Predicting the visual world: silence is golden

Christof Koch and Tomaso Poggio

In predictive coding, only unexpected input features are signaled to the next stage of processing. Rao and Ballard use this approach to model extra-classical receptive field effects.

In Arthur Conan Doyle's story *Silver Blaze*, Inspector Gregory from Scotland Yard asks Sherlock Holmes "Is there any point to which you would wish to draw my attention?", to which Holmes replies "To the curious incident of the dog in the night-time." "The dog did nothing in the night-time." "That was the curious incident." The point of this well known exchange is, of course, that the dog did not bark because the criminal was a person it knew well and expected. Predictive coding of sensory stimuli in nervous systems has a similar flavor. It is an encoding strategy in which predictable features in the input are suppressed, and only the unexpected is signaled to the next stage of processing.

Almost from the start of the information age, theorists have argued that such a coding strategy is very efficient and is likely to be widely used in natural sensory systems. In particular, predictive coding has been invoked to explain the detailed shape of the spatio-temporal receptive fields of neurons in the retina and lateral geniculate nucleus of mammals (for review, see ref. 1). The visual cortex, however, with its complex and sometimes highly nonlinear receptive-field properties, has proven resistant to these ideas. No more, though. The study by Rao and Ballard[2] in this issue of *Nature Neuroscience* provides a detailed account of how predictive coding can explain extra-classical receptive-field effects of neurons in primary visual cortex (V1) and beyond.
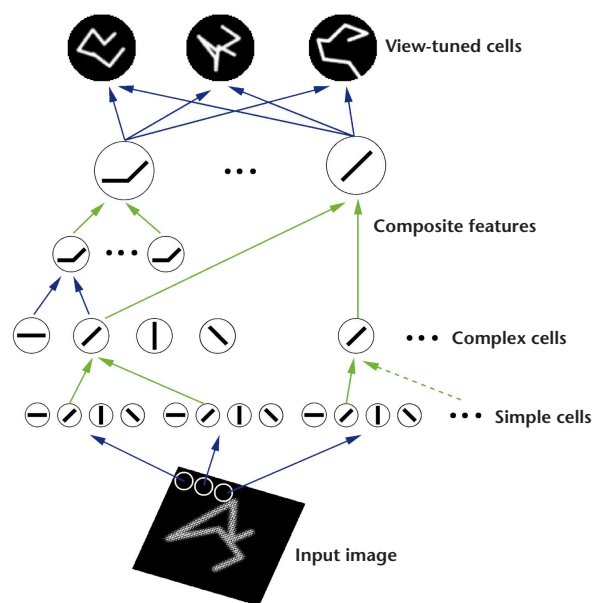
The basic idea behind hierarchical predictive coding is simple. It starts with the question of how images of natural scenes should be analyzed. An obvious answer is

Christof Koch is at the Computation and Neural Systems Program, California Institute of Technology, Pasadena, California 91125, USA. Tomaso Poggio is at CBCL in the Brain Sciences Department and in the AI Lab at MIT, Cambridge, Massachusetts 02142, USA. email: koch@klab.caltech.edu and tp@ai.mit.edu

in terms of frequently occurring features, such as blobs of opposing polarity, oriented edges, curved line segments and so on. A significant fraction of all images can be decomposed in terms of these more elementary features. In other words, a clever visual system that grew up in the natural world would not be surprised to find these features in real images. What is informative, however, are deviations from the norm, and in predictive coding, only these unexpected features are signaled to the next stage. Applied iteratively, this strategy leads to the following hierarchical network (see Fig. 1 in Rao and Ballard): at each stage of processing, the input from the previous stage is analyzed in terms of certain learned features. Each stage signals to the next the difference between these expected image features and the actual image, and each stage sends back to the one below it the

expected, or predicted, image features. Due to the convergence of spatially adjacent modules onto the next higher level, the receptive fields become progressively larger as one ascends the hierarchy. Learning of the synaptic weights of the units in each stage—which can be thought of as their receptive fields—is unsupervised, that is, it does not require an external 'teacher' that rewards or punishes individual choices.

Using a two-stage prediction network, Rao and Ballard demonstrate how units in their first stage come to show 'end-stopping' (a defining feature of the 'hypercomplex' cells of Hubel and Wiesel). Such a cell responds strongly to an appropriately oriented bar of a certain length. As the length of the bar is increased—extending into an area surrounding the cell's classical receptive field—the response decreases until the cell ceases to fire. End-stopping in these predictive-coding units results from the character of the natural images used in training. As Rao and Ballard show directly, their training images (natural scenes inhabited by animals, trees, rocks and so on) contain edges at different orientations but at a preferred scale; short bars are much less likely to occur than longer ones. During training, units in the second stage of the system come to expect such elongated edges and signal this to the lower stage. The firing rate of first-

Fig. 1. A purely feedforward model[10] can account for the available quantitative data on view-tuned inferotemporal cells. The model is an hierarchical extension of the classical Hubel and Wiesel approach of building complex cells from simple cells. It is the first demonstration that hierarchical schemes of biological neurons can account quantitatively for both the physiology and the psychophysics of high-level object recognition. The model consists of sequences of layers with linear (blue arrows) and nonlinear operations (green arrows), similar to logical AND and OR gates. These two types of operations respectively provide pattern specificity and transformation invariance. The nonlinear MAX operation, similar to a winner-take-all over all inputs of the cell, is key to the model's properties and is quite different from the basically linear summation of inputs usually assumed for complex cells. (Riesenhuber and Poggio, unpublished).

news and views

stage neurons is the difference between this expectation and the actual image. As the bar becomes long enough to intrude into the cell's extra-classical receptive field, the retinal stimulus agrees with the expected stimulus, and the cell, like Arthur Conan Doyle's dog, remains silent. The authors discuss how predictive coding can explain context-dependent effects found physiologically when stimuli extend into the extra-classical receptive field of V1 cells.

