17 Object and Scene Perception

OWEN LEWIS AND TOMASO POGGIO

17.1 Introduction

The collection of questions humans can answer about the visual world is vast and diverse. Some can be answered at a glance: "Is this a picture of a car?" "Was this picture taken outdoors?" Others require more cognitive computation, more viewing time, and possibly links to nonvisual cognitive systems: "Is everyone is this picture looking at the camera?" "Are the people in this image friends?" "What will happen next in this scene?"

These two classes of questions are hypothesized to be handled by two different processing modes in the brain. As we will see, the brain's visual areas are organized into a hierarchy or chain, with each area passing information to its immediate neighbors. Fast visual processing is thought to occur in a single feedforward pass through this chain, with information moving exclusively from lower areas to higher ones. More extended processing, by contrast, brings feedback connections into play, allowing information to circulate among the visual areas in loops. After reviewing the anatomy of visual cortex in section 17.2, the organization of this chapter follows the feedforward/feedback distinction, with sections 17.3 and 17.4 focusing on feedforward processing, and section 17.5, on feedback.

The particular focus of our feedforward discussion is classification problems. In object classification, a model assigns to an image the category label of its principal object, for example, "dog" or "car." In scene classification, the label identifies the picture's larger-scale gist: Is it an indoor or outdoor scene, a natural scene or an urban one? Classification is an important building block for complex visual tasks and is a rich and difficult problem in its own right: only very recently have models emerged that can begin to account for human recognition performance, and even today, important open problems remain.

Classification has also been a major focus of recent computer vision research, and the interaction between this field and computational neuroscience has been fruitful for both. Indeed, the currently dominant class of computer vision models, convolutional neural networks (CNNs), is directly inspired by the architecture of visual cortex. Partly due to this computer vision connection, many neural models of feedforward visual processing can handle difficult classification problems in large image data sets, an important engineering step that at the same time has the scientific advantage of facilitating direct comparison with humans' recognition performance.

The feedback models we present aim to go beyond classification, allowing all visual areas to interact recurrently, blending the different sorts of information extracted by each to form the kind of complete scene interpretation necessary to support a full range of visual tasks. This full scene interpretation is a more daunting problem than the classification, and our understanding of how it is solved in the brain is not yet as fully developed. While full scene interpretation in natural images remains beyond the grasp of the neural models we present, these models can tackle important subproblems, and they account for a range of physiological and imaging experimental results.

17.2 Visual Cortex

Humans' and animals' ability to process objects and scenes is supported by a large swathe of specialized neural machinery located in visual cortex, near the back of the brain (see figure 12.11 in chapter 12 of this volume). In this section, we lay out the functional and anatomical properties of visual cortex, giving a set of desiderata and structural constraints that will shape the models we develop in the rest of the chapter. This section's core message is that the visual cortex forms a hierarchy of interconnected areas, and that useful representations are built up sequentially, with low-level areas contributing the building blocks to the more high-level and semantically meaningful representations found further upstream.

Visual cortex is traditionally divided into two processing pathways: the dorsal, or "where," pathway, thought to represent the location and motion of objects and to support interaction with visual input via eye and body

movements, and the ventral, or "what" stream, thought to support our core tasks of object and scene recognition. Our focus will be on the ventral stream, which encompasses four areas, V1, V2, V4, and inferotemporal cortex (IT).

Within each area, individual cells are responsive to stimuli presented in a small region of space, called the cell's receptive field. V1, V2, V4, and parts of IT show retinotopic organization, meaning that cells whose receptive fields are close to each other are themselves physically close together in the brain.

Across areas, the key organizational hypothesis is that the ventral stream forms a hierarchy; although all four areas are richly interconnected (Kravitz et al., 2013), they are commonly pictured as comprising a chain,

$$V1 \leftrightarrow V2 \leftrightarrow V4 \leftrightarrow IT\ ,$$

in which each area interacts primarily with its immediate neighbors. V1 takes input from the lateral geniculate nucleus (LGN), the structure that processes information immediately after it leaves the retina, and IT passes its output to nonvisual areas in prefrontal cortex. Within the chain, projections from earlier to later areas are called feedforward, and those from later to earlier areas are called feedback. For example, V2 sends feedforward projections to V4 and feedback projections to V1.

Justification of the hierarchical picture comes partly from anatomical studies tracing the projections that connect areas, and partly from functional considerations. For instance, the activation induced by a stimulus spreads sequentially up the hierarchy, emerging in V1 an average of 34–97 ms after stimulus onset, in V2 after 82 ms, and in V4 after 104 ± 23.4 ms in anaesthetized macaques (Schmolesky et al., 1998). A number of other quantities vary systematically along the ventral stream: receptive field sizes get larger from V1 to IT, and the number of neurons in successive areas becomes smaller. Most importantly, though, the hierarchical structure of the ventral stream is reflected in the representations it produces. As information passes through the ventral stream, each area computes a new representation, encoded in the activities of that area's cells. The higher the area, the better able its representation is to support tasks like classification.

A number of experimental techniques are available to study the brain's activity in general, and the representations computed along the ventral stream in particular. At the single-cell level, the tool of choice, electrophysiology, involves implanting electrodes in the brain of an animal—often a macaque monkey—and recording the electrical activity induced by presentation of different stimuli. Electrophysiological recordings have unparalleled temporal and spatial precision, but they are difficult and time-consuming to perform, meaning that most studies can examine at most a few hundred cells. Functional magnetic resonance imaging (fMRI) is a noninvasive neuroimaging technique that monitors the oxygen consumed by blocks of tissue a few millimeters on a side; oxygen intake is a correlate of neural activity. Compared to electrophysiology, fMRI sacrifices spatial and temporal resolution (fMRI scanners obtain a series of images every 2 seconds) but is quicker to perform and safe to use on awake human subjects as they perform behavioral tasks. Finally, magnetoencephalography (MEG) and electroencephalography (EEG) are other imaging techniques, also suitable for human subjects, that give millisecond temporal resolution. However, these methods rely on electrical or magnetic fields recorded at the scalp, making their spatial resolution poor.

Once one of these methods has been used to capture neural responses within an area, it remains to characterize the information that this area contains. For instance, we are often interested in determining whether an area's representation is sufficient to classify a stimulus based on its object or scene category. One relatively new analysis technique, neural decoding, underlies many of the studies reported in this chapter. The idea is that if a given representation contains category information, it should be possible to read this information out directly using a classifier. Concretely, suppose an experiment showed images $I_1, \ldots, I_n$, from categories $c_1, \ldots, c_n$, and that image I_i elicited neural response R_i. The classification problem is then to learn a mapping directly from neural response to category label. As usual in machine learning, the set of pairs (R_i, c_i) are divided into a training set used to train a classifier and a test set used to validate its performance. If good classification is possible, one may conclude that the representations R_i contain category information. Generally, decoding studies use linear classifiers, extracting category information in a way similar to how downstream neural populations may do so.

Several notes are in order. First, the fact that it is possible for a classifier to extract category information from an area's activity pattern does not mean that the brain itself extracts this information. Relatedly, the presence of category information does not rule out the possibility of other sorts of information being housed in a given region as well. Last, a negative decoding result, a failure to read out category from activity, does not necessarily imply an absence of information: using finer-grained information or different algorithms, the brain may access information our classifiers cannot.

17.2.1 Representation by Area This collection of experimental and analytical techniques allows us to explore how information evolves as it passes along the ventral stream. Our particular focus in this section will be on the ability of each area's representation to support classification, though one may hypothesize that classification supporting representations also tends to carry semantic information important for fuller scene interpretation as well.

We begin with V1. The properties of the V1 representation are described in detail in chapter 12, but to briefly summarize, V1 cells are often modeled as Gabor filters, selective for bar-like stimuli at particular positions, orientation, and scales. The V1 representation is already more abstract than one using pixels, supporting edge detection and object segmentation, for instance, but it lacks the high-level semantic content and robustness needed for effective classification.

At the opposite end of the ventral stream, the IT representation contains a remarkable amount of category information. fMRI studies, for instance, have found subareas of IT specifically responsive to certain object classes, such as faces, places, and bodies. In a decoding study Hung et al. (2005) classified objects into one of eight categories using a representation consisting of electrophysiological recordings from approximately 250 IT neurons, achieving a high mean accuracy of 94% (with chance being 1/8 = 12.5%).

In a combined MEG and fMRI study, Cichy et al. (2014) explored the intermediate processing between V1 and IT. Taking advantage of MEG's millisecond temporal resolution, the authors were able to pose a large ensemble of decoding problems, one for each time point after stimulus onset. By examining the errors made in each of these problems, they conclude that representations at later times are better able to make fine classification judgments. For instance, decoding performance for the superordinate judgment of whether or not a stimulus was animate was highest at 157 ms, while the peak accuracy for the finer judgment of whether or not an animate stimulus contained a body occurred later, at 170 ms.

Exactly how these population-level decoding results arise from the response properties of individual cells in the various brain areas is not completely clear. A simplified picture has neurons in each successive area becoming responsive to more and more complex features, formed as combinations of the features represented at the layer below. V2, for instance, contains cells responsive to angles and junctions formed by combinations of the bar-like features preferred by V1 cells (Ito and Komatsu, 2004). Farther up, a number of studies have argued that V4 represents still more complex curvature features (Carlson et al., 2011). IT contains cells responsive specifically to certain high-level object classes, for instance faces. At each stage, though, most neurons are responsive to a variety of stimuli, arguing against a one-to-one correspondence between cells and high-level features. Overall, response properties of single cells are unlikely to tell the whole story: the full discriminative power shown by the higher-level ventral stream areas emerges from the conjunctive activity of the area's whole population.

17.2.2 Invariance To begin to get a more fine-grained picture of how neural representations come to support classification, we turn our attention to invariance to nuisance transformations such as translation and scaling. These transformations can radically change the pixel content of an image while preserving its class label, giving rise to difficult situations in which two images with very different pixel-level content should be assigned to the same category.

