H

Hierarchical Models of the Visual System

Matthew Ricci and Thomas Serre
Department of Cognitive, Linguistic, and Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, RI, USA

Synonyms

Deep learning architectures; Hubel and Wiesel model; Large-scale models of the visual system; Simple-to-complex hierarchies; Ventral stream

Definition

Hierarchical models of the visual system are neural networks with a layered topology. The receptive fields of units (i.e., the region of visual space to which units respond) at one level of the hierarchy are constructed by combining inputs from units at a lower level. After a few processing stages, small receptive fields tuned to simple stimuli get combined to form larger receptive fields tuned to more complex stimuli. Such an anatomical and functional hierarchical architecture is a hallmark of the organization of the visual system. In feedforward networks, information flows in a bottom-up fashion – from lower to higher processing stages. In feedback networks, information is able to dynamically reenter processing stages via recurrent connections. Feedback connections can be broadly divided between horizontal or lateral connections within processing stages and top-down connections from higher onto lower processing stages.

Since the pioneering work of Hubel and Wiesel (1962), a variety of hierarchical models have been proposed, from relatively small-scale models of the primary visual cortex to very large-scale (system-level) models of object and action recognition, which account for visual processing in entire visual streams. The term "model of the visual system" is generally reserved for architectures that are constrained in some way by the anatomy and the physiology of the visual system (with various degrees of realism). Deep convolutional networks are architecturally similar neural networks that have led to impressive results in a wide range of engineering disciplines, from computer vision to natural language processing, and artificial intelligence more broadly.

Detailed Description

The feedforward flow of visual information in the ventral stream of the visual cortex is carried by ascending projections from the retina, through the lateral geniculate nucleus (LGN) of the thalamus to the primary visual cortex (V1) and extrastriate visual areas, V2 and V4, and culminating in the inferotemporal (IT) cortex. In turn, IT provides a

D. Jaeger, R. Jung (eds.), *Encyclopedia of Computational Neuroscience*,
https://doi.org/10.1007/978-1-4614-7320-6_345-2

major source of input to the prefrontal cortex (PFC), which is involved in linking perception to memory and action (see DiCarlo et al. 2012, for review). A single feedforward pass through the visual cortex rapidly produces a coarse image representation sufficient for rapid object classification. Evolutionarily, such rapid feedforward processing might have been enabled the split-second detection of predators necessary for survival (Thorpe et al. 2001).

The function of feedback/recurrent connections are less well understood. By "feedback," we mean those connections which carry information either from a higher to a lower cortical area (top-down feedback) or within a cortical area (lateral or horizontal feedback). Top-down feedback is mediated by descending projections connecting, for example, V2, V4, and IT to V1 or V1 to LGN (see Pennartz et al. 2019, for a review). These descending projections generally outnumber ascending ones. It is speculated that feedback serves primarily to modulate rather than drive activity in lower areas (Gilbert and Li 2013). Accordingly, circuits with recurrent connections have been implicated in the dynamic maintenance of image representations via attention and working memory (Gazzaley and Nobre 2012) and other modulatory processes (Lamme et al. 1998; Angelucci and Shushruth 2013; Gilbert and Sigman 2007).

The goal of hierarchical models of the visual cortex is to explain how cortical anatomy and physiology, particularly the hierarchical arrangement of visual areas, give rise to complex visual behaviors including object recognition. Such models can be broadly classified into feedforward and feedback varieties and have their origin in computational neuroscience, where biological realism is the primary concern. Recently, however, computer vision models, notably deep convolutional neural networks (CNNs), have emerged as the best predictors of neural activity (see Serre (2019) for a review), despite their being largely unconstrained by biology.

First, we will provide an overview of feedforward hierarchical models of the visual cortex. One of the primary objectives of these models is to explain rapid visual categorization (see Serre (2016) for a review). Consequently, feedforward hierarchical models must address the ability of feedforward cortical pathways to extract features which are both selective for natural object recognition and invariant to irrelevant image transformations, including pose and illumination. The balancing of these two factors is sometimes called the "invariance-selectivity" trade-off (Geman 2006). Second, we will turn to feedback models of visual processing. These network models seek to explain how an initial, coarse feedforward representation can be dynamically manipulated with executive, mnemonic, and other processes. Finally, we will briefly discuss the problem of learning in hierarchical models.

Feedforward Hierarchical Models

Feedforward hierarchical models of the visual system have a long history starting with Marko and Giebel (1970)'s homogeneous multilayered architecture and later Fukushima (1980)'s neocognitron. One of the key principles in the neocognitron and other modern hierarchical models originates from the pioneering physiological studies and models of Hubel and Wiesel (1962). In these networks, the receptive fields of units at one level of the hierarchy are constructed by combining inputs from units at a lower level. Numerous feedforward models of the ventral stream of the visual system have been described since the neocognitron to account for the organization and the neurophysiology of the ventral stream of the visual cortex. These models can be coarsely divided into conceptual proposals (Biederman 1987; Perrett and Oram 1993; Hochstein and Ahissar 2002) and neurobiological models (e.g., Wallis 1997; Mel 1997; Riesenhuber and Poggio 1999; Ullman et al. 2002; Thorpe 2002; Amit and Mascaro 2003; Wersing and Koerner 2003; Serre et al. 2007; Masquelier and Thorpe 2007). Similar hierarchical models have also been proposed to explain motion processing in the dorsal stream of the visual cortex (e.g., Simoncelli and Heeger 1998; Grossberg et al. 1999; Perrone and Thiele 2002; Giese and Poggio 2003; Rust et al. 2006;

Jhuang et al. 2007; Pack and Born 2008; Mineault et al. 2012).

