Ann. N.Y. Acad. Sci. ISSN 0077-89

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Special Issue: *The Year in Cognitive Neuroscience*
REVIEW

Beyond the feedforward sweep: feedback computations in the visual cortex

Gabriel Kreiman[1] and Thomas Serre[2]

[1]Children's Hospital, Harvard Medical School and Center for Brains, Minds, and Machines, Boston, Massachusetts. [2]Cognitive Linguistic and Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, Rhode Island

Addresses for correspondence: Thomas Serre, Cognitive Linguistic and Psychological Sciences, Carney Institute for Brain Science, Brown University, 190 Thayer St, Providence, RI 02912. thomas_serre@brown.edu; Gabriel Kreiman, Children's Hospital, Harvard Medical School and Center for Brains, Minds, and Machines, 3 Blackfan Circle, Boston, MA 02115. gabriel.kreiman@childrens.harvard.edu

Visual perception involves the rapid formation of a coarse image representation at the onset of visual processing, which is iteratively refined by late computational processes. These early versus late time windows approximately map onto feedforward and feedback processes, respectively. State-of-the-art convolutional neural networks, the main engine behind recent machine vision successes, are feedforward architectures. Their successes and limitations provide critical information regarding which visual tasks can be solved by purely feedforward processes and which require feedback mechanisms. We provide an overview of recent work in cognitive neuroscience and machine vision that highlights the possible role of feedback processes for both visual recognition and beyond. We conclude by discussing important open questions for future research.

Keywords: deep learning; neural networks; machine vision; visual reasoning; categorization; grouping

Introduction

The anatomy of the primate visual system suggests an intricate network of over 30 or so interconnected visual areas, each one encompassing millions of neurons within highly specialized circuitry.[1] The neural dynamics resulting from such a network should theoretically be quite complex.[2] However, anatomical evidence suggests a clear hierarchical organization between visual areas, resulting in a feedforward versus feedback separation in terms of the connectivity patterns.[1,3,4] Such patterns of connectivity, in turn, constrain visual processing dynamics to be roughly composed of an early "bottom-up phase" primarily carried by feedforward processes during the first 150 ms after visual onset followed by a late "reentrant" phase carried by feedback processes[5] (but see also Ref. 6 for evidence of early contributions of feedback on neural responses).

A growing body of literature suggests that bottom-up processing enables the visual system to build an initial coarse visual representation before more complex visual routines are implemented. This base representation can be computed via an initial feedforward sweep of activity through the visual system and is sufficient for rapid categorization tasks.[7,8] Visual processing can be interrupted after the initial bottom-up phase and, while this interruption may prevent the visual input to reach consciousness,[5] the initial computations nonetheless allow the completion of certain visual tasks, such as speeded visual recognition.[9–11]

At the neurophysiology level, it has been shown that the early response of neurons in intermediate and higher visual areas contains enough information for decoding image category almost readily from the onset of the visual response both during passive[12,13] and active[14] presentations. Consistent with this idea, a recent monkey electrophysiology study has also shown that images that are behaviorally more difficult to classify by human observers tend to take longer to be reliably decoded, possibly

doi: 10.1111/nyas.14320

requiring additional feedback processes beyond this initial response.[15]

Human observers make recognition mistakes under these conditions, but these errors do not appear to be randomly distributed across images as would be expected from motor errors or guessing. Instead, there appears to be a systematic pattern of behavioral decisions—with some images being consistently classified correctly or incorrectly across human observers.[7,16] This pattern of correct and incorrect answers suggests an underlying visual strategy implemented in the bottom-up phase, which appears to be largely shared between human and nonhuman primates.[14,17,18]