To isolate the effect of invariance on classification performance, Anselmi et al. (2013) looked at several classification problems in which all images were rectified to a canonical position, scale, and orientation before classification. In this regime, even the simplest pixel representations were able to achieve good classification performance. This suggests that, at least for the classes studied, much of the performance difference between lower and higher ventral stream representations may be due to the latter's invariance properties.

A particular benefit of invariant representations is that they reduce the sample complexity of the learning process, the number of labeled images a learner has to see before being able to categorize new examples. Intuitively, if a learner requires a separate example for each pose in which an object could appear, the total number of examples needed becomes extremely large. Anselmi et al. (2013) quantifies this intuition in the case of translation. Consider a collection of objects of $p \times p$ pixels, which may appear anywhere in an image of size $rp \times rp$. Given an invariant representation, it requires $O(p^2)$ examples to learn a linear classifier, as compared to $O(p^2r^2)$ in the noninvariant case.

Like category selectivity, invariance builds along the ventral stream. Decoding is again an important tool. In using decoding to test for invariance to position, say, the classifier is trained on representations elicited by images shown at positions P_1, …, $P_{(n-1)}$ and tested on images shown at position P_n. Thus, the classifier is never shown examples of the position in which it is tested; good classification performance can only be attributed to information being preserved under positional shifts.

Using physiological data from IT, Hung et al. (2005) find decoding performance decreased by less than 10% in the presence of changes in translation and scale. In an MEG study Isik et al. (2014) show that invariance emerges in stages over the course of visual processing, with invariance to size arriving first, followed by invariance to position. In addition, invariance to smaller transformations precedes invariance to larger transformations.

Invariance also increases at the single-cell level, with responses in V1 being invariant to at most small transformations while some IT cells show invariance to fairly large degrees of rotation, position, and scale change (Ito et al., 1995; Hung et al., 2005). Additionally, Rust and DiCarlo (2010) show an increase in invariance from V4 to IT.

17.2.3 Summing Up: Lessons for Models The experimental results summarized above pose a clear set of desiderata for the computational models we will study in the remainder of this chapter. At a minimum, the models should mimic cortex's hierarchical structure, arriving at the top level with a representation sufficient for classification. Experimental findings about invariance give an important clue about how these representations can be built. Sections 17.3 and 17.4 encode these insights in models of the visual system's initial feedforward processing stage, arriving at algorithms that mimic humans' ability to classify objects in scenes.

17.3 Feedforward Models

The past five years have seen dramatic progress in feedforward models of visual classification, yielding algorithms that can compete with humans on challenging tasks. Importantly, the class of models, convolutional neural networks (CNNs), responsible for this success, are inspired by the architecture of visual cortex: even in no-holds-barred computer vision, where performance, rather than biological fidelity, is the measure of success, neuroscientific insights have played an important developmental role.

Progress in CNNs, and the history of their rise to prominence, is encapsulated in the results of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), an annual contest in which models compete on a data set notable for its size, containing 1,000 image categories, each with between a few hundred and more than a thousand example images. In ILSVRC, models are judged on their ability to assign to an image a correct object-category label; example categories include "Afghan hound," "electric fan," "yurt." The competition uses natural images harvested from the Internet and not cropped to isolate the target item: to correct for the fact that an image may contain multiple object classes, models are allowed to produce five candidate labels and are considered to achieve a correct classification if any one of these five matches the target. All statistics reported here are from the survey paper (Russakovsky et al., 2014).

In 2010, the first year the contest was held, the winning model attained 28.2%, an impressive result given the difficulty of the task, but significantly short of the 5.1% error achieved by a human labeler. The following year saw a modest performance improvement, with the winning model's classification error dropping to 25.8%. Both this model and the 2010 winner are representative of the computer vision techniques popular at the time, representing an image in terms of the orientation statistics of its spatial gradients.

In the following year, 2012, CNNs entered the scene, cutting error rates dramatically, with Krizhevsky et al.'s (2012) AlexNet attaining 16.4% error. In subsequent years, CNNs have continued to dominate the

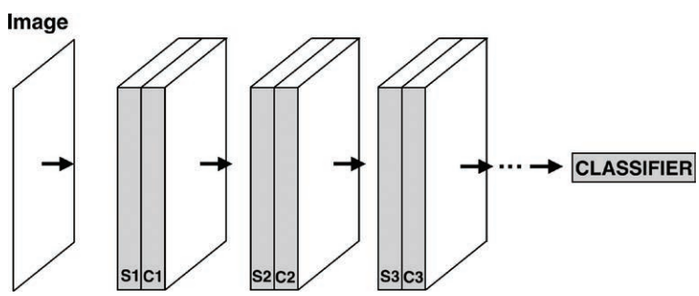

Figure 17.1 The convolutional neural network (CNN) architecture. Inspired by Hubel and Weisel's findings in V1, a CNN is composed of a series of layers, each with a simple (S) and complex (C) sublayer. Simple cells build selectivity for complicated features, and complex cells build translation invariance. As it passes through the network, an image undergoes a series of transformations, the end result of which is a representation that can be fed into a classifier. This applies more or less directly to HMAX and LeNet; more recent models include simple-only layers.

competition, becoming by far the most popular model class and achieving steadily decreasing error rates: 11.7% in 2013 and 6.4% in 2014. He et al. (2015) claimed 4.94% on the 2012 data set, beating reported human performance.

The success of CNNs is not limited to ILSVRC; it is now safe to say that they are the tool of choice for the field of object and scene recognition as a whole. They have also made an impact in related fields, for instance, being used to compute visual features for game playing (Mnih et al., 2015), and have attracted considerable investment from industry. As described below, CNNs also account well for humans' recognition behavior and provide unprecedented matches with electrophysiological data.

17.3.1 Convolutional Neural Networks CNNs have a long history prior to their recent climb to computer vision preeminence. The first CNN was Fukushima's 1980 model, the Neocognitron (Fukushima, 1980, 1988). In the 1990s, Yan LeCun and collaborators developed the LeNet family of CNNs (LeCun et al., 1998) and used them successfully for tasks like handwritten digit recognition. In 1999, Riesenhuber and Poggio proposed HMAX, a CNN specifically designed to model mammalian visual cortex (Riesenhuber and Poggio, 1999). HMAX continued to be developed in the 2000s (Serre et al., 2007b). While the different CNNs developed over the years are distinguished by various design choices, most share a core architecture: this will be the focus of this section. Section 17.3.3 presents HMAX, whose more explicitly biological orientation motivates a few important points of difference.

A series of classic experiment done by David Hubel and Thorsten Wiesel in the 1960s provide the principal neuroscientific inspiration for the CNN architecture. Hubel and Weisel recorded from V1 cells in anesthetized cats as the animals viewed patterns of light displayed on a screen. They found two main types of cells: simple cells responded to bar-like patterns at a particular orientation and position on the screen. Complex cells also responded to bars and also had a preferred orientation, but had a small degree of position invariance: their responses were preserved under small translations of the stimulus.

Mimicking the hierarchical organization of visual cortex, a CNN is composed of a succession of layers. Extrapolating Hubel and Wiesel's findings, each unit in the network is of either simple or complex type: simple cells build selectivity for complicated features whereas complex cells build translation invariance. Classically, the network is a composed of a series of layers, each with a simple and complex sublayer; complex cells take input from their own layer's simple cells, which, in turn, take input from the complex cells in the layer before. More recently (Krizhevsky et al., 2012; Szegedy et al., 2014), many models modify this schema by postfixing or interspersing some number of layers that contain only simple cells; we stick to the classical structure here.

Just as category information is exposed as an image passes through the areas of the ventral stream, a CNN's layers enact a similar series of transformations, culminating in a top-layer representation suitable for input to a classifier. This architecture is shown in figure 17.1. We now explore the complex and simple cell types in turn.

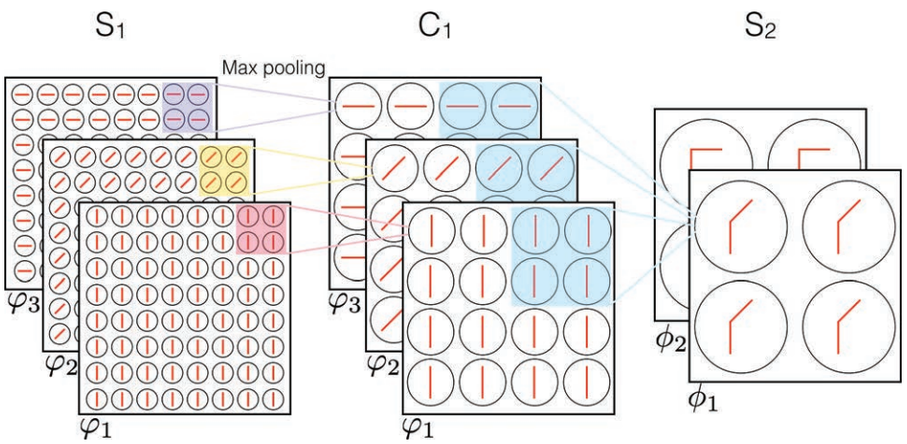

Figure 17.2 A segment from the early layers of a convolutional neural network. Each complex (C) and simple (S) sublayer consists of a collection of feature maps, shown in the figure as sheets, consisting of cells responsive to the same feature at different positions. Complex cells build position invariance by pooling over inputs from a single feature map in the previous sublayer. Simple cells build selectivity for complicated features by combing inputs from multiple feature maps. As shown in the figure, receptive field sizes increase farther up the network.

Consider first building a V1 complex cell like the ones Hubel and Wiesel observed; we will denote such a cell by $C_p^1(\theta_i)$, where the superscript "1" tells us that the cell lives in the first layer, and the subscript p indexes the cell's position within the layer. Because of the network's retinotopic organization, p also picks out a position in the input image where the cell's receptive field is centered. The argument θ_i is the orientation for which the cell is selective. To denote the activation of a cell, rather than its label, we use boldface: $\boldsymbol{C}_p^1(\boldsymbol{\theta}_i)$.