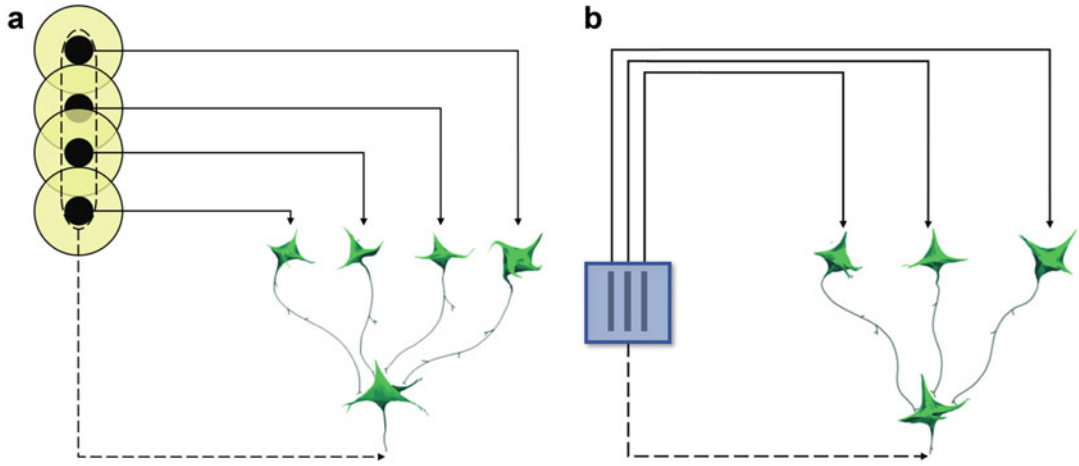
Somewhat independently, convolutional neural networks (CNNs) and other deep learning architectures have been developed in computer vision (see LeCun et al. 2015, for review). CNNs typically involve two processing stages: a feature extraction stage, in which visual information is passed through layers of computing units having small receptive fields, and a classification stage, in which the resultant features are combined by "fully connected" units having global receptive fields. In the most common computer vision setting, the parameters of these models are gradually adjusted by the backpropagation algorithm to minimize classification error on a large data set of labeled images. After supervised training, the network approximates a function from the space of natural images to a set of discrete object labels. These neural networks do not mimic the organization of the visual system in detail, but biology is often cited as a source of inspiration, and they nevertheless provide a better fit to experimental data than earlier biological models (Yamins et al. 2014).

Modeling of feedforward pathways in the visual cortex begins with the description of primary visual cortex by Hubel and Wiesel (1962). Hubel and Wiesel famously described two functional classes of cortical cells: simple cells, which respond to oriented stimuli (e.g., bars, edges, gratings) at particular orientations, and complex cells, which, while also tuned to oriented stimuli, tend to have larger receptive fields and exhibit some tolerance to the exact position of the stimulus within their receptive fields. A V1-like circuit is shown in Fig. 1, which connects a complex cell to an array of simple cells tuned to the same feature at neighboring locations.

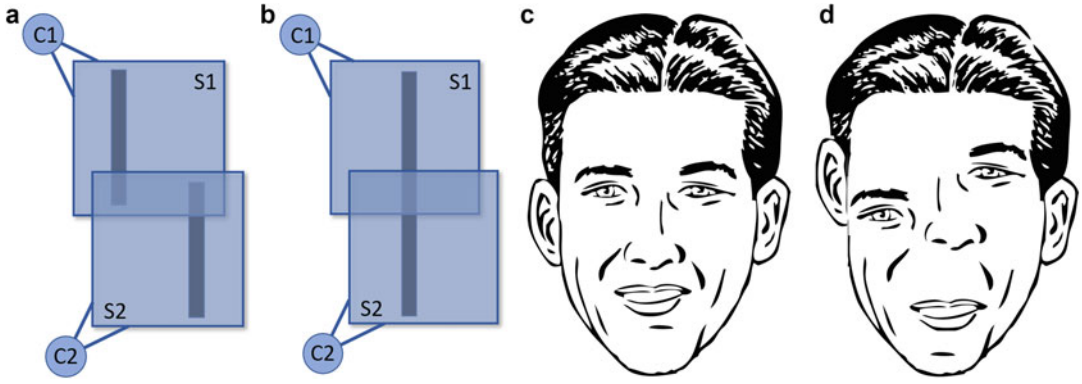
Though such models help to explain the basic invariance properties of key neural circuits, they do not explain the careful balance of invariance and selectivity needed to solve the problem of natural object recognition. To understand the limitation of these circuits more clearly, consider the following classic example of Geman (2006) (Fig. 2). A collection of simple cells are tuned to vertical bars. One subset, S1, of these cells feeds into complex cell, C1, and an overlapping subset, S2, feeds into complex cell C2. One of the purposed functions of complex cells in the visual system is the gradual introduction of invariance, since, for example, C1 and C2 will be as active when presented with two bars in the configuration of Fig. 2a as they will when presented with the configuration of Fig. 2b. In other words, the activation patterns of C1 and C2 cannot distinguish between a broken and a continuous line. This problem generalizes: a V1-like circuit consisting of a few face features with tolerance to small translations will tend to detect faces (Geman 2006) even when it should not (Fig. 2c vs. 2d). Such a circuit appears to be insufficient to fully achieve selective and invariant object recognition.

Contemporary feedforward models of the visual cortex partially resolve this "invariance-selectivity" dilemma by introducing further, "deeper," processing stages, each with numerous learned filters. These models are an extension of the Hubel and Wiesel's circuit from V1 to higher areas of the ventral stream. They come in numerous forms, differing in their specific wiring and neuronal operations. However, common to most of these network models are three key ingredients: (1) a cascade of linear filters each followed by (2) a pointwise nonlinearity which introduces tolerance to noise and (3) gradual, intermittent max/average pooling for translation invariance (see Mallat 2016, for a theoretical justification of these three parts). The result is an architecture which gradually builds hierarchical compositions of visual features up to and including natural objects. Such a feature hierarchy is depicted for the case of Hmax, a classical hierarchical model of visual processing (Riesenhuber and Poggio 1999; Serre et al. 2007), in Fig. 3.

A general wiring diagram of a feedforward model of the visual system is shown in Fig. 4. A layer of simple cells is conceived as a bank of feature detectors, each selecting for one of K different features within a local neighborhood centered at spatial location, u, in the layer's input (see Ullman and Soloviev 1999, for a discussion of the biological realism of this architecture). Each simple cell computes a weighted sum of afferent unit activities,