Starting with Fukushima's neocognitron,[19] computational models constrained by the anatomy and physiology of the visual cortex (VC) (see Refs. 20–22 for reviews) account relatively well for this pattern of behavioral responses.[7] These network models process information sequentially—through a bottom-up cascade of filtering, rectification, and normalization operations—providing computational evidence for the feedforward hypothesis.[22] Interestingly, further developments of these early computational models have led to modern deep convolutional neural networks (DCNNs), which have powered recent breakthroughs in computer vision[23] as well as many other domains. Although these network models are not constrained by experimental data, they have nonetheless been shown to provide an even better fit than earlier models to both behavioral[18,24,25] and electrophysiological[26,27] data (but see Ref. 28). These network architectures now achieve accuracy well beyond those of earlier computational models of the VC and are on par with or better than human accuracy during unspeeded image categorization tasks for both object[29] and face[30] recognition.

Despite these successes, it is also becoming increasingly clear that current DCNNs remain outmatched by the power and versatility of the primate brain (see Ref. 31 for a recent review). The gap between humans and machine vision is particularly obvious when scrutinizing the results of current automatic image captioning systems (Fig. 1). Although such algorithms are reasonably good at recognizing the presence of certain objects in the scene, they often fail miserably at flexibly interpreting the fundamental gist of complex visual scenes, human actions, social interactions, and events depicted in images. To date, no known artificial system is capable of passing a visual Turing test as defined in Ref. 32.

We attribute these limitations to the fact that current systems only perform classification—in a processing mode akin to preattentive bottom-up processing. In image categorization or face identification, for instance, a category label gets associated with an image. In object detection and localization as well as in instance segmentation, image regions containing an object of interest get associated with a bounding box or a segmentation mask and a category label. In dense labeling tasks, such as semantic image segmentation tasks, every pixel gets assigned a category label. There is obviously much more to scene understanding and visual cognition than mere classification. Many visual analysis problems require a level of abstraction that transcends object recognition or naming (i.e., image classification). For instance, humans can easily answer questions about spatial relations (e.g., whether something is above, to the right, etc., of another thing) or shape relations (e.g., whether two or more shapes are the same or different up to a transformation, including rotation, etc.), even for unfamiliar shapes.[33]

Think about many of the visual reasoning tasks that one must solve daily to plan actions, or to manipulate objects, such as when finding out which of two keys will fit into a particular lock or which piece of a puzzle is the missing piece. According to Ullman, visual cognitive tasks can be decomposed into a sequence of simpler elementary operations, including, for example, visual search, texture segregation, and contour grouping.[34] These elementary operations, or visual routines, can be dynamically and flexibly assembled to solve a myriad of complex, abstract, and open-ended visual reasoning tasks. Assigning a category label to a particular image region is but one of the many visual routines needed for scene understanding.

The limitations of current computational models underlie critical aspects of visual cognition that are not accounted for by purely feedforward networks. Bottom-up processing may not be sufficient for more general visual reasoning tasks, which may necessitate bringing in feedback signals. Indeed, neuroscience evidence suggests that feedback modulation of neural responses takes place after some delay (see Refs. 6 and 35 for reviews; see also Ref. 6). The challenge is to identify which neural