To make our task explicit, we need to wire up $C_p^1(\theta_i)$ in such a way that it activates to a bar at orientation θ_i appearing anywhere in its receptive field; we will denote the receptive field $R(p)$. Hubel and Wiesel proposed a simple way in which this could be done. They hypothesized that $C_p^1(\theta_i)$ receives input from a collection of simple cells, all also with orientation selectivity for θ_i, whose smaller receptive fields cover $R(p)$. Thus, if any one of these afferent cells becomes active, it indicates the presence of a θ_i-oriented bar somewhere in $R(p)$. All $C_p^1(\theta_i)$ has to do, then, is perform an OR-like operation over its input signals, becoming active if any one of them is active. In CNNs, the role of the OR operation is often played by the max function, yielding the so-called max-pooling activation function for V1 complex cells: $\boldsymbol{C}_p^1(\boldsymbol{\theta}_i) = Max_{q\in R(p)}\boldsymbol{S}_q^1(\boldsymbol{\theta}_i)$.

Complex cells in higher layers of the network work similarly. The complex cells in an arbitrary layer l perform max pooling over a localized collection of l's simple cells all tuned to the same feature ϕ_i. In higher layers, these features are more complicated than the orientations we have considered so far:

$$\boldsymbol{C}_p^l(\phi_i) = Max_{q\in R(p)}\boldsymbol{S}_q^l(\phi_i). \tag{17.1}$$

This expression requires that, for each position p, there does in fact exist a family of identically tuned simple cells at a range of positions in $R(p)$. This motivates a key architectural assumption of CNNs: simple cells are organized into "sheets," called *feature maps*, which tile the input plane with cells with the same tuning. In figure 17.2, feature maps are shown as stacked arrays. A single layer is composed of multiple feature maps. As also shown in figure 17.2, the fact that complex cells pool over multiple simple cells means that the complex cells' receptive fields are larger.

We now consider simple cells, again using an example of an early network layer, this time the second layer. Layer 2 simple cells are responsive to patterns such as corners and T-junctions, which can be constructed as combinations of the oriented bars selected for by their layer-1 complex cell afferents. Simple cells are similar to complex cells in that their afferents are spatially localized but differ in the fact that they take input from all of the previous layer's feature maps rather than just one.

In sum, a layer-2 simple cell's inputs come from complex cells detecting line segments at a full range of orientations and a small range of positions. To become selective to a feature like a corner, the cell has to activate strongly only if a certain combination of these segments, for instance a 90° one and 0° one, are present in the image. Whereas complex cells played an OR-like role, simple cells have to behave more like an AND over their inputs. To allow only particular inputs to activate our simple cell, we introduce connection weights modulating the influence these inputs have. Inputs corresponding to features included in our target pattern get high weights; other features get low ones. In real networks, as we will see, simple cell weights and the features they define are learned automatically rather than being set by hand.

Considering a simple cell $S_p^l(\phi_i)$ in an arbitrary layer l, suppose that the previous layer has n feature maps, corresponding to features $\varphi_1, \ldots, \varphi_n$. The afferents of $S_p^l(\phi_i)$ are then a subset of layer $l-1$'s complex cells, localized in space to $R(p)$, but spanning the full range of feature maps. Denote this set

$$A\left(S_p^l(\phi_i)\right) = \left\{C_q^{l-1}(\varphi_j) | q \in R(p),\ j = 1\ldots n\right\}.$$

One of these afferent cells, $C_q^{l-1}(\varphi_j)$, is connected $S_p^l(\phi_i)$ via the weight $W_{(q,\varphi_j)\to(p,\phi_i)}$. With this (somewhat cumbersome) notation in place, we can present the simple cell activation function, the counterpart of equation 17.1:

$$\boldsymbol{S}_p^l(\phi_i) = \eta\left(\sum_{A\left(S_p^l(\phi_i)\right)} W_{(q,\varphi_j)\to(p,\phi_i)}\boldsymbol{C}_q^{l-1}(\boldsymbol{\varphi}_j) + b_{\phi_i}\right). \tag{17.2}$$

Here, η is a nonlinear function, such as a sigmoid or hyperbolic tangent, and b_{ϕ_i} is a bias term. Overall, then, a simple cell sums the activations of its afferents, weighting each activation by the corresponding connection weight, adds a bias term, and applies a nonlinearity. As with complex cells, a simple cell's receptive field spans the receptive fields of its afferents and is therefore larger.

To take its place in the larger network architecture, $S_p^l(\phi_i)$ must itself be part of a feature map. In other words, there must exist simple cells $S_q^l(\phi_i)$ for all other positions q. Since ϕ_i-specificity is defined by a pattern of weights, a feature map is nothing other than an array of simple cells with the same pattern of input weights, placed at different positions in the layer; the weights are often said to be *tied*. Thus, computing the activations of all the cells in a feature map is mathematically

equivalent to convolving the layer's input with the linear filter defined by its feature's weight pattern, adding the bias b to the result, and applying η. This is the origin of the name "convolutional neural network." Simple cell weights and biases are free parameters in a CNN and must be learned during a training period; the procedure is outlined below.

With simple and complex cell types in hand, a modeler can assemble an entire network by specifying a number of parameters: the number and sizes of its layers, the number of its feature maps, the sizes of its receptive fields, and so on. Classical CNNs such as HMAX, LeNet 5, and the Neocognitron make choices roughly similar to visual cortex, having around four layers and sticking to the alternating simple and complex sublayer scheme that we have used in this section. More recently, architectures have grown more exotic; GoogLeNet, winner of the 2014 ILSVRC, for instance, has 22 layers, and Simonyan and Zisserman (2014) present another high-performing model with 16–19. As described above, many modern models also include some number of simple-cell-only layers. The question of why such apparently nonbiological choices, particularly as regards layer numbers, yield state-of-the-art recognition performance is an interesting one that will require further investigation.

17.3.2 Learning in CNNs Having specified the structure of a CNN, it remains to learn the simple cell connection weights. A number of learning techniques are possible. HMAX, for instance, uses a simple unsupervised learning scheme, described below. However, by far the most common approach in performance-optimized CNNs is to optimize a supervised learning objective function using gradient descent. These methods define an error function that measures the model's performance on a training data set of labeled images, compute the gradient of this error function with respect to each of the model's weights, and perform standard gradient descent.

The output layer of a CNN has one unit for each possible class label an input image could be assigned. The activation in the yth output unit in response to an image X represents the model's belief that y is X's correct label. We denote by $f(X; w)$ the output activations induced by an image X fed through a network, parameterized by weights w. Given an image-correct label pair (X_i, y_i) from the training set, a perfectly correct network would produce $f(X_i; w) = \delta_{y_i}$, a vector uniformly zero aside from a one in the y_ith position, indicating a complete concentration of belief on the correct label. A model accrues error to the extent to which it deviates from this desired output. Specifically, we define an error function $E(w)$ measuring the difference between $f(X_i; w)$ and δ_{y_i}. A simple L_2 distance illustrates the point, though more complex measures such as cross-entropy are generally used in practice.

$$E(w) = \sum_i \| f(X_i; w) - \delta_{y_i} \|_2^2 . \tag{17.3}$$

Given a choice of error function, optimization proceeds by gradient descent.

A given weight w is updated by

$$w_j^{t+1} = w_j^{t+1} + \alpha \frac{\partial E}{\partial w_j} \tag{17.4}$$

where α is a learning rate. The actual computation of the gradient is usually done by an algorithm called backpropagation. While it works very well in practice, backpropagation is generally considered neurally implausible, and we do not present it here. Nevertheless, even backpropagation-trained networks can be neuroscientifically interesting, as they make it possible to test hypotheses about general network structure, and choice of learning objective, network properties that can be chosen independently of any specific learning process.

17.3.3 HMAX While the CNNs used for tasks like the ILSVRC are inspired by neuroscience, their main objective is classification performance. HMAX, by contrast, is explicitly designed as a cortical model, drawing design choices from the experimental literature wherever possible. HMAX differs from the CNN template presented so far in several ways. First, HMAX simple cells pool over scales as well positions, building invariance to both changes in size and translations. Second, while other models learn their simple cells' weights at every layer, HMAX layer-one simple cells are hand set to have the Gabor-like tuning found in V1. The first layer is composed of Gabor filters at 16 different scales, from 7×7 pixels to 37×37, and at four different orientations, 0°, 45°, 90°, and 135°.

In addition, HMAX's simple cell activation function is somewhat different from equation 17.2. To avoid the extra notation associated with different scales, we will drop some of our earlier indexing, considering an arbitrary simple cell S_i. As before, S_i receives inputs from a set $A(i)$ of complex cells from the previous layer. Its activation is given by

$$\boldsymbol{S_i} = \exp \frac{-1}{2\sigma^2} \sum_{j \in A(i)} (w_{ji} - \boldsymbol{C_j})^2 . \tag{17.5}$$

The interpretation here is that the synaptic weights w_{ji} define a "template": S_i's excitation is when the

activations in $A(i)$ match this template exactly, and it falls off in a Gaussian way as they deviate from it.

As measured by performance, probably the important distinguishing feature of HMAX is its learning rule. Proceeding with the activation function interpretation above, learning in HMAX consists of choosing an afferent template for each of its simple cells. HMAX accomplishes this with a simple, unsupervised sampling procedure. For each simple cell S_i, the model picks an (unlabeled) image from a data set provided, which, when presented to the network, induces a particular pattern of activation in the complex cells in S_i's afferent set, $A(i)$, with $\boldsymbol{C}_j$ being the induced activation in the jth simple cell. This activation pattern is stored as the template; $w_{ji} := \boldsymbol{C}j$. In other words, each simple cell becomes tuned to the neural image of a particular image patch arising in the training set.