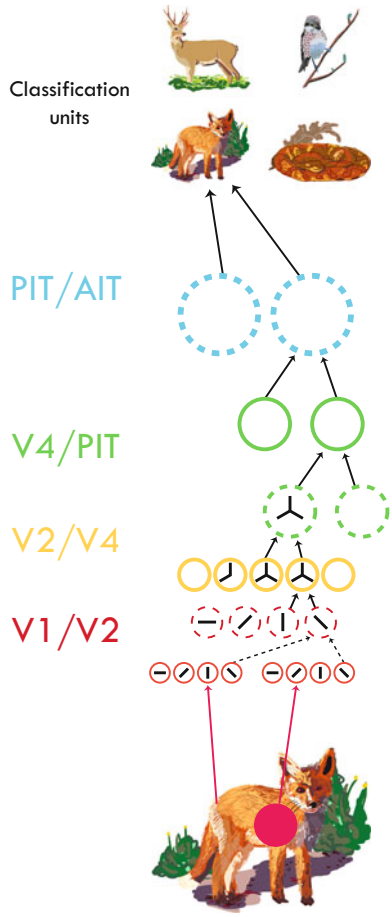
Hierarchical Models of the Visual System, Fig. 1 Hubel and Wiesel model. (**a**) Receptive field (RF) of a simple cell (lower green cell) obtained by selectively pooling over afferent center-surround cells (upper green cells) aligned along a preferred axis of orientation (vertical shown here). (**b**) At the next stage, a complex cell (lower green cell) RF can be obtained by selectively pooling over afferent simple cells (upper green cells) with the same preferred orientation (vertical). Shown here is a complex cell RF obtained by pooling over position to build tolerance to translation of the preferred stimulus, but a more complete model of a complex cell would also include pooling over simple cells tuned to slightly different spatial frequency and phases (Rust et al. 2005; Chen et al. 2007). (Modified from Hubel and Wiesel (1962))



Hierarchical Models of the Visual System, Fig. 2 The invariance-selectivity trade-off. Two complex cells C1 and C2 pool over overlapping sets of simple cells tuned to vertical bars. Being translation invariant, the cells respond identically to the broken bars in (**a**) as they to the continuous bar in (**b**). Similarly, a collection of complex cells tuned to face parts may erroneously detect a full face (**c**) if facial components are subject to small translations (**d**). See Geman (2006) for details. Arguably, the visual system avoids these false alarms for thousands of fine-grained object categories while remaining invariant to irrelevant image nuisances, like illumination, contrast, and pose.

$$y_u^{(k)} = \sum_{v \in \Omega_u} w_v^{(k)} x_v, \qquad (1)$$

where k indexes the particular feature to which the simple cell is tuned, Ω_u are the locations of afferent units in the input pool, x_v are those units' activities, and w_v are the synaptic weights connecting each afferent to the simple cell target. The scalar $y_u^{(k)}$ is analogous to an average firing rate of the simple cell. Since feature detectors are replicated at each spatial location, the output of a simple cell layer can be modeled as a convolution of the afferent signal, x, with a bank of features $\{w^{(k)}\}_{k=1}^{K}$ of small support:



Hierarchical Models of the Visual System, Fig. 3 Sketch of the Hmax hierarchical model of visual processing. Acronyms: V1, V2, and V4 correspond to primary, second, and fourth visual areas and PIT and AIT to posterior and anterior inferotemporal (IT) areas, respectively (tentative mapping with areas of the visual cortex shown in color, parietal cortex and dorsal stream not shown). The model relies on two types of neural operations: a max-like pooling operation (shown in dash circles) over similar features at different positions and scales to gradually build tolerance to affine transformations and a convolution (also called tuning) operation (shown in plain circle) over multiple features to increase the complexity of the underlying representation. Since it was originally developed Riesenhuber and Poggio (1999), the model was shown to explain a number of new experimental data (see Serre and Poggio 2010, for a review). However, today this model and similar architectures have been superseded by deep convolutional networks that have been shown to provide a better fit to neural data along the ventral stream of the visual cortex (Serre 2019)

$$y^{(k)} = w^{(k)} \star \mathrm{x}.$$

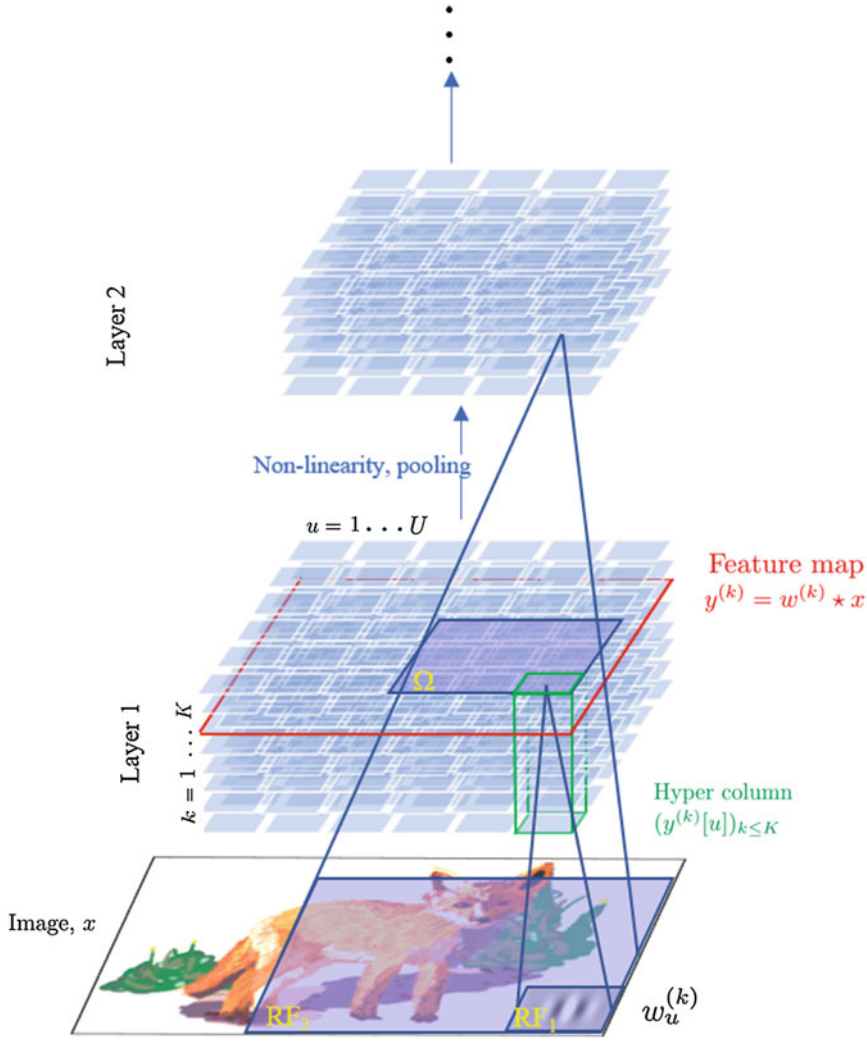
The pattern of activations $y^{(k)}$ for a given k is called a feature map (Fig. 4, red rectangle). Note that we have written the output of a layer of simple cells as a 2D convolution, though, in general, the summation can be taken across many dimensions, including time, input channels, eye dominance, etc.