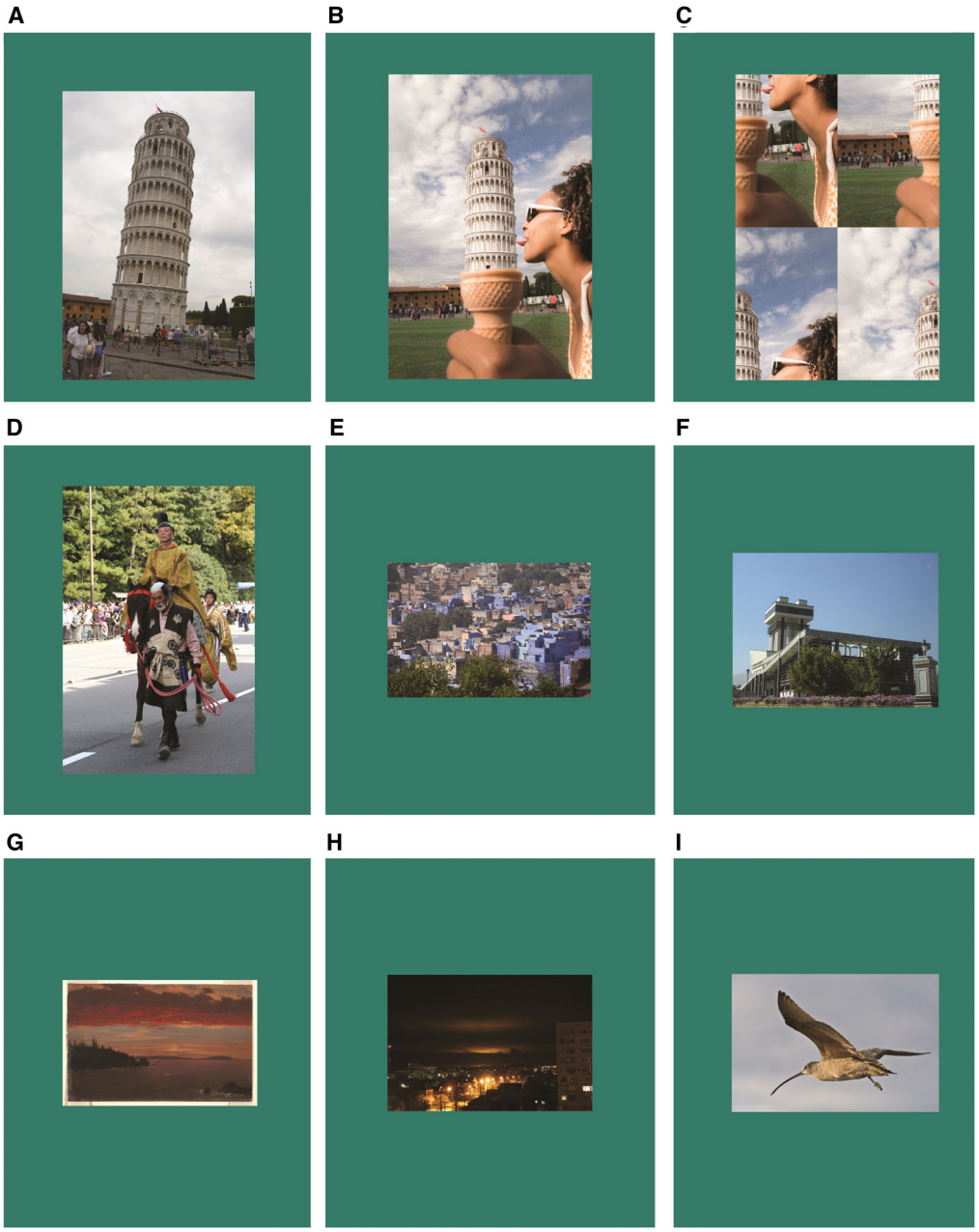


Figure 1. Current image captioning efforts illustrate exciting progress and how far we still need to go. (A–C) Example of how an image captioning system (Microsoft Cognitive Services) describes three pictures, using the Microsoft CaptionBot system (https://www.captionbot.ai/). (A) "I think it's a group of people standing in front of Leaning Tower of Pisa"; (B) "I think it's a person standing in front of Leaning Tower of Pisa"; and (C) "I think it's a person standing in front of a building." (D–I) Captions automatically generated by @picdescbot, a bot that describes random pictures from Wikimedia Commons also using Microsoft Cognitive Services (https://picdescbot.tumblr.com/about). Images posted on July 8, 2019, with the following captions (D–F): "A group of people riding horses on a city street," "a large body of water with a city in the background," and "a small clock tower in front of a house." Images posted on July 7, 2019, with the following captions (G–I): "A cat lying on top of a mountain," "a view of a city at night," and "a bird flying over a body of water."

computations are critical to visual understanding beyond rapid visual categorization, in contrast to aspects of biological computations that represent implementation details but are not critical to account for cognitive functions. The goal of our review is to bring together recent exciting and complementary developments in computational cognitive neuroscience, with behavioral and neurophysiological results as the first step toward a unifying theory for how our visual system integrates bottom-up sensory inputs with top-down mnemonic and cognitive processes.