This learning scheme has two advantages over the supervised one presented above. First, it avoids the biological implausibility associated with backpropagation. Second, the fact that it is unsupervised means that it does not require a large database of natural images, likely better matching the conditions faced by a young human or animal during development. These advantages, though, come at the price of reduced performance. As we will see below, HMAX achieves respectable recognition performance, competing with humans on some tasks. However, it lags significantly behind performance-optimized CNNs in most classification problems. The difference between HMAX's unsupervised learning rule and the supervised backpropagation used by most other models is likely the main reason for this deficit.

17.3.4 Assessing CNNs A number of experimental methods are available to help us to study the performance and properties of CNNs and to assess the similarity of CNNs to the mammalian visual cortex.

A first question is behavioral: How well can CNNs account for human recognition performance? We have already seen, CNN performance compares favorably to humans in the difficult ILSVRC recognition task. A more controlled experiment compared HMAX to humans on a simpler visual task, determining whether or not an image contains an animal. In difficult cases, humans can deploy a number of strategies to make this determination, for instance, moving their eyes for detailed examination of particular image regions or using prior knowledge about where in a natural scene an animal might hide. However, strategies like this require time and feedback processing and are outside the portion of visual processing that CNNs attempt to model, namely, the initial feedforward pass through the ventral stream. In an attempt to level the playing field by limiting humans to feedforward processing as well, Serre et al. (2007a) used a paradigm called *masking*, in which each target image is only shown for 20 ms and is followed by a mask, a noise image modified to share the target's low-level statistics. The motivation for this procedure is that lower visual areas will be occupied with feedforward processing of the mask images during the time period when feedback processing of the target could ordinarily occur, thereby preventing this feedback processing from taking place.

Under these conditions, human subjects achieved 80% accuracy on the animal/no-animal classification task. HMAX achieves a similar 82% accuracy. Moreover, the pattern of errors made by the model was qualitatively similar to that of humans: humans and the model tended to find the same images difficult.

To get a finer-grained picture of the properties of a trained CNN that supports recognition performance, Zeiler and Fergus (2014) use a clever algorithm to visualize the features preferred by units at different layers. Applying this algorithm to a CNN similar to the first ILSVRC winner (Krizhevsky et al., 2012), they find features qualitatively similar to those detected by human macaque visual cortex (figure 17.3). Unlike HMAX, in which the layer-1 Gabor filters are hand chosen, the weights at all layers of this CNN are learned during training. Even so, edge-like tuning still emerges in the network's first layer. The second layer selects for features involving multiple edge transitions, such as corners, similar to hypothesized V2 features. The highest layers of the network, 4 and 5, are selective for recognizable objects and object parts, such as animal faces in layer 4 and whole animals in layer 5, suggesting a similarity to mammalian IT. Roughly, then, the features a given CNN layer selects for are qualitatively similar to those selected for by a biological layer at a similar point in the hierarchy.

The feature-preference correspondence between biological neurons and artificial CNN units suggests a more direct comparison: How well can CNN activations predict actual physiological data? Yamins et al. (2014) recorded the responses of macaque IT and V4 units and compared the results to CNN activations induced by the same images. To make the comparison, Yamins et al. compare each biological neuron to the linear combination of artificial units that best matches its responses. Using this procedure, Yamins et al. find that the CNN's top layer explains 48.5 ± 1.3% of the explainable variance in the biological IT cells' firing patterns, and the model layer below explains 51.7 ± 2.3% of the variance of biological V4 firing rates. To put this number in context, it is comparable to the matches other encoding

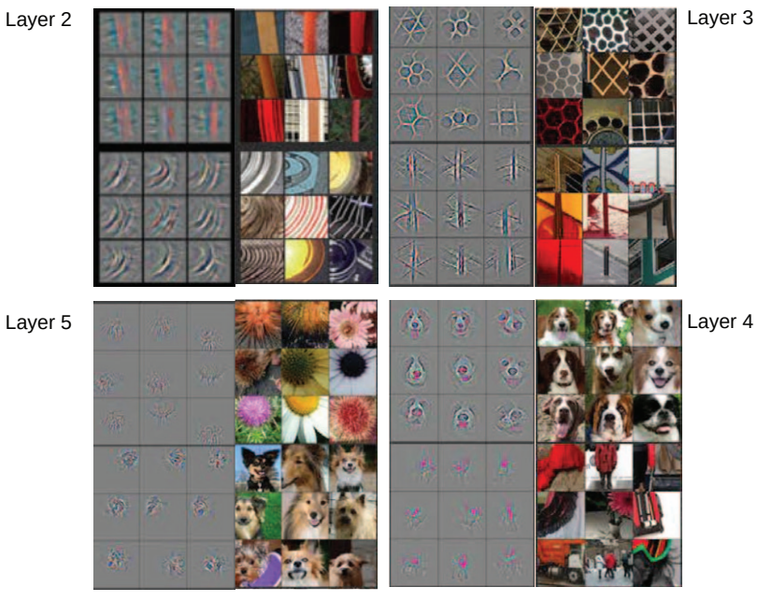

FIGURE 17.3 Zeiler and Fergus (2014) developed a deconvolution algorithm to visualize the features that best activate the units in a CNN. Here, the algorithm is applied to AlexNet, the best-performing CNN in ILSVRC 2012. Each 3×3 grid corresponds to one feature map. The images on the left of each panel show algorithmic reconstructions of the image features that cause high activations, and the color images on the right show patches in which these patterns occur. As in primate visual cortex, units in higher layers are responsive to complex stimuli. Reproduced with permission and modified from Zeiler and Fergus (2014).

models have achieved to the much better understood lower visual areas.

Another physiological question one can ask about CNNs concerns the biological reality of their basic operations. The simple cell activation function in equation 17.2 is a standard and widely accepted model of neural activation throughout the brain, but a CNN's complex cells' max-pooling operation is more exotic. Does it actually occur in cortex? Lampl et al. (2004) recorded from V1 complex cells and found that for 80% of them, a max-like activation function fit their responses better than a linear model. Knoblich et al. (2007) present a circuit-level model of how max pooling might occur.

Given these impressive results, it is natural to wonder if CNNs now offer a full account of feedforward object classification. While they undoubtedly make substantial progress toward this goal, a few caveats remain. First, while CNNs achieve human-like total error on tasks like the ILSRVC, the patterns of errors they make deviate somewhat from those made by humans (Russakovsky et al., 2014). Russakovsky compared human performance to GoogLeNet, the most recent ILSRVC winner. More than half of human errors are attributable to knowledge deficits as opposed to visual errors: humans may not realize that the correct label is an option and may struggle with fine-grained and obscure classes, specific species of dogs, for example. CNNs, by contrast, have more purely visual problems, finding it difficult to label objects that take up a small percentage of the image as a whole, and struggling with images distorted by Instagram-style filters. CNNs also struggle with abstract representations of objects such as drawings or statues. Taking these last two sources of error together, one might hypothesize that, compared to humans, a CNN is more likely to classify based on textural rather than structural features of an image. Relatedly, Szegedy et al. (2013) demonstrate that CNNs can be fooled by adversarial examples, images constructed by taking an image the model originally classified correctly, and adding a specially designed perturbation. While the resulting image is indistinguishable to human eyes from the original, the model nevertheless labels it incorrectly.

Second, achieving high classification accuracy with biological training mechanisms remains an open challenge. As mentioned, the popular backpropagation algorithm relies on non-neural mechanisms. More generally, commonly used supervised learning paradigms require very large amounts of labeled training data, arguably significantly more than human learners need. See Durbin and Rumelhart (1989), O'Reilly (1996), and Balduzzi et al. (2014) for proposed biologically plausible backpropagation alternatives and section 17.6.1 for further discussion of small-sample learning.

So feedforward classification is not without its loose ends. Nevertheless, CNN modeling has been a significant advance. Indeed, it is arguably the most successful effort to date in solving a difficult biological problem in an at least broadly biological way.

17.4 Feedforward Scene Recognition

Models like HMAX and other CNNs are generally used for object recognition, as in the animal-detection task described above, but in principle, nothing prevents them from categorizing scenes as well. Indeed, a natural conjecture is that scene recognition is nothing other than repeated object recognition—we might recognize a street scene, for instance, by realizing that it contains cars, pedestrians, and buildings in particular arrangements. But recent work gives compelling experimental evidence that this object-based view of scene perception is incorrect, or at least incomplete. Rather, the human object recognition apparatus is supplemented by a distinct set of processes housed in a distinct set of brain areas that quickly extract the global geometric properties of a scene, its "gist," that support fast recognition of a scene's semantic category and its functional properties. This section reviews this experimental work on gist extraction and then shows that its findings have been captured in computational models.

If the properties of a scene that support classification are not the identities of its constituent objects, what are they? The evidence suggests that scene classification relies on a collection of global geometric properties, such as a scene's openness, the extent to which the horizon is visible; its depth; and its navigability. In addition to their supporting role in classification, these properties have clear ecological relevance in their own right.

Greene and Oliva (2009) showed subjects masked images of natural scenes for variable amounts of time and examined which properties subjects were able to extract. They found that a scene's global properties became available earlier (mean = 34 ms) than did its basic-level semantic category ("mountain," "ocean," etc.; mean = 50 ms), consistent with a picture in which global properties underlie category judgments. To show the sufficiency of these properties for scene classification, Greene and Oliva (2006) had human subjects evaluate the prevalence of each of seven global properties in each database of natural scene images and trained a classifier to predict the category of a scene given only its properties, as coded by humans. They found a good correlation between human performance and that of the classifier, and a similar distribution of mistaken classifications.

Other timing studies examine when object information enters the mix, and they argue for the following approximate ordering: global properties, then scene category, then individual object categories. In a study by Fei-Fei et al. (2007), subjects generated free descriptions of scenes they had viewed for various durations. While the portions of their descriptions involving objects changed and became more detailed for longer viewing times, subjects' high-level scene classifications remained relatively constant. This is consistent with an interpretation in which the scene category is extracted quickly while object identities require longer viewing to emerge in full.

