. For example, visual object recognition requires encoding identity information independent of identity-irrelevant dimensions, such as spatial position, rotation, or pose. Frequently, these nuisance covariates inherit intrinsic symmetries from the physical reality. Representations that incorporate these known underlying symmetries can not only learn more efficiently but also generalize at an abstract level. For example, if we know to represent the effect of in-plane rotations by an affine transformation, we can predict the appearance of objects for arbitrary rotation angles and novel objects.

In the context of deep learning, symmetry-aware representations help improve the data efficiency, with example approaches including convolutions for translation equivariance, data augmentation, and geometric deep learning (Bronstein, Bruna, LeCun, Szlam, & Vandergheynst, 2017; Olah, Cammarata, Voss, Schubert, & Goh, 2020).

In visual neuroscience, an influential symmetry-aware theory is invariant object recognition, which posits that neural representations underlying object recognition ignore identity-irrelevant transformations (DiCarlo & Cox, 2007). The ventral visual stream, particularly the inferior temporal (IT) cortex, is widely suggested to be the basis of invariant object recognition due to its selectivity for natural image features and encoding of identity-predictive features. However, perceptually, humans and other primates can also identify identity-irrelevant features such as rotation in addition to object identity, and neural responses in IT cortex also encode information about identity-orthogonal features, such as the position, pose, and size of objects (Hong, Yamins, Majaj, & DiCarlo, 2016). Moreover, humans can also generalize rotations to new object exemplars (Biederman & Bar, 1999), which is a difficult task for deep learning models optimized for invariant object recognition (O'Connell et al., 2023). Thus, rather than invariance, equivariant representations may provide a more symmetry-aware explanation that also accounts for transformation decoding and generalization.

We thus investigated the primate ventral visual stream as well as deep neural network (DNN) models of visual object recognition to assess the degree to which equivariance, specifically for 3-dimensional object rotation, characterizes their representational structure. Our findings provide evidence to suggest that although equivariance, rather than invariance, is a more complete theory of visual object perception, neither IT cortex nor leading Imagenet-trained deep neural network models contain representations sufficient for supporting equivariant object perception.

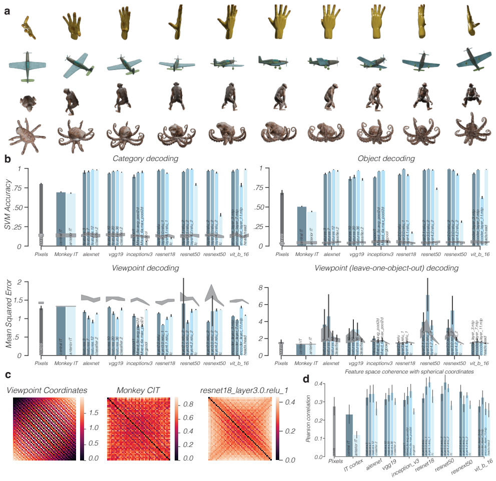

Figure 1: a, Examples of stimuli rendered from different viewpoints. b, Decoding accuracy. Gray lines denote permutation distribution. c, Representational dissimilarity matrices. d, RDM correlation to coordinate space.

Methods

Stimuli We selected 20 objects from the Objaverse-XL 3D object set. We sampled 50 camera positions that were (approximately) evenly spaced on a sphere and used Blender to render images of each object from each of these 50 positions, yielding a total of 1000 unique images 1.

Neural Data Collection We measured electrophysiological responses in IT cortex of one monkey while viewing images, presented for a duration of 150ms with a 150ms interstimulus interval, in serial presentation. The monkey was trained to maintain fixation centrally during image presentation, and fixation was monitored using an Eyelink eyetracking system. Each image was repeated 10 times within a recording session. We recorded from two chronically implanted electrode arrays in the left hemisphere, one floating microelectrode array in central IT and one Neuropixels probe in anterior IT.

Results

Decoding information

What information is encoded in the neural representation that can be accessed via a linear readout? We used a cross-validated linear SVM to predict the category of each image from neural responses and compared its performance to a permutation distribution from refitting the classifier with randomly shuffled labels. We found that category information could be accurately predicted substantially above chance-level from IT neural responses as well as DNN model responses (Fig. 1b). We similarly found accurate decoding of object identity. Finally, we used ridge regression to estimate the 3D viewpoint coordinates corresponding to each image. Using random cross-validation, we found that most DNN model features were able to predict viewpoint substantially above chance. However, when holding out one object and training on the remaining objects, we found that no model's features predicted viewpoint above chance, suggesting that although object-specific, but not object-general, viewpoint information is encoded in pretrained DNN features. This finding is consistent with prior literature suggesting that DNN models fail to generalize to out-of-class viewpoints (O'Connell et al., 2023).

Comparing similarity structures

To assess the degree to which neural and DNN representations capture the similarity structure between different viewpoints of the same object, we performed a representational similarity analysis. We first computed the distance between each pair of viewpoints in the 3D coordinate space (Fig. 1c, left). Then, for each object, we computed the distance between the neural response to each pair of viewpoints of that object and averaged over all objects to yield a representational dissimilarity matrix over viewpoints (Fig. 1c, middle, right). We then correlated these RDMs with the RDM generated from the viewpoint coordinate to measure the degree to which each representational space captured the relative distances between all viewpoints. To determine a baseline level of representational similarity, we also computed the representational dissimilarity in the pixel-space of the images. We found that most models performed no better than pixel-wise similarity.