The vector $\left(y_u^{(k)}\right)_{k=1}^{K}$ for a fixed location, u, is called a hypercolumn (Fig. 4, green prism) and represents an encoding of the afferent activity with respect to the features $\{w^{(k)}\}$. In the primary visual cortex, these local descriptors are well modeled by Gabor wavelets (as shown in RF_1 in Fig. 4), tuned to different orientations and scales, and these filters often emerge in models trained on natural images, especially when activity in the network is encouraged to be sparse (Olshausen and Field 1996; Lee et al. 2008). Iteratively combining the supports of filters throughout the hierarchy which feed into a given unit yields its receptive field; e.g., in Fig. 4, the shaded region RF2 is the union of receptive fields of its afferent units in Ω. Comparing the size of RF_1 and RF_2, we see that iterative pooling in this manner gradually increases the size of receptive fields as one proceeds up the feature hierarchy.

The output of a simple cell layer is typically passed through a pointwise nonlinearity, ρ, called an activation function in order to provide some robustness to noise and increase the expressiveness of the visual representation:

$$\begin{aligned} \widehat{y}^{(k)} &= \rho\left(y^{(k)}\right) \\ &= \rho\left(w^{(k)} \star \mathrm{x}\right). \end{aligned} \tag{2}$$

Typical choices for ρ include rectification, logistic, or hyperbolic functions. The standard choice in computer vision for ρ is zero rectification. In this context, a scalar bias, b, on the output of a simple cell layer functions as a spiking threshold, since $y_u = \max(0, (w*x)_u - b) > 0$ only when $(w*x)_u > b$. A model neuron which passes a linear combination of its afferents through a pointwise



Hierarchical Models of the Visual System, Fig. 4 Feedforward wiring diagram. Hierarchical models of the visual system are characterized by multiple stages of processing whereby units in one stage (shown as squares) pool over the response of units from the previous stage. Each stage computes a convolution of the previous stage with a bank of K filters, and a unit's activity is the output of this convolution at a given location. The convolution with a given filter is called a *feature map* (e.g., the outputs of cells in the red square). The set of feature activations at a given location is called a *hypercolumn*. The output of a layer is passed through a pointwise nonlinearity and max or average pooling, after which it forms the input for another bank of filters. See text for details

nonlinearity is called a linear-nonlinear (LN) neuron (Simoncelli et al. 2004). The LN model has been shown to account for a host of experimental data (Rieke et al. 1997), and it has been shown that in many cases, biophysically more realistic, spiking neuron models can be reduced to a simple LN cascade (Ostojic and Brunel 2011).

Extensions of the LN cascade include the addition of a normalization stage (Heeger 1993; Carandini and Heeger 1994), in which the response y_u of the neuron is divided by a factor that typically includes the summed activity of a pool of neurons:

$$y_u = \frac{\sum\limits_{v \in \Omega_u} (w_v x_v)^p}{\epsilon + \left(\sum\limits_{v' \in \Omega'_u} x_{v'}^q \right)^r}, \tag{3}$$

where $\epsilon \ll 1$ is a constant to avoid zero-division. The pool of neurons Ω'_u used for normalization may correspond to the same pool of neurons Ω_u which shape the classical receptive field or may extend beyond to account for extra-classical receptive field effects (Series et al. 2003). Normalization circuits were originally proposed to explain the contrast response of cells in the primary visual cortex and are now thought to operate throughout the visual system and in many other sensory modalities and brain regions (see Carandini and Heeger 2012, for a review).

Complex cell layers filter their input with a fixed low-pass or max filter, the latter of which is the standard in computer vision. If a max-pooling unit at location u receives input from a pool of afferent units Ω_u, then the unit outputs

$$y_u = \max_{v \in \Omega_u} \{x_v\}, \tag{4}$$

where x is itself usually the activity of a simple cell layer. Interestingly, maximum, Gaussian, sigmoid, and other types of cell tuning can all be approximated by various forms of Eq. 3 depending on the values of the static nonlinearities p, q, and r in the underlying neural circuit. By adjusting these nonlinearities, Eq. 3 can approximate better a maximum or other tuning functions (see Kouh and Poggio 2008, for details).

While recent work has suggested that simple and complex cells may represent the two ends of a continuum instead of two discrete classes of neurons (see Ringach (2004) for a discussion), this dichotomy is probably not critical for hierarchical models of the visual system. Indeed, some computational models do not distinguish between simple and complex cell pooling (O'Reilly et al. 2013).

Deep convolutional neural networks with an architecture similar to that of Fig. 4 have achieved impressive results in image classification, by some measures surpassing human performance (He et al. 2015). Many of the most powerful contemporary networks are "ultra-deep," extending the linear-nonlinear cascade to hundreds or even thousands of layers (He et al. 2016). Often, these models are only superficially deep, as they employ "skip" connections to bypass several layers, similar to the ascending projections from V2 to IT or V1 to V4 (Nakamura et al. 1993). During training, these "residual networks" (He et al. 2015) essentially learn to adjust their processing depth to best suit a given task. For example, imagine that pairs of linear-nonlinear stages are grouped into blocks indexed by t, so that the output of the tth block is given by the nonlinear operator A_t:

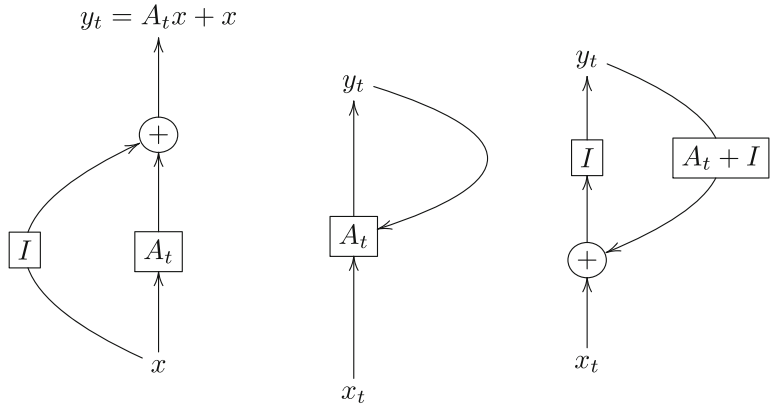
$$A_t x = \rho(w_{t,2} \star \rho(\mathbf{w_{t,1}} \star \mathbf{x})), \tag{5}$$

where $w_{t,1}$ and $w_{t,2}$ are two kernels of synaptic weights interleaved by pointwise nonlinearities. Skip connections add the input to this operator to the output: $A_t x + x$. Suppose that the true function to be learned by this block is $F(x)$. In an ultra-deep system, the optimal function for this block may very well be the identity map, $F(x) = x$. Bypass connections in residual nets (Fig. 5a) accelerate the learning of these identity maps since they allow the training algorithm to alternatively learn zero "residual" functions

$$F(x) - x = A_t x. \tag{6}$$

After training, many of the A_t will have converged to the zero operator, allowing data to essentially skip a layer as the network performs a feedforward pass.