In addition to limiting the processing time available to viewers, other studies have restricted the information available in the frequency domain. Schyns and Oliva (1994) showed subjects images from which all frequencies above two cycles per degree had been eliminated. After this modification, objects generally appear only as unidentifiable blobs, but subjects' ability to identify the scene category was largely unimpaired. A similar low-pass-filtered image is shown in figure 17.4A.

Further evidence for scene processing as distinct from object processing comes from neuroimaging studies showing a distinct network of brain areas underlying scene perception. Imaging work has found at least three areas that preferentially respond to scenes or places over objects: the parahippocampal place area (PPA), retrosplenial cortex, and the occipital place area. In a particularly striking finding, Epstein and Kanwisher (1998) showed that PPA activation to an image of a room was more or less unchanged by removing all the moveable objects (furniture etc.), further evidence of a geometric rather than object-based encoding. Decoding studies with fMRI also support this view. For instance, Park et al. (2011) were able to decode scene information from all three of the scene-specific areas mentioned, as well as from the object-selective area lateral occipital cortex (LOC), but the pattern of errors the classifier made showed a sensitivity to spatial properties (open vs. closed) in the scene areas, as compared to content in LOC.

17.4.1 Feedforward Models of Scene Perception

These experimental findings suggest some ways that a representation designed to support scene classification should differ from one for object classification. First, as indicated by humans' ability to classify scenes using only low spatial frequencies, a scene representation should need to be only weakly spatially localized. Second, and relatedly, the scene representation should be able to be relatively low dimensional, as shown by the good performance of the model in Greene

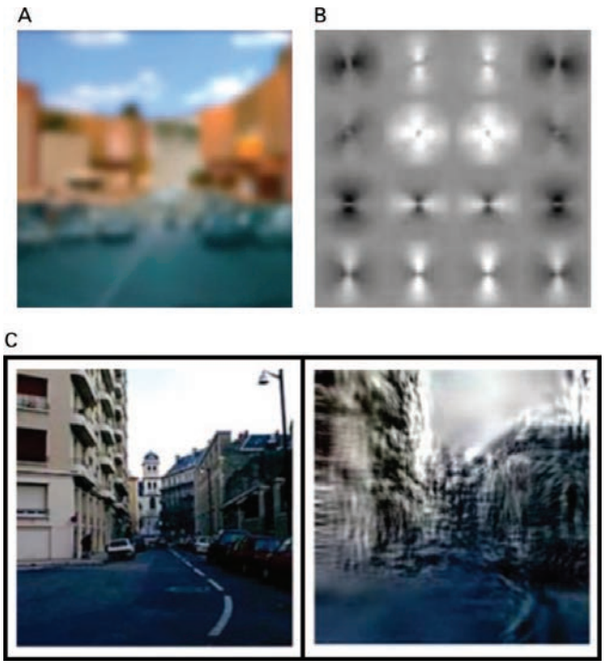

FIGURE 17.4 (A) Despite the fact that low-pass filtering has removed object information, it is still easy to recognize that the image shows a street scene. (B) A global feature template, one principal component of a multiscale filter-bank representation of a database of natural images. The image is divided into 16 subregions, and an output value is computed for each. The polar plot within subregions shows how the output is computed for different orientations and spatial frequencies: angle in the plot represents orientation, and distance along the radius represents spatial frequency, with low frequencies near the center. White regions correspond to positive outputs, black, to negative. (C) A noise image (right) coerced to share global features with a target image (left). The comparison shows that global features preserve the scene's large-scale geometry even while losing its finer details. Reproduced with permission and modified from Oliva and Torralba (2006).

and Oliva (2006) using only seven human-identified features.

Most models of scene processing stem from Oliva and Torralba (2001). Here we present a slightly more neural variant by the same authors, Oliva and Torralba (2006). The model begins with a V1-like representation given by the output of a bank of Gabor filters spanning a range of orientations and scales and positioned densely across the image. A first reduction step coarsens the spatial resolution of this representation, by dividing the image into $N \times N$ windows and averaging the output of each filter within each window. This reduces the dimensionality of the representation to $N \times N \times S \times R$, where S is the number of spatial scales in the filter bank, and R is the number of orientations.

The second reduction step computes this representation for each of a large number of images in a data set and compresses the resulting set of vectors with principal components analysis. Each principal component is a weighted combination of filter outputs at each of the $N \times N$ windows. These weighted combinations are then "fused" into what the authors call global feature templates (GFTs), which span the whole image; an example is shown in figure 17.4B. The first M of these GFTs can then be applied to a new scene image, by applying the original filters and weighting the results, and the output fed to a classifier to extract global properties like openness, depth, and so forth. To get a feel for the representation that the GFTs compute, figure 17.4C shows a noise image modified in such a way that it induces the same GFT responses as a target image of a natural scene. While the modified noise image loses the fine details present in the target, it maintains enough of its global gist to support easy classification.

Since Oliva and Torralba (2001), a number of authors have proposed different models of scene-level processing. See, for example, Renninger and Malik (2004) and Siagian and Itti (2007) and references therein.

Scene classification is a worthy end in itself, but scene information can also be useful in the context of object detection. An image's scene category contains a lot of information about which objects it is likely to contain and where they are likely to appear. A street scene, for instance, is likely to contain multiple cars, while a mountain or forest scene probably contains none. Furthermore, cars appear at predictable (vertical) locations in street scenes; a car is unlikely to appear in the sky, for instance. Torralba et al. (2010), building on Murphy et al. (2003), use scene-level information to adjust the outputs of local object detectors, ensuring that the final detections contain a plausible number of objects of each class and that these objects appear in plausible locations.

Similar intuitions underlie the model in Torralba et al. (2006), which uses scene features to predict human eye movements during visual search. Subjects were told to find an instance of a target object class, "car," for example, and their eye movements were tracked as they looked for it. If humans use their knowledge of the target image's class, together with quickly extracted scene gist features from the given image, then a model incorporating gist features should better predict their fixations than one without.

The tendency of a human to fixate a location X is modeled as

$$S(X) = L(X)^{-\gamma} P(X \mid G, O = 1) \qquad (17.6)$$

where $L(X)$ is the local saliency of the location X. Heuristically, L measures the extent to which X is unexpected given the rest of the image; salient regions stand out. This local contribution is traded off via the exponent γ with a global term, $P(X|G,O=1)$ denoting the probability that an instance of the target class is present at the location X, given the image's global scene-gist features G, and assuming that at least one target class instance is present in the image as a whole ($O = 1$). As expected, the combined global–local model significantly better predicted which image regions subjects were likely to fixate than a local-only one.

17.5 *Feedback Connections*

While the foundations of the feedforward models presented in section 17.3 were already in place by the end of the 1960s, an understanding of visual cortex's feedback connections has been more elusive. It is generally agreed that feedback connections mediate attentional processing, carrying signals about stimulus salience and task relevance from frontal and parietal areas (see chapter 19, "Saccades and Smooth Pursuit Eye Movements," and chapter 4, "Neural Rhythms"), but there has been less consensus about the role of feedback connections in this chapter's core themes of object and scene recognition. Here, we will focus on a class of models that has come to the fore over the last fifteen years or so, which views feedback connections as primarily generative in nature, enabling the brain to synthesize as well as consume images.

That the brain can accomplish this generation is intuitively clear from everyday experiences like dreaming and visual imagination and from clinical observations of visual hallucinations. Sufferers of Charles Bonnet syndrome, for instance, report extremely vivid hallucinations which they have trouble distinguishing from reality, and which imaging studies have shown to have very similar neural signatures to normal visual processing (Ffytche, 2005). But even granting the visual cortex's generative capacity, it still remains to explain how this capacity could be used to support recognition.

17.5.1 Generative Models for Recognition

The spiritual father, arguably, of the generative approach to vision is Hermann von Helmholtz, who in 1867 advanced the theory of "perception as unconscious inference." On this view, the input the visual system receives is the result of a causal process, the imaging process mapping world states (particular configurations of objects, illumination, etc.) to two-dimensional images, or collections of LGN cell responses. The goal of visual cortex is then to invert this process, inferring from an image the world state that caused it. Note the ambiguity inherent in this problem: any image is, in general, consistent with many high-level causes. For instance, a percept of a given size might be caused by a large, distant object or by a small object close at hand.

The purely feedforward models we have studied so far approximate a solution to this inverse problem by learning a direct mapping from percept to causes, a purely bottom-up approach. Given a model of the imaging process, a purely top-down process is also possible: one could guess high-level causes and check them by running the imaging process and seeing if the result matched the observed percept. This approach has the advantage of being robust to low-level ambiguity, but in practice, of course, pure trial and error is hopelessly inefficient. Practical generative vision models use a mixed approach in which bottom-up and top-down information is progressively integrated.

Reflecting the ambiguity inherent in the inverse inference problem, generative models in vision are generally probabilistic. Sticking with our chain view of visual cortex, in which areas interact primarily with their immediate neighbors, allows us to factor the joint probability of the activities in all cortical areas, and of the input data in the LGN, as follows:

$$P(LGN,\ V_1,\ V_2,\ V_4,\ IT) = P(LGN\ |V_1)P(V_1|V_2)P(V_2|V_4)P(V_4\ |IT)P(IT). \tag{17.7}$$

The conditional probabilities $P(V_i|V_j)$ encode a model of the statistics of natural images given their content. For instance, $P(V_4\ |IT)$ indicates which V_4 responses the model expects given the presence of a particular configuration of objects, encoded as IT responses. These probabilities are acquired through experience; this is the learning problem. Once the relevant distributions are in place, we can ask Helmholtz's inversion question: which higher-level causes were likely to have generated a given LGN input? In the probabilistic context, this is called the inference problem, and it amounts to approximating or maximizing the following posterior distribution:

$$P(V_1,\ V_2,\ V_4,\ IT\ |\ LGN) = \frac{P(V_1,\ V_2,\ V_4,\ IT,\ LGN)}{P(LGN)} = \frac{1}{Z}P(V_1|V_2)P(V_2|V_4)P(V_4\ |IT)P(IT)\cdot \tag{17.8}$$

Here, we have treated the probability of the LGN input as a normalizing constant and used uation 17.7 to perform the factorization in the last step.