Note that the system adapts to the statistical properties of its training set. Change these, for instance by exposing it to nothing but random dot patterns, and end-stopping should disappear (because such patterns do not contain edges at a preferred scale). Ultimately, according to predictive coding theory, the existence of end-stopped cells in cortex reflects the way the visual world is structured (in which longer edges are more common than shorter ones).

Predictive coding is a general framework for interpreting information processing in complex natural and artificial systems, and many mechanisms may be seen in this light. For instance, lateral inhibition very early in the visual system can be understood as predicting deviation from uniformity. Lateral inhibition between motion detectors (or between cells pooling motion information across the retina and individual motion detectors) has been postulated to explain the fly's and quite possibly the cortex's detection of relative motion[3]. In these and other cases, although predictive coding has not been explicitly used, the resulting models can be framed in such terms.

The architecture of Rao and Ballard's neural network, with strong feedforward and feedback connections, is very reminiscent of the widespread reciprocal connections between cortex and thalamus and between cortical areas (see also ref. 4). Their model predicts that feedback pathways (whether within or between cortical areas) are critical to the normal functioning of the system. Deprived of this feedback, and therefore of predictions from the higher level, cells respond promiscuously and lose much of their selectivity, in agreement with experimental data.

In predictive coding, the commonplace view of sensory neurons as detecting certain 'trigger' or 'preferred' features

is turned upside down in favor of a representation of objects by the absence of firing activity. This appears to be at odds with single-neuron data indicating that neurons along the ventral pathway in the macaque monkey, extending from V1 to inferior temporal cortex, respond with vigorous activity to ever more complex objects, including individual faces or paperclips twisted in just the right way and seen from a particular viewpoint[5]. Training the monkey over weeks and months usually increases the incidence of neurons with highly specific receptive fields[6], rather than decreasing their number as would be expected if cortex were implementing a predictive coding strategy (because the system would come to expect these images). In addition, what about all of the functional imaging data from humans revealing that particular cortical areas respond to specific image classes, such as faces or three-dimensional spatial layout[7]? Is it possible that this activity is dominated by the firing of inhibitory feedback cells actively expressing an error signal, a discrepancy between the input expected by this brain area and the actual image? Alternatively, might predictive coding apply primarily to the context-dependent effects found in the extra-classical surround of neurons?

More familiar are purely feedforward models in which individual units respond to specific features in the image with elevated activity[8–10]. This framework easily accommodates a hierarchy of levels containing units of ever-increasing complexity, starting with oriented simple cells in V1 and ultimately leading to face cells in IT or place cells in the hippocampus. Appropriate nonlinear operations at each processing stage assure that the resulting units show the observed invariance to object size, rotation in depth and retinal location (**Fig. 1**). In principle, feedforward and predictive-coding feedback networks should respond in a qualitatively similar manner to select features, although the former are likely to react much more rapidly when confronted with complex images because they require no multiple iterations to converge. This is compatible with human data showing that frontal cortex can analyze complex scenes within an amazingly short 150 milliseconds[11].

The function of top-down connections within a primarily feedforward architec-

ture could be to modulate responses, as in selective attention or imagery. Such a modulatory role of feedback connections in the adult is compatible with the absence of any strong loops between two or more cortical or corticothalamic areas[12]. Another possibility is that feedback connections are necessary to enable the system to learn the appropriate feature vectors during development or learning, even though it might not need such feedback otherwise.

It will be critical to unravel the precise function of corticocortical feedback projections and their biophysical mode of operation, whether linearly subtractive as in Rao and Ballard's model or more modulatory multiplicative (or divisive)[13]. This most likely awaits new pharmacological or molecular tools that can delicately, deliberately and transiently inactivate selected feedback pathways.

In summary, it seems clear that predictive coding can be used to interpret certain features of cortical cells, in particular context-dependent responses. To what extent such an information coding strategy might extend to feature selectivity remains unclear; the brain may use multiple coding strategies for different tasks. As usual, only more experiments, guided by the sort of insights provided by Rao and Ballard, will help unravel the complexities and multiple facets of information processing in the brain.

1. Atick, J. J. *Network* **3**, 213–251 (1992).
2. Rao, R. P. N. & Ballard, D. H. *Nature Neurosci.* **2**, 79–87 (1999).
3. Poggio, T., Reichardt, W. & Hausen, K. *Naturwissenschaften* **68**, 443–446 (1981).
4. Rao, R. P. N. & Ballard, D. H. *Neural Comput.* **9**, 721–763 (1997).
5. Logothetis, N. K, Pauls, J. & Poggio, T. *Curr. Biol.* **5**, 552–563 (1995).
6. Kobatake, E., Wang, G. & Tanaka, K. *J. Neurophysiol.* **80**, 324–330 (1998).
7. Epstein, R. & Kanwisher, N. *Nature* **392**, 598–601 (1998).
8. Fukushima, K. *Biol. Cybern.* **36**, 193–202 (1980).
9. Perret, D. & Oram, M. *Image Vision Comput.* **11**, 317–333 (1993).
10. Riesenhuber, M. & Poggio, T. *AI Memo* **1629**, CBCL Paper 160 (1998).
11. Thorpe, S., Fize, D. & Marlot C. *Nature* **381**, 520–522 (1996).
12. Crick, F. & Koch, C. *Nature* **391**, 245–249 (1998).
13. Koch, C. *Biophysics of Computation* (Oxford Univ. Press, New York, 1998).