Despite being unconstrained by neurophysiology, currently the best feedforward models of visual cortical activity are deep convolutional networks developed by the computer vision community. Early computational neuroscience work started with AlexNet (Krizhevsky et al. 2012) and ZFNet (Zeiler and Fergus 2014), which were shown to improve the fit to neural data in intermediate and higher areas of the ventral stream of the visual cortex even if they were optimized for task performance (e.g., image classification



(a) Feedforward residual circuit (b) Standard RNN circuit (c) Recurrent residual circuit

Hierarchical Models of the Visual System, Fig. 5 Residual and recurrent circuits. (**a**) A residual net contains skip connections that allow a copy of activity x to bypass the tth block of layers, here encoded in the operator A_t. The copied activity, x, is then added to the output of the bypassed layers, A_{tx}, in the hopes that the model will more easily learn that the A_t operator is often the zero operator. (**b**) A recurrent circuit runs in time and updates its persistent state by integrating incoming information: $y_t = A_t y_{t-1} + x_t$. (**c**) A residual net can be reformulated as a particular recurrent network (Liao and Poggio 2016), where the tth residual block is the state of the recurrent circuit at time t.

accuracy) instead of neural prediction directly (Cadieu et al. 2014; Khaligh-Razavi and Kriegeskorte 2014). A recent study (Cadena et al. 2019) has shown that intermediate layers from the VGG network of Simonyan and Zisserman (2014) provide a better fit to V1 monkey electrophysiology data compared to simpler linear-nonlinear models. Work by Hong et al. (2016) has also shown that multiple image properties beyond object categories (e.g., object position, 3D size, and pose) remain relatively well encoded in higher processing stages in both neural and CNN representations. This provides further evidence that visual hierarchies are able to learn visual representations that are invariant to task-irrelevant transformations while maintaining information for task-relevant ones.

Further evidence for hierarchical processing in object recognition comes from studies that have shown that the depth of convolutional layers that provides the best goodness of fit with brain data increases along the ventral visual stream (Cichy et al. 2016; Devereux et al. 2018; Guclu and van Gerven 2015; Kalfas et al. 2017; Khaligh-Razavi and Kriegeskorte 2014; Yamins et al. 2014). Similar results were also reported for scene recognition (Cichy et al. 2017; Greene and Hansen 2018) and action recognition (Güçlü and Gerven 2017) with spatiotemporal CNNs trained for action recognition (Tran et al. 2015). Interestingly, the goodness of fit between brain data and fully connected layers tends to be lower than with convolutional layers (Kalfas et al. 2017), a result consistent with a behavioral study that has compared CNNs with human behavioral decisions during a rapid categorization task (Eberhardt et al. 2016). This result also seems consistent with a recent object naming study (Devereux et al. 2018) that has shown that a network model of semantics, explicitly trained to learn a mapping from the convolutional layers of a CNN onto object semantic attributes, was better able to explain functional magnetic resonance imaging (fMRI) activation patterns in higher visual areas compared to either convolutional or fully connected layers. CNNs have also been used to synthesize patterns of fMRI activations, which were then used to reproduce classic functional brain-mapping experiments, from recovering retinotopic maps in early visual areas to replicating the known faces-versus-places BOLD contrast in higher areas (Eickenberg et al. 2017).

Recurrent Models

To date, most existing hierarchical models of visual processing – from the perspectives of both biological and machine vision – are instances of feedforward models (Serre 2019). These models have been useful in exploring the power of fixed hierarchical organization as originally suggested by Hubel and Wiesel (1962). However, the limitations of feedforward networks, in terms of their correspondence to both cortical function and anatomy, are becoming increasingly obvious. For example, convolutional networks can be easily tricked into incorrectly detecting objects in ways that biological vision cannot, either by placing familiar objects in unfamiliar arrangements (Fig. 6a; see Rosenfeld et al. (2018)) or even by adding minute perturbations to images (Szegedy et al. 2013; Geirhos et al. 2018). Evidently, deep feedforward nets have not fully resolved the invariance-selectivity dilemma and its attendant high false alarm rate. These networks also struggle to learn simple visual reasoning tasks which are trivial for humans, notably those in which arbitrary objects must be compared (Fig. 6b; see Kim et al. 2018). This limitation suggests feedforward models of the visual system cannot easily emulate humans' ability to flexibly construct innumerable structured descriptions of the visual world (Fodor and Pylyshyn 1988; Geman et al. 2015). Naturally, feedforward models do not help explain the copious feedback and lateral connections in the visual cortex or the executive and mnemonic processes they support. Contemporary CNNs can also have hundreds of processing stages, counting skip connections, whereas the ventral stream contains a dozen or fewer. In fact, the classification accuracy of feedforward networks begins to decorrelate with human behavior after a few layers (Eberhardt et al. 2016).

Could the functional and anatomical limitations of feedforward models of the visual system be connected? Geman (2006), for instance, claims that selectivity in natural vision arises not from the learning of a large dictionary of complex visual features but rather from the dynamical construction of representations out of simple parts using recurrent connections. Recent work in recurrent neural networks (RNNs) has partially corroborated this intuition and shown that models with feedback connections are both functionally superior to feedforward networks on some tasks (Linsley et al. 2018) and better predictors of cortical activity (Nayebi et al. 2018; Kar et al. 2019).