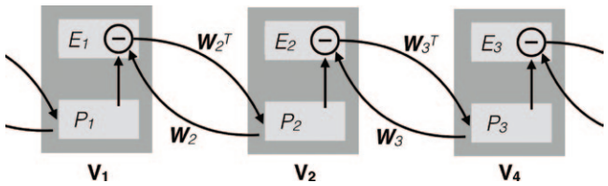

FIGURE 17.5 A possible neural architecture for predictive coding. Each area contains two populations of neurons: P_i predicts the activity of the downstream area, sending the predictions via the synaptic weights matrix W_i, and E_i computes the error made by the predictions coming from the next higher area.

In general, generative approaches to recognition share this basic structure and problem definition but differ in the way in which the optimization is performed and in how the distributions are parametrized.

17.5.2 PREDICTIVE CODING When you tell a friend about your day, you do not catalog routine minutiae. Rather, you only mention the few unusual events that distinguished this particular day from any other, relying on your friend's prior knowledge of your daily habits to fill in the details you left out. This is the intuition underlying predictive coding, a technique originally developed for speech processing in the 1960s: a sender need only transmit those parts of a signal that are unexpected given a predictive model possessed by the receiver.

Predictive coding underlies several models of visual cortex and other brain areas (Mumford, 1992; Friston, 2005). Our exposition most closely follows the model developed in Rao and Ballard (1999). In Rao and Ballard's model, higher visual areas play the role of the receiver in the sketch above, and lower areas, the sender. Higher areas use feedback connections to transmit their generative predictions of the next lower level's activity, and lower areas send up the errors in these predictions via feedforward connections.

In the simplest version of the model, the predictions made by the ith area is the linear function of the area's own activities:

$$\boldsymbol{W_i} V_i \tag{17.9}$$

where $\boldsymbol{W_i}$ is a matrix of synaptic weights. The $\boldsymbol{W_i}$ are structured so that each cell in area V_i sends connections to a spatially localized subset of cells in the layer below, the size of this subset increasing with i, corresponding to the fact that cells in higher areas have larger receptive fields.

With a Gaussian noise model, equation 17.9 gives the following conditional probabilities:

$$P(V_i | V_{i+1}) = \mathcal{N}(\boldsymbol{W}_{i+1} V_{i+1}, \mathrm{diag}(\sigma_i^2)) \tag{17.10}$$

where σ_i^2 is a layer-dependent noise variance. With this formulation, we can solve both the learning problem (finding W) and the inference problem with gradient descent; we focus on the inference problem here. From equation 17.8, passing to log space and differentiating gives the following:

$$\frac{\partial}{\partial V_i} \log P(\{\mathrm{V}\}_i | LGN) = \sum_k \frac{\partial \log P(V_k | V_{k+1})}{\partial V_i} \tag{17.11}$$

$$= \frac{\partial \log P(V_{i-1} | V_i)}{\partial V_i} + \frac{\partial \log P(V_i | V_{i+1})}{\partial V_i}.$$

After plugging in equation 17.10, we obtain the following update rule:

$$V_i \leftarrow V_i + \alpha \left(\frac{1}{\sigma_{i-1}^2} \boldsymbol{W_i}^T (V_{i-1} - \boldsymbol{W_i} V_i) + \frac{1}{\sigma_{i+1}^2} (\boldsymbol{W}_{i+1} V_{i+1} - V_i) \right). \tag{17.12}$$

This is a quantitative expression of the predictive coding intuition explained above. The second term in the sum is the error in the predictions coming from the higher area V_{i+1} and is available locally to V_i. The first term, the error in V_i's own predictions, has to be passed back up from V_{i-1} where it is computed.

These considerations suggest a neural implementation like the one shown in figure 17.5, adapted from Friston (2005). Each visual area is composed of two subpopulations: the prediction neurons P_i send predictions via feedback connections, and the error detecting neurons E_i compute the error in predictions coming down from the area above. For a more detailed attempt to map the predictive coding hardware onto neural machinery, see Bastos et al., (2012).

17.5.3 PREDICTIVE CODING: SIMULATIONS AND PREDICTIONS In simulations with a truncated version of the model above, encompassing LGN, V1, and V2, Rao and Ballard (1999) derive a number of properties of V1 cell responses, particularly the so-called extra-classical receptive field effects. As described in section 17.2, a cell's classical receptive field is the region of space in which a presented stimulus can evoke a response. Many V1 cells have in addition an extra-classical receptive field (ECRF), which is a larger area outside the classical receptive field where stimuli can modulate the cell's firing. Two examples are shown in figure 17.6. End-stopped cells are responsive to bar-like stimuli and show enhanced responses when a bar terminates in their ECRF. Another effect, surround suppression, occurs when a cell's firing decreases when the dominant orientation in its classical receptive field extends through its ECRF. In general, extra-classical effects occur at

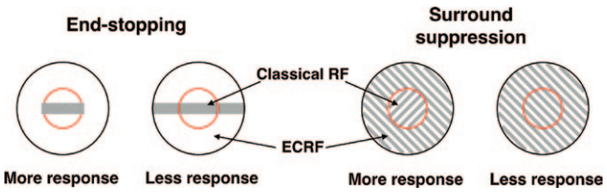

FIGURE 17.6 Extra-classical receptive fields (ECRFs). In both figures, the cell's classical receptive field (RF), the region in which a presented stimulus can directly induce firing, is shown in red, and its ECRF, a larger area in which stimuli can modulate firing, is shown in black. End-stopped cells show a decrease in firing when a bar extends beyond their classical RF. Surround suppression occurs when a cell fires less strongly when the orientation present in its classical RF continues in its ECRF.

around 80–100 ms after stimulus onset, as compared to 66 ms for classical V1 responses, suggesting a role for feedback, or at least lateral, processing.

After training their network on a database of natural images, Rao and Ballard found that a large proportion of their error-correcting neurons displayed end-stopping effects. An analysis of natural image statistics reveals why this should be so. Edges in natural images generally continue smoothly; short, isolated bar segments are rare. During training, the simulated V2 neurons adapt to this regularity and shape their predictions of V1 responses accordingly. Thus, short bars contained within a V1 cell's ECRF trigger a prediction error and cause increased activity in the V1 error-detecting cells. Recall that the receptive field of each simulated V2 neuron contains several V1 neurons; this group constitutes its central cell's ECRF.

The analysis for surround suppression is similar. By and large, the dominant orientation in a natural image changes slowly and smoothly over space, so this is what V2 cells come to predict. When this prediction is upheld, error-detecting neurons fire less, resulting in the observed suppression.

Subsequent studies have used fMRI to investigate predictive coding. Murray et al. (2004) used stimuli of a variety of types, each containing a collection of elements. In one class of stimuli, the elements could be interpreted as forming a coherent group of some kind, while, in the other class, they appeared essentially random. A collection of line segments forming a coherent two-dimensional shape, for instance, would be a member of the first class, while a collection of scattered segments would be a member of the second. In fMRI experiments, the groupable stimuli were found to induce increased response in higher visual areas, such as the object-responsive area LOC, and reduced responses in V1. This finding is consistent with predictive coding: groupable stimuli admit a higher-level explanation that is able to "predict away" lower-level responses. The authors point out, however, that the result is also consistent with an alternative "sharpening" explanation, in which lower-level responses that are consistent with higher-level predictions are enhanced, and others reduced, perhaps as a way to produce a sparser representation overall.

More directly related to our concern with object recognition, Egner et al. (2010) showed subjects images of either faces or houses. Before each image, subjects saw a colored frame that was stochastically predictive of the image category to follow: a face followed green frame with probability 0.25, a yellow frame with probability 0.5, and a blue frame with probability 0.75. Egner et al. recorded the activation in the FFA, a fusiform area selective for faces. Consistent with other studies, they found some FFA activity in response to house stimuli, but this activation was much higher, almost as much as for faces, when these stimuli were unexpected. The predictive coding explanation of this result is that the FFA activation is an error-correcting response induced by the violated face expectation. The authors rule out an alternative explanation that face expectation uniformly enhances FFA activation, perhaps as a result of anticipatory neural activity. On this explanation, activation on face-expected, face-present trials should be at least as high as on face-unexpected, face-present ones, the opposite of the actual finding that face-present activation increased the face expectation.

For a recent review of other predictive coding, including proposals about how it could function in cognitive domains other than vision, see Clark (2013).

17.5.4 OTHER APPROACHES TO INFERENCE Schemes like Rao and Ballard's are perhaps the best-developed approaches to inference in generative vision, but there is no shortage of other candidates. We focus in particular on one algorithm, known as belief propagation or message passing, in part because it has been proposed as a general algorithm for neural probabilistic inference, applicable to areas beyond visual cortex. As the name suggests, the message-passing algorithm is based on communication between probabilistically related variables, making it particularly amenable to neural interpretations. The specific approach we sketch here was developed in Lee and Mumford (2003). For another approach, see Rao (2007).

Lee and Mumford move to a slightly different formalism from the one presented above, replacing the conditional probabilities with potential functions $\phi(V_i, V_j)$, which measure the undirected compatibility of the representations in a pair of adjacent areas.

