Dispatch

Object Recognition: Complexity of Recognition Strategies

Philippe G. Schyns

Primate brains and state-of-the-art convolutional neural networks can recognize many faces, objects and scenes, though how they do so is often mysterious. New research unveils some of the mystery. revealing unexpected complexity in the recognition strategies of rodents.

Object recognition is a pervasive process that fascinates and puzzles in equal measure. It fascinates because, within a few hundred milliseconds, the human brain will recognize its own face in the mirror, its sex, age and emotional expression, but also mundane objects such as a favorite coffee mug, pair of shoes and car, or more complex everyday scenes such as an office, a neighborhood or city. And this will occur in bright daylight, on a dark rainy day, or from close up or far away. Similarly, deep convolutional neural networks, an influential development in computer vision, can now visually recognize objects with a performance level that even exceeds that of humans. So, recognition fascinates because wet and silicon brains perform apparently effortless recognition under a wide variety of circumstances. But recognition continues to puzzle because we still do not understand how it works. In this issue of *Current Biology*, Djurdjevic *et al*. [1] address the challenge by mathematically modelling complex recognition strategies in rodents.

How should we approach recognition to understand how it works? A first important aspect to consider is the recognition task itself. Suppose a biological or artificial brain must discriminate an orange from a pear. In principle, their respective shapes (round *versus* not), colors (orange or green) and textures (orange peel or grainy) could each be diagnostic. That is, the system could focus on any of these features to correctly discriminate oranges and pears. But if the task changed to discriminating an orange from a peach, shape would be less useful than color and texture. Similar examples can easily be conjured up for more complex face, object and scene discriminations. The key point is that recognition tasks are thought to compare stimulus information with memorized information to accomplish recognition behavior. We are unlikely to understand how recognition works in biological and artificial brains if we do not understand what information the system minimally processes in the task — that is, the diagnostic information [2].

Diagnostic information is not a novel idea — it can be traced back to Lorenz and Tinbergen’s [3,4] seminal studies in ethology — but it is often neglected, even in state-of-the-art recognition studies. One reason is that diagnostic information can be structurally complex — for example, think about the diagnostic information of an art-deco building — and few techniques exist to reveal hidden information structures from two-dimensional images. A notable exception is the so-called reverse correlation approach developed in psychophysics ([5], for review see [6]). In reverse correlation, a small amount of noise can be added to each pixel of an image — for example, representing a face [7], an object or a scene [8], or nothing at all [9] — or noise can be multiplied with an image to mask its information contents [10]. In both cases, additive or multiplicative noise introduces random variations in the input image. These variations can facilitate or hinder the recognition task under study. After many such experimental trials, two stacks emerge: the random variations that facilitated the recognition task and those that hindered the task. For each image pixel, we can easily compute — for example, via linear regression, correlation [6] or mutual information [11] — whether the random variations produced by added noise or multiplicative masking affected recognition performance.

The outcome is a classification image — also called a ‘decision template’ — that summarizes the groups of pixels (the image features) that facilitated or hindered recognition. Such pixel-based reverse correlation techniques are powerful because they are data-driven and so can agnostically reveal the information structures diagnostic of the task. The real-world is three-dimensional, however, and pixel-based techniques operate on the two-dimensional projections of three-dimensional faces, objects and scenes — in other words, on two-dimensional images. As such, classification images are not tolerant to common changes in the three-dimensional world: rotation of objects in the plane, in depth, translations, changes of size and so forth.

To address these issues, one can develop a generative model of the visual information — for example, a three-dimensional face [12], object [13] or scene [14] — and introduce random variations on the generative parameters of the model — such as random movements of facial muscles, random changes of the limb sizes of an animal, random changes of building height — and measure their effects on task performance. Such generative approaches have two main advantages. First, the ‘classification mode’ could require estimation of fewer parameters in the three-dimensional generative space than the total number of pixels that must be estimated in a two-dimensional classification image. Second, the classification model could be tolerant to three-dimensional changes in ways that a two-dimensional classification image simply cannot be. There is, however, a significant trade-off to these gains. In the three-dimensional generative space, each parameter is an explicit hypothesis on the diagnostic information that might be important for the task — for example, movements of facial muscles, limb sizes of an animal or building heights. With each hypothesis, we are eroding the agnostic stance of the two-dimensional pixel-based approach, where novel, unanticipated diagnostic features can emerge in the classification image, in a data-driven manner. This can occur because added pixel noise can in principle represent any image feature. In the three-dimensional generative space, there will be no such surprises. Diagnostic features can only lie within the range of parametric generation [15].