An RNN is built from circuits with cyclical connections. For example, if x_t is a signal at time t, then such a circuit could compute

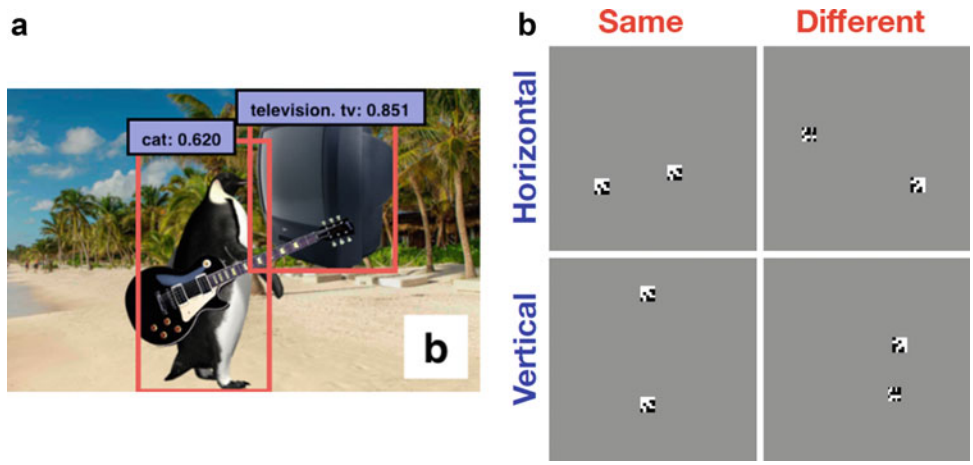
$$y_t = A_t y_{t-1} + x_t, \tag{7}$$

where A_t is a potentially nonlinear, time-dependent operator (Fig. 5b). y_t is a persistent or hidden state of the circuit which accumulates information from x_t over time. An additional operator can provide a readout or classification layer: $o_t = B_t y_t$. The principal advantage of networks built from these circuits is that the system can perform stateful computations based on the information stored in the persistent activity y_t. For example, given a memory-data combination, (y_t, x_t), a network could learn to sequentially suppress or enhance different regions of the visual field, giving rise to exactly the kind of attentional mechanisms capable of disambiguating a cluttered scene (Treisman and Gelade 1980; Evans and Treisman 2005), as in Fig. 6a, or detecting the relations in Fig. 6b (Donderi and Zelnicker 1969; Clevenger and Hummel 2014).

Importantly, a version of the dynamical system in Eq. 7 is equivalent to the ultra-deep residual networks discussed above. If $x = x_0$, $x_t = 0$ when $t > 0$, $y_0 = 0$, and we consider the circuit in Fig. 5c with operator $A_t + I$, then we find

$$\begin{aligned} y_1 &= (A_1 + I)0 + x_0 = x \\ y_2 &= (A_2 + I)y_1 + 0 = A_2 x + x \\ y_3 &= A_3 y_2 + y_2 \\ &\vdots \end{aligned}$$

We observe that the operator A_t corresponds to the function embodied by blocks in a residual network, so that the tth block in the hierarchy is also the tth sequential operation in the recurrent circuit of Fig. 5c. In other words, an ultra-deep residual network can be "folded" into a shallow



Hierarchical Models of the Visual System, Fig. 6 Limitations of feedforward models. (**a**) A feedforward model can be trained to high accuracy on hundreds of object classes, and yet the resulting network still misclassifies objects when placed in unfamiliar arrangements. Here, the YOLO object recognition and localization deep network (Redmon et al. 2016) mistakes a penguin for a cat because of a partially occluding guitar. (See Rosenfeld et al. (2018) for a systematic analysis of this effect; image credit: Junkyung Kim.) (**b**) Kim et al. (2018) found that feedforward models could easily solve spatial reasoning problems, like determining whether two random patterns are arranged vertically or horizontally (vertical axis), but had a comparatively more difficult time solving same-different tasks such as determining

recurrent circuit (see Liao and Poggio 2016, for details). This suggests that the relatively shallow visual cortex could match the performance of ultra-deep feedforward systems by using recurrent connections through time.

Contemporary recurrent models of the visual cortex elaborate on the basic RNN structure by hierarchically stacking recurrent circuits, introducing lateral connections within a layer and top-down connections between layers, and allowing the system to gate activity. This last operation, known primarily for its use in long short-term memory (LSTM) networks (Hochreiter et al. 1997) and, later, gated recurrent units (GRUs) (Cho et al. 2014) for natural language processing, enables a network to selectively delete and append information to its persistent state. Linsley et al. (2018) recently developed a recurrent network model with lateral connectivity and gating to mimic the contour integration and object segmentation capabilities of humans in cluttered scenes (Grossberg and Mingolla 1985; Field et al. 1993; Grossberg and Raizada 2000; Grossberg and Williamson 2001). Further, Nayebi et al. (2018) and Kar et al. (2019) found that convolutional networks augmented with recurrent connections had improved performance in image classification and could predict recordings of primate IT activity better than a feedforward baseline. These attempts to model the visual system with recurrent networks are promising, though this line of research is still in its infancy compared to feedforward modeling.

Learning

Modeling the visual system with recurrent networks is a tantalizing proposition, since RNNs are Turing-equivalent and can therefore theoretically compute any function given enough time (Hyötyniemi 1996). This is to say nothing of how this function is learned, however. Learning is the domain where differences between computational and biological models are at their starkest. In particular, CNNs are typically trained by supervision with gradient descent on millions of labeled images, whereas training in the visual system likely occurs by some combination of unsupervised and reinforcement learning.

Differences with biological vision seem all the greater in light of recent evidence indicating that face preference (Reid et al. 2017), category selectivity (van den Hurk et al. 2017), and basic reasoning (Martinho and Kacelnik 2016) can require little to no visual experience to function properly.

Even the basic mechanisms of error propagation seem difficult to implement in a biologically plausible manner. Backpropagation of error has traditionally been considered biologically unrealistic since it would seem to require that feedforward connections be mirrored by identical feedback connections. Any mechanism for computing error in cortex would have to know the weights of the feedforward pathway, a dilemma known as the "weight transport" or "weight symmetry" problem (Grossberg and Mingolla 1987). However, Liao et al. (2015) found that feedforward and feedback weights need not be symmetric for successful training as long as the network was trained with batch normalization (Ioffe and Szegedy 2015) and the magnitudes of gradient updates were discarded. Further, Bengio et al. (2015) argues that spike-timing-dependent plasticity (STDP; see Sjöström and Gerstner (2010)) is equivalent to gradient descent for a particular variational algorithm.

Though these results assuage some fears about basic error propagation in the cortical setting, they do not address the implausibility of direct supervision. A complete computational model of the visual system will have to integrate the feedforward and feedback mechanisms discussed above with methods from unsupervised learning, for example, predictive coding (Rao and Ballard 1999). See Marblestone et al. (2016) for an extensive review on the neuroscientific implementation of deep learning algorithms.

Cross-References

- ▶ Deep Learning Network
- ▶ Feedforward Network
- ▶ Recurrent Network

References

Amit Y, Mascaro M (2003) An integrated network for invariant visual detection and recognition. Vis Res 43(19):2073–2088

Angelucci A, Shushruth S (2013) Beyond the Classical Receptive Field: Surround Modulation in Primary Visual Cortex . In J. S. Werner L. M. Chalupa (Eds.), The New Visual Neurosciences (pp. 425–444). Cambridge: MIT Press.