Lee and Mumford propose an inference mechanism that combines belief propagation with another algorithm popular in machine learning and computer vision called particle filtering. Particle filtering replaces the continuous distributions over V_i with a discrete approximation, taking the form of weighted samples: $\sum_{k=1}^{n} w_k \tilde{V}_i^k$ where the $\tilde{V}_i^k$, called particles, are a chosen set of possible activation values. Belief propagation then allows neighboring areas to update these distributions by passing messages. Specifically, each area passes up this message:

$$M_{V_i \to V_{i+1}}\left(\tilde{V}_{i+1}^k\right) = \sum_j M_{V_{i-1} \to V_i}\left(\tilde{V}_i^j\right) \phi(\tilde{V}_i^j, \tilde{V}_{i+1}^k). \quad (17.13)$$

Similarly, the downward massages take the form

$$M_{V_i \to V_{i-1}}\left(\tilde{V}_{i-1}^k\right) = \sum_j M_{V_{i+1} \to V_i}\left(\tilde{V}_i^j\right) \phi(\tilde{V}_i^j, \tilde{V}_{i-1}^k). \quad (17.14)$$

Focusing on the upward case, the kth component of the message $M_{V_i \to V_{i+1}}\left(\tilde{V}_{i+1}^k\right)$ can be understood as the expected compatibility of the particle $\tilde{V}_{i+1}^k$ with the activity in area V_i, where the expectation is taken with respect to incoming messages from the other direction. We have presented the standard version of the algorithm, called the sum product algorithm; Lee and Mumford use a variant called the max-sum algorithm, in which the sum is replaced with a max.

After one round of message passing, each particle $\tilde{V}_i^k$ has received "scores" from the messages coming from above and below. These are multiplied to give the particle's weight:

$$w_i^k = \frac{1}{Z} M_{V_{i+1} \to V_i}\left(\tilde{V}_i^k\right) M_{V_{i-1} \to V_i}\left(\tilde{V}_i^k\right). \quad (17.15)$$

As the model's beliefs evolve, many of the original particles become overwhelmingly unlikely, as manifested in low weights. At each iteration, a new set of particles is formed by resampling with replacement particles according to their weights. This will tend to reinforce particles that are doing well while killing off ones that are doing poorly.

The basic form of the message-passing algorithm, propagation of locally computed information, makes it attractive from a neural point of view. One sticking point, though, is how the particles are represented. Since the algorithm requires each brain area to maintain multiple independent hypotheses (particles) simultaneously, the question arises how they can be kept internally coherent and mutually distinct. Lee and Mumford tentatively suggest a few options such as the synchronous firing, but at present no option constitutes a fully fleshed out proposal.

Comparing this picture with Rao and Ballard's model, presented above, we see that the top-down messages can still be interpreted as predictions, but the bottom-up messages are no longer directly interpretable as error signals. In addition, allows more complex interactions between areas than Rao and Ballard's. This makes the Lee and Mumford model more expressive, but at the cost of less tractable inference.

17.5.5 Feedback Processing and Priors In addition to providing an efficient representation, integrating generative top-down signals with bottom-up ones is an effective way of dealing with the ambiguities inherent in low-level cues. Yuille and Kersten (2006) argue that low-level determinations are difficult to make using only local low-level information but are relatively easy given a high-level analysis. McDermott (2004) asked human subjects to determine whether a junction of edges was present in a given image region. This is easy when the whole image is visible; an object's location determines the locations of its edges. However, when McDermott restricted his subjects' information to low-level cues by making them view the image through a small, 13 ×13 pixel window, their junction-recognition performance was poor.

A prediction, then, is that feedback processing should be especially prevalent when low-level cues are ambiguous or absent. The Kanizsa square (figure 17.7) is a well-known visual illusion, which admits two interpretations: a white square partially obscuring four black circles, or four mutilated circles. Lee and Mumford

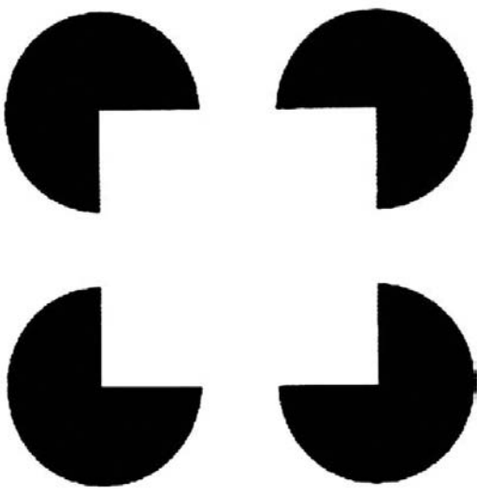

Figure 17.7 The Kanizsa square is an ambiguous stimulus, which can either be interpreted as a white square partially occluding four circles, or as four circles with missing pieces. Feedback processing induces V1 cells to respond to illusory contours defining the square.

recorded from V1 as monkeys viewed this stimulus, finding equally strong responses to the illusory contours between the circles as to the real ones within the circles, suggesting an interpretation in which higher-level processes settled on the "occluding square" interpretation and used feedback connections to fill in the "missing" contour information in V1. Importantly, V1 cells responded to illusory contours after V2 cells, consistent with the feedback story.

In a study using more natural stimuli, Tang et al. (2014) used subdural electrodes to record from the brains of human epilepsy patients undergoing surgery as they viewed two ordinary and heavily occluded natural images. Object recognition was possible in the occluded images—occlusion was calibrated so that subjects achieved approximately 80% accuracy in a behavioral identification task—but most (all but 18%, on average) low-level information was obliterated by the occluders. However, activity in category-selective IT cells was significantly delayed in the occluded condition, suggesting a reliance on preceding recurrent processing.

17.5.6 Feedforward and Feedback Processing: Summing Up Feedforward and feedback models present substantially different pictures of visual processing. How do the two fit together? The mainstream view is of a first, purely feedforward pass sufficient for relatively straightforward classification tasks followed by more sustained feedback processing supporting more difficult or ambiguous determinations.

A first line of evidence for this view comes from the timing of human recognition. In an EEG study, Thorpe et al. (1996) found category-selective information arising in the prefrontal cortex (after visual cortex) in only 150 ms. Historically, this timing has been a challenge for feedback models, given their more extensive processing demands. Lee and Mumford (2003) point out, however, that if recurrent processing occurs continuously in local feedback loops, rather than requiring multiple complete top-to-bottom passes, then recognition at the 150-ms timescale is not unrealistic. More recent time studies pose challenges for this argument, though. A recent behavioral study (Potter et al., 2014) used a rapid serial visual presentation (RSVP) paradigm in which a sequence of images was shown for only 13 ms each, after which subjects were asked whether an image matching a description (e.g., "bear catching fish") was present in the sequence. Even under these extremely demanding time constraints, subjects' performance was significantly better than chance. Short presentation times alone are consistent with feedback models, but the RSVP paradigm poses a challenge: in the time after presentation of the first image, early visual areas are occupied with analysis of subsequent images and are therefore unavailable for feedback processing of the first one.

Other evidence comes from physiology. In a decoding study, Hung et al. (2005), for instance, found that category information was present in the first IT spikes appearing in response to a stimulus, arguing against a feedback picture in which initial IT activity is fed back to and processed by earlier areas before a final interpretation is reached.

As argued above, though, fast, purely feedforward processing is insufficient for many tasks, particularly those with ambiguity, like occlusion. Precise characterization of the tasks supported by the two processing regimes is an important question that awaits future work.

17.6 Conclusion

As we have seen, several well-established frameworks exist for understanding the various processing stages in the ventral visual pathway. However, several important questions remain unanswered. Here we focus on two areas we think will be important for future research.

17.6.1 Digging Deep: Learning From Few Examples While the remarkable developments in computer vision over the last few years owe much to new algorithms and representation schemes, much credit must also go to the availability of huge databases of labeled images, and the computing power required to process them. This state of affairs contrasts markedly with human vision: we can often learn to recognize a new object (say an iPhone circa 2005) from only one example, not hundreds or thousands. Even as children, we hear our parents explicitly identify only a relatively modest number of objects in the world. It is an important challenge, therefore, both for the development of the next generation of computer vision systems and for genuine understanding and replication of the animal visual system, to come up with computational systems that can learn from similarly compact collections of data.

As we have seen, building in invariance to nuisance transformations is one route to data-efficient models. HMAX and other CNNs can build invariance to changes and scale, but we can hope for more gains by accounting for other transformation types. Anselmi et al. (2013) develops a more general theory of invariant representations.

Transformations like translation, scaling, and rotation are class general; they apply to all object classes in

the same way, a property that does not hold for transformations in general. Changes in expression, for instance, are a class of transformation that apply to faces alone. Less obviously, three-dimensional rotation in depth followed by projection onto the two-dimensional image plane is also class-specific: a long, thin object, for instance, rotates differently from an approximately spherical one. Anselmi et al. present a two-stage model to tackle class-specifc transformations. The first stage, corresponding to areas V1–V4, is an HMAX-style CNN. The second stage, corresponding to anterior IT, consists of a collection of modules each of which handles class-specific transformations for one important object class. The face patches in primate brains are hypothesized to be examples of such modules. The resulting architecture achieves competitive performance in face processing tasks and accounts for a number of experimental findings. It predicts, for instance, the existence of the mirror-symmetrical cells found in macaque face areas that respond equally to a face as to an axis-flipped version.

Another approach to improving data efficiency (Salakhutdinov et al., 2011) uses hierarchical probabilistic models. A nonvisual parable conveys the main idea. Suppose you are given a number of opaque bags of colored marbles. You empty out the first one, seeing that it contains only red marbles. Further exploration shows that all the marbles in second are blue, and all in the third are yellow. You then draw a single green marble from the fourth bag. Even without seeing any other bag-four marbles, you can be reasonably sure that they are green: After all, all of the other bags were monochromatic. Examples like this are formulated with hierarchical probability models: the individual bags are probability distributions, in this case distributions over marble colors, whose parameters are themselves drawn from a higher-level distribution. Thus, learning something about a few of the bags—for instance, that they are monochromatic—tells you something about the generative process for bags in general and allows you make hypotheses about new examples. Salakhutdinov et al. transfer this intuition to vision, replacing bags with object categories and marbles with image features. Thus, information about the feature distributions for a few classes is information about how features are distributed for classes in general. As with the single green marble in the example, this higher-level information, combined with a single image from a new class, constrains the new class's structure, making it easier recognize further examples.