The role of recurrence in visual recognition

Computational flexibility

Some of the most successful vision systems in many pattern recognition tasks consist of purely feedforward architectures where information flows in a single bottom-up sweep from pixels to category decisions. In stark contrast, biological architectures are characterized by pervasive feedback (also called recurrent) connectivity (Fig. 2A). A recurrent neural network (RNN) can be "unfolded" to create an equivalent purely feedforward network that performs the same computation by adding extra layers for each recurrent step (Fig. 2B). If we constrain the number of weight parameters of the unfolded network to be the same as the folded version, that is, we impose weight sharing, the two networks will carry the same computations. In other words, the same computations can be carried by a single-layer recurrent network requiring N recurrent computational steps and an $(N + 1)$–layer feedforward network with identical weights across layers.

Interestingly, several successful approaches to vision involve such feedforward architectures, where the same weights are reused recursively several times to increase the depth of visual processing. Indeed, the first texture discrimination algorithms were recursive,[36] and related ideas have also been applied to the recognition of dynamic texture.[37] Similarly, a hierarchical extension of the classic wavelet transform where the transform is applied recursively (also known as the scattering transform) has been shown to yield significant improvements in texture categorization.[38] Such recursive architectures can be implemented by RNNs within a single fully recurrent layer of processing. More recently, it has been shown that forcing recursivity into state-of-the-art DCNNs led to networks that perform better on image categorization tasks with fewer parameters.[39,40]

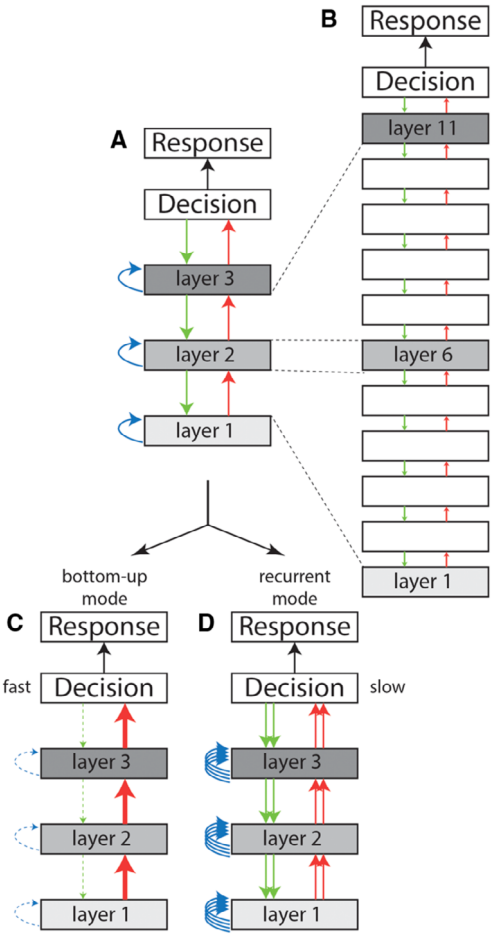


Figure 2. Recurrent networks show greater parameter efficiency and computational flexibility. (A) Schematic illustration of a three-layer network showing bottom-up (red), horizontal recurrent (blue), and top-down (green) connections. The top layer sends signals to a decision process that evaluates how confident the network is about the solution and decides whether to emit a response or continue processing by sending top-down signals that interact with the horizontal recurrent computations to enhance the solution. (B) Schematic illustration of an 11-layer network where each of the horizontal computations in part A is unfolded to generate four steps of feedforward operations with weight sharing. (C–D) The network in (C) can be flexibly utilized in a rapid bottom-up mode (C) or in a slow(er) recurrent mode (D).