Bengio Y, Lee D-H, Bornschein J, Lin Z (2015) Towards biologically plausible deep learning. Learning. arXiv:1502.04156 [cs.LG]

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94(2):115–147

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, Ecker AS (2019) Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput Biol 15(4):e1006897

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10(12):e1003963

Carandini M, Heeger DJ (1994) Summation and division by neurons in primate visual cortex. Science 264:1333–1336

Carandini M, Heeger DJ (2012) Normalization as a canonical neural computation. Nature Reviews Neuroscience 13(1):51–62. https://doi.org/10.1038/nrn3136

Chen X, Han F, Poo M-m, Dan Y (2007) Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). Proc Natl Acad Sci 104(48):19120–19125

Cho K, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder decoder for statistical machine translation

Cichy RM, Khosla A, Pantazis D, Torralba A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6:27755

Cichy RM, Khosla A, Pantazis D, Oliva A (2017) Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. NeuroImage 153:346–358

Clevenger PE, Hummel JE (2014) Working memory for relations among objects. Atten Percept Psychophys 76:1933–1953

Devereux BJ, Clarke A, Tyler LK (2018) Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Sci Rep 8:10636

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73(3):415–434

Donderi DONC, Zelnicker D (1969) Parallel processing in visual same-different. Percept Psychophys 5(4):197–200

Eberhardt S, Cader J, Serre T (2016) How deep is the feature analysis underlying rapid visual categorization ? In: Lee D, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Neural information processing systems. Curran Associates, Red Hook, pp 1100–1108

Eickenberg M, Gramfort A, Varoquaux G, Thirion B (2017) Seeing it all: convolutional network layers map the function of the human visual system. NeuroImage 152:184–194

Evans KK, Treisman A (2005) Perception of objects in natural scenes: is it really attention free? J Exp Psychol Hum Percept Perform 31(6):1476–1492

Field DJ, Hayes A, Hess RF (1993) Contour integration by the human visual system: evidence for a local "association field". Vis Res 33(2):173–193

Fodor JA, Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. Cognition 28(1–2):3–71

Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36:193–202

Gazzaley A, Nobre AC (2012) Top-down modulation: bridging selective attention and working memory. Trends Cogn Sci 16(2):129–135

Geirhos R, Temme CRM, Rauber J, Schütt HH, Bethge M, Wichmann FA (2018) Generalisation in humans and deep neural networks. In: NeurIPS. Curran Associates, Red Hook

Geman S (2006) Invariance and selectivity in the ventral visual pathway. J Physiol Paris 100(4):212–224

Geman D, Geman S, Hallonquist N, Younes L (2015) Visual Turing test for computer vision systems. Proc Natl Acad Sci 112(12):3618–3623

Giese MA, Poggio T (2003) Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci 4(3):179–192

Gilbert CD, Li W (2013) Top-down influences on visual processing. Nat Rev Neurosci 14(5):350–363

Gilbert CD, Sigman M (2007) Brain states: top-down influences in sensory processing. Neuron 54(5):677–696

Greene MR, Hansen BC (2018) Shared spatiotemporal category representations in biological and artificial deep neural networks. PLoS Comput Biol 14(7)

Grossberg S, Mingolla E (1985) Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. Psychol Rev 92(2):173–211

Grossberg S, Mingolla E (1987) Neural dynamics of surface perception: boundary webs, illuminants, and shape-from-shading. Comput Vis Graphics Image Process 37(1):116–165

Grossberg S, Raizada RD (2000) Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. Vis Res 40(10–12):1413–1432

Grossberg S, Williamson JR (2001) A neural model of how horizontal and interlaminar connections of visual cortex develop into adult circuits that carry out perceptual grouping and learning. Cereb Cortex 11(1):37–58

Grossberg S, Mingolla E, Pack C (1999) A neural model of motion processing and visual navigation by cortical area MST. Cereb Cortex 9(8):878–895

Güçlü U, Gerven MAJV (2017) Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage 145:329–336

Guclu U, van Gerven MAJ (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci 35(27):10005–10014

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Computer Vision and Pattern Recognition; Artificial Intelligence; Learning, Santiago, Chile, IEEE, pp 2026–1034. Retrieved from http://arxiv.org/abs/1502.01852

He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. CoRR, abs/1603.05027

Heeger DJ (1993) Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. J Neurophysiol 70(5):1885–1898

Hochreiter S, Hochreiter S, Schmidhuber J, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hochstein S, Ahissar M (2002) View from the top: hierarchies and reverse hierarchies in the visual system. Neuron 36(5):791–804

Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. Nat Neurosci 19(4):613–622

Hubel D, Wiesel T (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154

Hyötyniemi H (1996) Turing Machines are Recurrent Neural Networks. In Alander J, Honkela T, Jakobsson M (eds), STeP'96 Genes, Nets and Symbols. Vaasa: The Finnish Artificial Intelligence Society, pp 13–24. Retrieved from http://lipas.uwasa.fi/stes/step96/step96/hyotyniemi1/

Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML'15: Proceedings of the 32nd International Conference on Machine Learning (pp. 448–456). Lille, France, Proceedings of Machine Learning Research

Jhuang, H., Serre, T., Wolf, L., Poggio, T. (2007). A biologically inspired system for action recognition. In Proceedings of the Eleventh IEEE International Conference on Computer Vision (pp. 1–8). Rio de Janiero, Brazil, IEEE. Retrieved from http://www.cnbc.cmu.edu/cns/papers/Jhuang_etal_iccv07.pdf https://arxiv.org/pdf/1811.09716.pdf

Kalfas I, Kumar S, Vogels R (2017) Shape selectivity of middle superior temporal sulcus body patch neurons. eNeuro 4(3):0113–0117

Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nature Neuroscience 22(6):974–983. https://doi.org/10.1038/s41593-019-0392-5

Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 10(11): e1003915

Kim J, Ricci M, Serre T, Serre T (2018) Not-So-CLEVR: learning same different relations strains feedforward neural networks. Interface Focus 8:2018011

Kouh M, Poggio T (2008) A canonical neural circuit for cortical nonlinear operations. Neural Comput 20(6):1427–1451

Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. In: Neural information processing system, Lake Tahoe

Lamme VAF, Supèr H, Spekreijse H (1998) Feedforward, horizontal, and feedback processing in the visual cortex. Curr Opin Neurobiol 8(4):529–535

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lee H, Ng AY (2008) Sparse deep belief net model for visual area V2. In: Advances in Neural Information Processing Systems 20. Vancouver, Curran Associates, pp 873–880. https://doi.org/10.1.1.120.9887

