Journal Club

Editor's Note: These short, critical reviews of recent papers in the *Journal*, written exclusively by graduate students or postdoctoral fellows, are intended to summarize the important findings of the paper and provide additional insight and commentary. For more information on the format and purpose of the Journal Club, please see http://www.jneurosci.org/misc/ifa_features.shtml.

Unwrapping the Ventral Stream

Jeremy Freeman and Corey M. Ziemba
Center for Neural Science, New York University, New York, New York 10003

Review of Rust and DiCarlo

It seems immediately obvious that the apple in my hand is the same apple I just picked up from the table. How can my brain recognize an object despite substantial variation in its location, size, and context? In a series of cortical areas known as the ventral stream, the brain performs sophisticated computations that allow us to select things we need to know for object identity (selectivity) while ignoring variations that do not matter (invariance). The brain's capacity to achieve both selectivity and invariance simultaneously remains a fundamental mystery in systems neuroscience.

We understand selectivity and invariance early in the ventral stream better than in its later stages. Cells in primary visual cortex (V1) show selectively for simple image features like orientation and spatial frequency, and some complex cells in V1 also show invariance, or insensitivity, to position. Because V1 is closest to the input, this simple (quasilinear) filtering of the input successfully describes the function of V1 cells, and cells can be well characterized by measuring responses to random white noise stimuli (Rust et al., 2005; Chen et al., 2007).

Further along the ventral stream, through areas V2, V4, and inferior temporal cortex (IT), neural populations gradually combine the simple features from V1 to represent more complicated patterns and objects. In the process, cells become selective for increasingly specific image features and invariant to increasingly complex transformations. Consequently, linking any particular cell's response to the input becomes more difficult. How should experimenters choose stimuli to characterize these cells? Stimuli used to study V1 (white noise) are insufficient because they rarely contain the special features that drive these cells to respond—a million white noise images will rarely include an apple. So-called natural images more likely contain such features. But the number of available natural images is nearly infinite, and even when a cell responds to a particular image, the feature that triggered the response may be unknowable because our capacity for modeling the structure of natural images remains incomplete (Simoncelli and Olshausen, 2001).

Rather than search for the features that drive each individual cell's particular response, we can instead assess the response of a population of cells, and replace a barrage of random natural images with a set of well defined image manipulations. The geometrical intuition behind this approach is to imagine a population of cells as carving up the set of all possible images into subsets, or manifolds (DiCarlo and Cox, 2007). Each manifold might, for example, correspond to images of some particular object, but contain many possible variations of that object (Fig. 1). A population of cells can discriminate among objects if objects from different manifolds consistently yield distinguishable patterns of neural responses. Starting with a natural object image, we can characterize a manifold by manipulating the image and asking whether the manipulated image is or is not still on the manifold. To operationalize "still on the manifold," we determine whether a population of cells can still discriminate an image from other images despite the manipulations. Manipulations that impair performance take us off the manifold, identifying directions of selectivity. Manipulations that do not affect performance keep us on the manifold, identifying directions of invariance. Rust and DiCarlo (2010) used this approach with impressive rigor to characterize both selectivity and invariance in areas V4 and IT.

To examine selectivity, Rust and DiCarlo (2010) presented scrambled versions of natural images. Scrambling preserved many low-level features, but disrupted complex feature conjunctions and destroyed object identity. The IT population robustly discriminated among different intact images, but poorly discriminated among scrambled images, demonstrating that scrambling disrupted features that IT neurons used to discriminate images. In contrast, the V4 population was better able to discriminate among the scrambled images, suggesting that neurons in V4 tend to encode features that were unaffected by scrambling.

To examine invariance, Rust and DiCarlo (2010) analyzed population responses to complementary transformations that preserved object identity, such as translation, scaling, and added background context. IT was reliably able to decode object

Received Nov. 27, 2010; revised Dec. 11, 2010; accepted Dec. 14, 2010.

J.F. and C.M.Z. are supported by National Science Foundation Graduate Research Fellowships. We thank Elisha Merriam and Michael Landy for helpful comments on earlier drafts.

Correspondence should be addressed to Jeremy Freeman, Center for Neural Science, New York University, 4 Washington Place, New York, New York, 10003. E-mail: freeman@cns.nyu.edu.