Assessing the Presence of An Equivariant Subspace

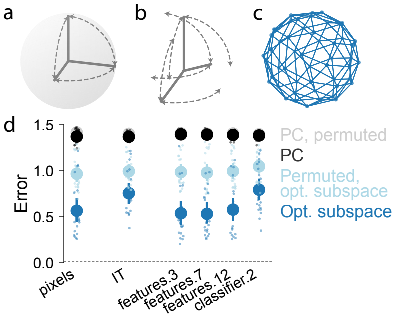

Figure 2: a,b, illustration of veridical (a) and possibly irregular (b) view representations. If a representation space is equivariant to the veridical space, the original pairwise relations (indicated by arrows) should pertain. c, illustration of the 50 camera views we sampled. d, View-to-view prediction error in various 3D subspaces (colors) of different image representations (x-axis) by applying veridical transformations (a,b). Small and large dots correspond to individual objects and object-averages.

One hypothesis of how a neuronal population can simultaneously represent information about object identity and rotation is to partition each type of information into orthogonal subspaces. Furthermore, if the rotation-specific subspace had the same geometry of the veridical transformations, it could support generalization to unseen views by extrapolating rotations from any view. In our stimulus set, the 50 object views were obtained from 50 camera locations placed on a sphere around the origin (Fig. 2c). The relation between each pair of views can be represented as a rotation matrix in this 3D latent space, and pairwise rotation matrices only depend on the relative positions (Figure 2a). We empirically tested whether the neural data contained a subspace equivariant to these view rotations. To do so, we optimized a projection matrix from the full neural state space (29 neurons) into a 3-dimensional latent space that aimed to preserve the pairwise linear transforms between views (Fig. 2b), by minimizing the prediction error between every pair of views while *freezing* the pairwise rotational matrices between positions in the latent 3D space (Fig. 2b). The error was calculated as the Euclidean distance between predicted and actual view vectors (averaged over 50 × 50 pairs; vectors were normalized to have an average scale of 1). Because any equivariant representation may, in general, be specific to an object, we separately optimized the projection matrix for each object.

We first established that the first 3 principal components of the neural state space did not contain equivariant representations. Applying the latent-space rotation matrices to the first 3 PCs, we could not predict the representation of one view from another better than chance (the same analysis after randomly permuting the 50 views). Meanwhile, the neural subspace optimized for equivariance showed some level of equivariance. Optimizing a 3D subspace (in the 29D neural representation space) lowered the error of predicting the representation of one view from another. This lower error is specific to the view structure, because view permutation led to higher view-to-view prediction even after optimization of the projection.

We conducted an analogous analysis in pixel space and four layers of AlexNet. To ensure meaningful comparisons, we used PCA to project each feature space to 29D, the same as the full neural state space. The amount of equivariance present in the optimal neural subspace is similar to that present in a late deep net layer (AlexNet 'classifier.2.' or pool 5) and lower than the equivariance in earlier AlexNet layers and in the pixels. This result is consistent with IT emphasizing object identity information and reducing view representations, perhaps by overrepresenting canonical views of an objects (e.g., a person viewed from the front rather than the back, top, or bottom).

Summary

Our results indicate that IT cortex is unlikely to be the basis of equivariant object recognition. This is in line with prior results suggesting that IT cortex encodes local complex visual features, rather than global 3D shape (Bonnen, Yamins, & Wagner, 2021; Jagadeesh & Gardner, 2022; Waidmann, Koyano, Hong, Russ, & Leopold, 2022; Xiao, Sharma, Kreiman, & Livingstone, 2023). We conclude by suggesting that equivariant, not invariant, object recognition ought to be the target of computational modeling efforts, therefore emphasizing the necessity of exploring other cortical regions, perhaps downstream of IT cortex, that might provide sufficient representations to support equivariant object recognition.

References

Biederman, I., & Bar, M. (1999). One-shot viewpoint invariance in matching novel objects. *Vision research*, *39*(17), 2885–2899.

Bonnen, T., Yamins, D. L., & Wagner, A. D. (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, *109*(17), 2755–2766.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, *11*(8), 333–341.

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

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, *119*(17), e2115302119.

O'Connell, T. P., Bonnen, T., Friedman, Y., Tewari, A., Tenenbaum, J. B., Sitzmann, V., & Kanwisher, N. (2023). Approaching human 3d shape perception with neurally mappable models. *arXiv preprint arXiv:2308.11300*.

Olah, C., Cammarata, N., Voss, C., Schubert, L., & Goh, G. (2020). Naturally occurring equivariance in neural networks. *Distill*, *5*(12), e00024–004.

Waidmann, E. N., Koyano, K. W., Hong, J. J., Russ, B. E., & Leopold, D. A. (2022). Local features drive identity responses in macaque anterior face patches. *Nature Communications*, *13*(1), 5592.

Xiao, W., Sharma, S., Kreiman, G., & Livingstone, M. S. (2023). Out of sight, out of mind: Responses in primate ventral visual cortex track individual fixations during natural vision. *bioRxiv*, 2023–02.