Djurdjevic *et al*. [1] trained rats to discriminate a target tripod stimulus (three lobes, Y-shaped) from 11 distractors with a broad range of image-based similarities with the target. Though all animals learned the task, their discrimination performance varied considerably. To understand how the rats discriminated, a hybrid reverse correlation technique used in a testing phase estimated their perceptual strategies — the features each rat used in the task. To this end, a three-dimensional generative model produced random variations in the orientation, size and aspect ratio of the three-dimensional geometrical primitives making up the tripod. The authors computed, for each rat, a pixel-based two-dimensional classification image to infer the features that facilitated and hindered the rat's discrimination performance. They found that good performers had a more complex perceptual strategy than poor performers: their classification images comprised an additional feature that enabled more effective rejections of certain distractors. In interesting validations, the authors first used the classification image of each rat as a perceptual filter applied onto each distractor to successfully predict its discriminability from the tripod target. Then they incorporated the rat's individual classification image into a decision model based on logistic regression to successfully predict discrimination performance on an independent data set.

Why are these results important? Recognition is implicitly cast as brain processes that operate on information. If we do not know explicitly, with a model, the information contents that wet and silicon brains process in each task, how can we understand and explain where, when and how specific circuits process these contents to produce behavior? Here, Djurdjevic *et al*. [1] modelled the information contents of a task and from its complexity predicted discrimination performance and modelled decision behavior. They demonstrated an unexpected level of complexity in rodents, including tolerance to changes in size and contours, that challenges the notion that they are not capable of advanced shape processing (see also [16] and [17] for a related approach applied to primates). Of course, our understanding of the diagnostic information in recognition tasks will remain bound by the generative models of visual information that we can imagine — or uncover from deep convolutional neural networks, which will also require some mathematical imagination. Generative models of visual information could be the next frontier to produce information processing models of the brain and behavior.

References

1. Djurdjevic, V., Ansuini, A, Bertolini, D. Macke, J.H. and Zoccolan, D. (2018). Accuracy of rats in discriminating visual objects is explained by the complexity of their visual strategies. Curr. Biol. *28,* xxx-xxx.

2. Schyns, P.G., Gosselin, F., and Smith, M.L. (2009). Information processing algorithms in the brain. Trends Cogn. Sci. *13,* 20–26.

3. Lorenz, K. (1970). Studies in animal behavior, (Cambridge, MA: Harvard University Press).

4. Tinbergen, N. (1951). The study of instinct, (New York: Oxford University Press).

5. Ahumada, A.J. and Lovell, J. (1971). Stimulus features in signal detection. J. Accoustical Soc. Am. *49*, 1751-1756.

6. Murray, R.F. (2011). Classification images: A review. J. Vision, *11,* 1–25.

7. Mangini, M.C., and Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. Cogn. Sci. *28*, 209–226.

8. Neri, P. (2014). Semantic control of feature extraction from natural scenes. J. Neurosci. *34*, 2374-2388.

9. Gosselin, F., and Schyns, P.G. (2003). Superstitious perceptions reveal properties of internal representations. Psychol. Sci. *14*, 505–509.

10. Gosselin, F., and Schyns, P.G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res. *41*, 2261–2271.

11. Ince, R.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., and Schyns, P.G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. Hum. Brain Mapp. *38*, 1541–1573.

12. Jack, R.E., Garrod, O.G., Yu, H., Caldara, R., and Schyns, P.G. (2012). Facial expressions of emotion are not culturally universal. Proc. Natl. Acad. Sci. USA *109*, 7241–7244.

13. Olman, C., and Kersten, D. (2004). Classification objects, ideal observers and generative models. Cogn. Sci. *28*, 227–239.

14. Greene, M.R., Botros, A.P., Beck, D.M., and Fei-Fei, L. (2014). Visual noise from natural scene statistics reveals human scene category representations. arXiv Preprint arXiv:1411.5331.

15. Jack, R.E. and Schyns, P.G. (2017). Towards a Social Psychophysics of Face Communication. Annu. Rev. Psychol. *68*, 269-297.

16. Vermaercke, B. and Op de Beeck, H.P. (2011). A multivariate approach reveals the behavioral templates underlying visual discrimination in rats. Curr. Biol. *22*, 50-55.

17. Popivanov, I.D., Schyns, P.G. and Vogels, R. (2016). Stimulus features coded by single neurons of a macaque body category selective patch. Proc. Natl. Acad. Sci. USA *113*, E2450-E2459.

Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. E-mail: philippe.schyns@glasgow.ac.uk

In Brief:

Primate brains and state-of-the-art convolutional neural networks can recognize many faces, objects and scenes, though how they do so is often mysterious. New research unveils some of the mystery, revealing unexpected complexity in the recognition strategies of rodents.