DOI:10.1523/JNEUROSCI.6191-10.2011

identity across a range of transformations, whereas the V4 population performed poorly. Although demonstrations of complex feature selectivity and invariant representations in V4 and IT already exist (Pasupathy and Connor, 2002; Zoccolan et al., 2007), these results provide the first systematic comparison of matched populations across the two different visual areas, documenting an increase in both selectivity and invariance.

What is the relationship between the properties of single cells and the capacity for selectivity and invariance at the population level? In the case of selectivity, Rust and DiCarlo (2010) focused on the property of receptive field size; they identified subpopulations of cells with specific receptive field sizes and measured the discrimination performance of these subpopulations. In both areas, populations of cells with bigger receptive fields were more sensitive to scrambling, probably because these cells encode large, spatially extended image structures that scrambling disrupts. Although the IT population as a whole was more sensitive to scrambling than the V4 population, subpopulations of V4 and IT cells with similarly sized receptive fields were similarly affected by scrambling, suggesting that those V4 and IT cells were in fact selective for similar kinds of image structures. Does this result then contradict the traditional view that selectivity qualitatively changes across visual areas? It may in fact only reflect a limitation of the scrambling manipulation. Much information is lost through scrambling; sensitivity to scrambling shows that cells use some or all of that information to discriminate intact images, but it cannot identify exactly what information cells use.

Rust and DiCarlo (2010) also examined single-cell properties that were important for invariance. One particularly interesting property they identified is the degree to which an individual cell's preference for some objects over others is maintained across various transformations; i.e., whether a cell's object preference is independent of how the object is presented. For example, if a cell prefers apples to oranges when the fruit is presented on the left, does it also prefer apples to oranges when the fruit is on the right? At the single-cell level, stability of object preference across transformations was more common in IT than V4, and this difference contributed to the IT population's ability to decode object identity across a range of transformations.

The stability of object preference across transformations is clearly crucial for invariant object recognition. Using preference stability to characterize a cell, however, is more like building a lookup table than a functional model. A lookup table catalogs the response to each of several specific stimuli, whereas a functional model operates directly on images and predicts the response to an arbitrary input. For a cell in V1, we could of course construct a lookup table by measuring the response to different orientations and positions, but we already have the more powerful functional model, which captures the response to these, and many other, stimuli. However, building functional models for areas deep in the ventral stream, far from the input, is challenging. Although the lookup table approach allows the demonstration of clear differences between areas, we should not think of the lookup table itself as a characterization of how a cell encodes the input, but rather a catalog of the cell's capabilities, and thus a guide to formulating proper functional models.

Like many studies of IT, this study used stimuli and manipulations with obvious behavioral relevance. An object does not look the same when it is scrambled, but does when it changes in location, size, or context. Object recognition is clearly a goal of the visual system, and it is likely to be realized at the top of the ventral stream, closest to areas involved in memory. Thus, it makes sense to use our intuitions about behavioral relevance when picking stimuli, but the reliance on such intuitions renders the conclusions somewhat IT-centric. The present paper demonstrates a robust difference between V4 and IT, but we primarily learn about things that IT can do and V4 can not, and learn less about the specific capabilities of V4.

How then can we probe areas where our intuition founders? Two related approaches are worth considering. One is to study a given stage of the ventral stream by using our knowledge of the previous stage. If we have a functional model for