17.6.2 Thinking Wide: Full Scene Interpretation in Natural Images We began this chapter with an overview of the range of questions humans can answer about the visual world but have focused the discussion so far on the areas best understood by neuroscience: feedforward object and scene classification and preliminary steps toward full scene interpretation made by feedback models. Here we discuss the prospects for modeling a wider range of visual tasks, ones requiring the analysis of scenes with multiple objects interacting in complex ways. We briefly survey promising computer vision approaches to these problems, approaches that may develop into fruitful sources of neuroscientific hypotheses.

A prerequisite, arguably, for human-level flexibility in visual scene analysis is the ability to recognize all the objects present in an image and to represent their spatial relations. One version of this task is called image parsing (Yao et al., 2010; Zhu and Mumford, 2007; Tu et al., 2005; Jin and Geman, 2006; Zhu et al., 2010; Socher et al., 2011). In language processing, one parses a sentence by recursively breaking it up into meaningful parts: a sentence is composed of phrases, which are themselves composed of smaller phrases, and so on. Images have similar hierarchical structure: an image depicts a scene, which is composed of objects, which are composed of parts, which are composed of primitive components like corners and edges. Given an input image, an image-parsing algorithm seeks to recover this hierarchal structure. Whereas the classification models we have seen so far assign an image a single label, image-parsing algorithms assign an image a whole parse tree.

Several models (Yao et al., 2010; Zhu and Mumford, 2007) define image grammars. Just as a string grammar is a generative model for the well-formed strings in a language, an image grammar is a rich, structured generative model for images. As in the feedback models surveyed above, image parsing is the task of inverting this generative model, going from an image to the sequence of choices that could have formed it. There are two main algorithmic challenges in this approach. In the learning problem, the model must infer the appropriate grammar to describe a collection of input images . In the inference problem, the model must find a parse to explain an image with respect to an already fixed grammar. Given the large number of possible grammars and parse trees, both of these problems are challenging.

Partly in response to these difficulties, other image-parsing approaches such as Zhu et al. (2010) and Jin and Geman (2006) eschew formal grammars, opting for more tractable recursive models amenable to efficient discriminative and dynamic programming algorithms. Socher et al. (2011) uses a neural network approach,

though not one with direct correspondence to the brain. This model takes as input an image oversegmented into regions, from each of which it extracts a feature vector. It then proceeds by merging regions pairwise, continuing until only one region remains in the image. The binary tree resulting from this process is the image's parse.

The heart of the model is a so-called recursive neural network (RNN) that takes as input two feature vectors and outputs a single merged feature vector of the same length. The RNN is used to evaluate candidate region merges. Given neighboring regions R_1 and R_2 with feature vectors F_1 and F_2, the model computes $F_3 = \mathrm{RNN}(F_1, F_2)$. A linear regression model then assigns a score to F_3, which measures the quality of merge. Given a current segmentation of the image into regions, the model considers each possible merge and chooses the one with the highest score, continuing in this way recursively. Each time a new merged region is created, it comes with its feature vector, F_3 above. In addition to being necessary for calculating further merges, this feature vector can be fed into a classifier assigning a category label to region. Thus, each node in the parse tree can be assigned a category.

In the context of generative models like image grammars, one often hears the slogan "vision as inverse graphics," an expression of the Helmholtzian view of perception as unconscious inference. Some promising recent models (Mansinghka et al., 2013; Kulkarni et al., 2015) take this slogan literally, viewing images as arising not from a grammar but from a probabilistic graphics program, in which parameter choices such as object positions, camera angle, and lighting conditions are stochastic. Given a generative model of this kind, an inference algorithm can be used to find the collection of parameter choices that best explain a given input image. While this inference problem is difficult, advantages of the inverse graphics approach include the fact that it can model a scene's three-dimensional structure and can exploit the expressivity and flexibility of a full programming language, potentially obtaining richer scene interpretations than those possible with grammar-like formalisms.

Mapping structured scene interpretation models to neural computation is an important outstanding challenge. In addition the purely visual aspects of this mapping, a full solution will have to address fundamental questions about how the brain represents structured information. In vision, these questions arise in even the simplest visual scenes. Consider, for instance, a ball inside a cup. We know how to represent the ball and the cup by feature vectors, and we may be able to extend the principle to find the vector for "inside" as well. The question, then, is how to combine all of these representations in a way that reflects the scene's structure, preserving, for instance, its distinctness from "cup inside ball." Considering neural representations of more complex objects, such as parse trees, only magnifies the problem. The question of neural representations of structured and relational data, sometimes called the "connectionist variable binding problem," has a long history, and a number of solutions have been proposed (Smolensky and Legendre, 2006; Shastri and Ajjanagadde, 1993; van der Velde and de Kamps, 2006). To date, though, none has emerged as a consensus solution. Open problems abound in this area, both for computational neuroscience in general and for fully structured vision in particular.

Acknowledgments This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., & Poggio, T. (2013). Unsupervised learning of invariant representations in hierarchical architectures. *arXiv preprint arXiv : 1311.4158*.

Balduzzi, D., Vanchinathan, H., & Buhmann, J. (2014). Kickback cuts Backprop's red-tape: Biologically plausible credit assignment in neural networks. *arXiv preprint arXiv : 1411.6191*.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695–711.

Carlson, E. T., Rasquinha, R. J., Zhang, K., & Connor, C. E. (2011). A sparse object coding scheme in area V4. *Current Biology, 21*(4), 288–293.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience, 17*(3), 455–462.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Durbin, R., & Rumelhart, D. E. (1989). Product units: A computationally powerful and biologically plausible extension to backpropagation networks. *Neural Computation, 1*(1), 133–142.

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience, 30*(49), 16601–16608.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598–601.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision (Charlottesville, Va.), 7*(1), 10.

Ffytche, D. H. (2005). Visual hallucinations and the Charles Bonnet syndrome. *Current Psychiatry Reports, 7*(3), 168–179.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–836.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119–130.

Greene, M. R., & Oliva, A. (2006). Natural scene categorization from conjunctions of ecological global properties. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 291–296). Mahwah, NJ: Erlbaum.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102.

Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *Journal of Neuroscience*, *24*(13), 3313–3324.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, *73*(1), 218–226.

Jin, Y., & Geman, S. (2006). Context and hierarchy in a probabilistic image model. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 2145–2152). Los Alamitos, CA: IEEE Computer Society.

Knoblich, U., Bouvrie, J., & Poggio, T. (2007). *Biophysical models of neural computation: Max and tuning circuits* (pp. 164–189). Berlin: Springer.

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience*, *31*(20), 7322–7333.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, *17*(1), 26–49.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. In B. Schölkope, J. Platt, & T. Hofmann (Eds.), Advances in Neural Information Processing Systems (pp. 1097–1105). San Mateo, CA: Morgan Kaufman; Cambridge, MA: MIT Press.

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. 2015 IEEE Conference on Computer Vision and Pattern Recognition.

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, *92*(5), 2704–2713.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448.

Mansinghka, V., Kulkarni, T. D., Perov, Y. N., & Tenenbaum, J. (2013). *Approximate Bayesian image interpretation using generative probabilistic graphics programs*. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), Advances in Neural Information Processing Systems (pp. 1520–1528). San Mateo, CA: Morgan Kaufman; Cambridge, MA: MIT Press.

McDermott, J. (2004). Psychophysics with junctions in real images. *Perception*, *33*(9), 1101–1128.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251.

Murphy, K., Torralba, A., & Freeman, W. T. (2003). Using the forest to see the trees: A graphical model relating features, objects, and scenes. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), (pp. 1499–1506). Advances in Neural Information Processing Systems Cambridge, MA: MIT Press.

Murray, S. O., Schrater, P., & Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. *Neural Networks*, *17*(5), 695–705.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

Park, S., & Chun, M. M. (2009). Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *NeuroImage*, *47*, 1747–1756.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from its spatial boundary: Complementary roles for the PPA and LOC in representing real-world scenes. *Journal of Neuroscience*, *31*(4), 1333–1340.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception & Psychophysics*, *76*(2), 270–279.

Rao, R. P. (2007). Neural models of Bayesian belief propagation. In K. Doya, S. Ishii, A. Pouget, & R. P. N. Rao (Eds.), *Bayesian Brain: Probabilistic Approaches to Neural Coding* (p. 239). Cambridge, MA: MIT Press.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*(19), 2301–2311.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2014). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv : 1409.0575*.

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995.

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011). Learning to share visual appearance for multiclass object detection. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1481–1488). IEEE.

Schmolesky, M. T., Wang, Y., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., et al. (1998). Signal timing across the macaque visual system. *Journal of Neurophysiology*, *79*(6), 3272–3278.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200.

Serre, T., Oliva, A., & Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6424–6429.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*(3), 417–451.

Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 300–312.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv : 1409.1556*.

Smolensky, P., & Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar: Vol. 1. Cognitive Architecture*. Cambridge, MA: MIT Press.

Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 129–136).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2014). Going deeper with convolutions. *arXiv preprint arXiv : 1409.4842*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv : 1312.6199*.

Tang, H., Buia, C., Madhavan, R., Crone, N. E., Madsen, J. R., Anderson, W. S., et al. (2014). Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron*, *83*(3), 736–748.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522.

Torralba, A., Murphy, K. P., & Freeman, W. T. (2010). Using the forest to see the trees: Exploiting context for visual object detection and localization. *Communications of the ACM*, *53*(3), 107–114.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S. C. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, *63*(2), 113–140.

Van der velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–70.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624.

Yao, B. Z., Yang, X., Lin, L., Lee, M. W., & Zhu, S. C. (2010). I2t: Image parsing to text description. *Proceedings of the IEEE*, *98*(8), 1485–1508.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014* (pp. 818–833). Springer.

Zhu, L., Chen, Y., Yuille, A., & Freeman, W. (2010). Latent hierarchical structural learning for object detection. In *2010 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1062–1069). New York: IEEE.

Zhu, S. C., & Mumford, D. (2007). *A Stochastic Grammar of Images*. Hanover, MA: Now.