Liao Q, Poggio T (2016) Bridging the gaps between residual learning, recurrent neural networks and visual cortex. Technical report, Massachusetts Institute of Technology

Liao Q, Leibo JZ, Poggio T (2015) How important is weight symmetry in backpropagation? Technical report 36

Linsley D, Kim J, Veerabadran V, Windolf C, Serre T (2018) Learning long-range spatial dependencies with horizontal gated recurrent units. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Neural information processing systems. Red Hook, Curran Associates, pp 152–164

Mallat S (2016) Understanding deep convolutional networks. Phil Trans R Soc A 374(20150203):1–17

Marblestone AH, Wayne G, Kording KP (2016) Toward an integration of deep learning and neuroscience. Front Comput Neurosci 10:1–41

Marko H, Giebel H (1970) Recognition of handwritten characters with a system of homogeneous layers. Nachr Z 23:455–459

Martinho A III, Kacelnik A (2016) Ducklings imprint on the relational concept of same or different. Science 353(6296):286–288

Masquelier T, Thorpe SJ (2007) Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Comput Biol 3(2):e31

Mel BW (1997) SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. Neural Comput 9:777–804

Mineault P, Khawaja F, Butts D, Pack C (2012) Hierarchical processing of complex motion along the primate dorsal visual pathway. Proc Natl Acad Sci 109(16): E972–E980

Nakamura H, Gattass R, Desimone R, Ungerleider LG (1993) The modular organization of projections areas V4 and TEO in macaques from areas VI and V2 to. The Journal of Neuroscience 13(9):3681–3691

Nayebi A, Bear D, Kubilius J, Kar K, Ganguli S, Sussillo D, DiCarlo JJ, Yamins DLK (2018) Task-driven convolutional recurrent models of the visual system. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Neural information processing systems. Curran Associates, Red Hook

O'Reilly RC, Wyatte D, Herd S, Mingus B, Jilk DJ (2013) Recurrent processing during object recognition. Front Psychol 4:1–14

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583):607–609

Ostojic S, Brunel N (2011) From spiking neuron models to linear-nonlinear models. PLoS Comput Biol 7(1): e1001056

Pack CC, Born RT (2008) Cortical mechanisms for the integration of visual motion. Elsevier, Oxford

Pennartz CMA, Dora S, Muckli L, Lorteije JAM (2019) Towards a unified view on pathways and functions of neural recurrent processing. Trends Neurosci 42:1–15

Perrett D, Oram M (1993) Neurophysiology of shape processing. Image Vis Comput 11(6):317–333

Perrone JA, Thiele A (2002) A model of speed tuning in MT neurons. Vis Res 42(8):1035–1051

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, IEEE. https://doi.org/10.1016/j.nima.2015.05.028

Reid VM, Dunn K, Young RJ, Amu J, Donovan T, Reissland N (2017) The human fetus preferentially engages with face-like visual stimuli. Curr Biol 27(12):1825–1828.e3

Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1997) Spikes: exploring the neural code. MIT Press, Cambridge, MA

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2(11):1019–1025

Ringach DL (2004) Mapping receptive fields in primary visual cortex. J Physiol 558(3):717–728

Rosenfeld, A., Zemel, R., Tsotsos, J. K. (2018). The Elephant in the Room. arXiv:1808.03305v1 [cs.CV]
Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005) Spatiotemporal elements of macaque v1 receptive fields. Neuron 46(6):945–956
Rust NC, Mante V, Simoncelli EP, Movshon JA (2006) How MT cells analyze the motion of visual patterns. Nat Neurosci 9(11):1421–1431
Series P, Lorenceau J, Frégnac Y (2003) The silent surround of V1 receptive fields: theory and experiments. J Physiol 97:453–474
Serre T (2016) Models of visual categorization. Wiley Interdiscip Rev Cogn Sci 7:197–213
Serre T (2019) Deep learning: the good, the bad, and the ugly. Annu Rev Vis Sci 5(1):399
Serre T, Poggio T (2010) A neuromorphic approach to computer vision. Commun ACM 53(10):54
Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T (2007) A quantitative theory of immediate visual recognition. Prog Brain Res 165:33
Simoncelli EP, Heeger DJ (1998) A model of neuronal responses in visual area MT. Vis Res 38(5):743–761
Simoncelli, E. P., Paninski, L., Pillow, J., Swartz, O. (2004). Characterization of Neural Responses with Stochastic Stimuli. In M. Gazzaniga (Ed.), The Cognitive Neurosciences (3rd ed., pp. 327–338). Cambridge: MIT Press
Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems Vol 1. Montreal, Canada, Curran Associates, pp 568–576
Sjöström J, Gerstner W (2010) Spike-timing dependent plasticity. Scholarpedia 5(2):1362. Revision #184913
Szegedy C, Zaremba W, Sutskever I (2013) Intriguing properties of neural networks. arXiv Preprint arXiv . . ., pp 1–10
Thorpe S (2002) Ultra-Rapid Scene Categorization with a Wave of Spikes. In: Bülthoff H.H., Wallraven C., Lee SW., Poggio T.A. (eds) Biologically Motivated Computer Vision. BMCV 2002. Lecture Notes in Computer Science, vol 2525. Springer, Berlin, Heidelberg
Thorpe SJ, Gegenfurtner KR, Fabre-Thorpe M, Bülthoff HH (2001) Detection of animals in natural images using far peripheral vision. European Journal of Neuroscience 14(5):869–876. https://doi.org/10.1046/j.0953-816X.2001.01717.x
Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: ICCV '15 Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, IEEE, pp 4489–4497
Treisman A, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 136:97–136
Ullman, S., Soloviev, S. (1999). Computation of pattern invariance in brain-like structures. Neural Networks, 12, 1021–1036.
Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5(7):682–687
van den Hurk J, Van Baelen M, Op de Beeck HP (2017) Development of visual category selectivity in ventral visual cortex does not require visual experience. Proc Natl Acad Sci 114(22):E4501–E4510
Wallis G (1997) Invariant face and object recognition in the visual system. Prog Neurobiol 51(2):167–194
Wersing H, Koerner E (2003) Learning optimized features for hierarchical models of invariant recognition. Neural Comput 15(7):1559–1588
Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci 111(23):8619–8624
Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Computer vision: ECCV 2014. Springer, Berlin, pp 818–833

Further Reading

Kreiman G (2008) Biological object recognition. Scholarpedia 3(6):2667
Poggio T, Serre T (2013) Models of visual cortex. Scholarpedia 8(4):3516