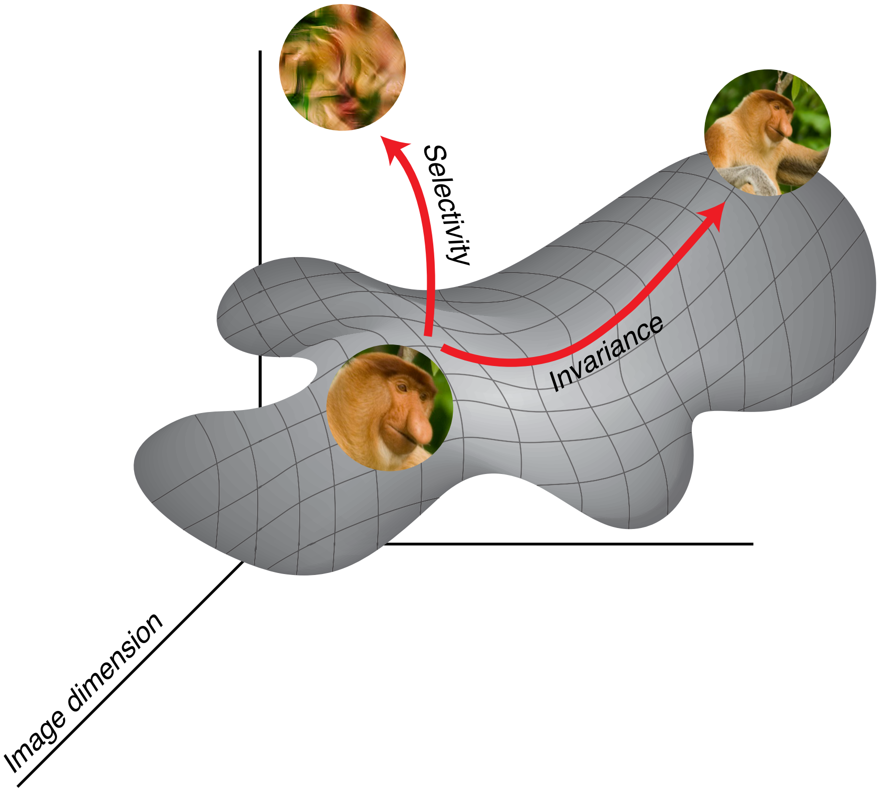

Figure 1. The geometry of selectivity and invariance. The three axes are three image dimensions (e.g., the values of three pixels in an image). Real images require several thousand dimensions, but we use three for simple visualization. Any point in the space corresponds to a different image. The gray surface represents a continuous subset, or manifold, of images of a particular object. If a hypothetical neural population effectively encodes this object's identity, all object images from this manifold will yield patterns of neural responses that are distinguishable from the patterns of responses induced by other sets of images. Moving along the surface of the manifold changes the image itself but maintains the ability of the neural population to discriminate the image from others. This is a direction of invariance. Moving away from, or orthogonal to, the surface of the manifold changes the image in a way that prevents the population from effectively discriminating. This is a direction of selectivity. The manifold shown here corresponds to a set of population responses that are selective for proboscis monkeys, not just for image patches with similar color and texture, but are also invariant to changes in size (near vs far) and context (face only vs face and body). Photo used with permission, courtesy of Paul Huggins (http://www.paulhugginsphotography.com/).

what cells in a given area represent, we can characterize a cell in a hierarchically higher area by showing stimuli that span the space represented by the cells in the lower area. Given our current knowledge, this technique may work well in areas immediately after V1 (Rust et al., 2006; Willmore et al., 2010), but not beyond. A second, more arbitrary approach is to construct a parameterized set of features, such as angles for V2 or shape and curvature for V4, and characterize how cells respond to their systematic variation (Pasupathy and Connor, 2002; Ito and Komatsu, 2004). No normative theory, however, constrains the choice of features from among the vast array of possibilities. A final approach, yet to be fully explored, is to study statistical properties of natural images and to use normative theories, like efficient coding, to identify stimulus dimensions that the midlevel visual system ought to encode (Karklin and Lewicki, 2009). As our understanding of the structure of natural images improves, theory, rather than intuition, can help choose the stimuli we use to investigate the murky middle of the ventral stream.

References

Chen X, Han F, Poo MM, Dan Y (2007) Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). Proc Natl Acad Sci U S A 104:19120–19125.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11: 333–341.

Ito M, Komatsu H (2004) Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. J Neurosci 24:3313–3324.

Karklin Y, Lewicki MS (2009) Emergence of complex cell properties by learning to generalize in natural scenes. Nature 457:83–86.

Pasupathy A, Connor CE (2002) Population coding of shape in area V4. Nat Neurosci 5:1332–1338.

Rust NC, DiCarlo JJ (2010) Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area V4 to IT. J Neurosci 30:12978–12995.

Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005) Spatiotemporal elements of macaque V1 receptive fields. Neuron 46:945–956.

Rust NC, Mante V, Simoncelli EP, Movshon JA (2006) How MT cells analyze the motion of visual patterns. Nat Neurosci 9:1421–1431.

Simoncelli EP, Olshausen BA (2001) Natural image statistics and image representation. Annu Rev Neurosci 24:1193–1216.

Willmore BD, Prenger RJ, Gallant JL (2010) Neural representation of natural images in visual area V2. J Neurosci 30:2102–2114.

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ (2007) Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci 27:12292–12307.