Given that it is possible to unfold recurrent connections to create a deeper network with identical computational prowess, why bother with recurrent connections? Recurrent networks offer

several advantages for biological organisms over purely feedforward architectures. First, recurrent networks are potentially *computationally more efficient*. The network in Figure 2A requires fewer units, synapses, and overall shorter wiring length than the one in Figure 2B. Limiting the number of cells and synapses and the overall wire length is particularly critical for biological systems, which have size and weight constraints; the brain is also the most expensive organ from an energetic standpoint and it must operate under a constrained budget.

In the engineering literature, there is also a growing realization that energy efficiency may be an appealing reason to prefer smaller networks. A recent study estimated that training a state-of-the-art deep neural network for natural language processing costs millions of dollars in cloud computing service—with a carbon footprint equal to about five times the emissions of a single car during its entire lifetime (or about 300 NY−SF flights)[41] (see also Ref. 42).

Even ignoring energy and size constraints, a critical advantage of recurrent networks is that they are *computationally more flexible*. The depth of processing required to solve different types of tasks may not be known ahead of time. While most computer vision tasks require training a network to solve a specific task (e.g., categorize images in ImageNet[43]), the brain needs to solve a possibly endless and constantly changing set of tasks. Unfolding a highly recurrent network to create a deeper feedforward network makes a commitment to a specific architecture and a given number of computational steps. Imagine that after you tried different architectures to label certain images, the dataset changes, but now you are stuck with the architectural choices. By and large, the adult brain's architecture is fixed: it is possible to add a few neurons (neurogenesis), some neurons die, and synapses come and go, but the overall number of layers and number of units per layer is to a first approximation essentially fixed. Recurrent connections offer the flexibility to potentially vary the depth of processing across tasks, without the need to change the architecture for each task.[a]

This computational flexibility to perform multiple and arbitrary recognition tasks carries additional benefits. Some tasks may be easier (i.e., require less processing depth) and can be solved in a faster fashion—possibly through a single feedforward sweep of activity—while other tasks may benefit from those additional computational steps afforded by recurrent connections.[45–47] An image could rapidly traverse through the architecture in Figure 2C to reach a decision stage. This decision stage (perhaps located in the prefrontal cortex (PFC)) can evaluate whether it has enough information to produce a response. If it does, then the problem is solved with just a rapid feedforward sweep. If it does not, then the decision stage may provide additional fast feedback signals through top-down connections to lower areas or wait for slower intra-areal horizontal feedback signals to provide additional elaboration and finally, produce a response.

This flexibility to use more or fewer computations, in real time and on demand, could at least partly account for the well-known speed-accuracy trade-offs in psychophysics experiments and also for the fact that certain easy problems might be solved in a rapid or speeded operation mode (Fig. 2C), whereas other tasks may be solved in a slower mode (Fig. 2D).[48] Indeed, a related idea referred to as adaptive computing is gaining traction in computer vision and natural language processing and is being actively explored both with feedforward[49] and recurrent networks.[50,51]

An experimental technique that has been used to impose rapid processing is *backward masking*. Shortly after flashing a stimulus, a noise mask is presented. The interval between the onset of the stimulus and the mask, generally referred to as stimulus onset asynchrony typically, encompasses between ∼50 and ∼100 milliseconds. Under these conditions, the mask purportedly interferes with and interrupts the interactions between recurrent signals and the incoming inputs, thereby emphasizing bottom-up processing of the stimulus[52–56] (but see Ref. 57 for a counterargument). It has been shown that, electrophysiologically, the initial sweep of rapid visually selective signals along the ventral VC (VVC) is unaffected by backward masking.[14]

[a]A related way to achieve flexibility is through bypass routes,[44] which allow the architecture to skip some of the processing stages,[22] and which may help alleviate the issue of a fixed architecture to some extent (at the expense of adding and training yet more connections).

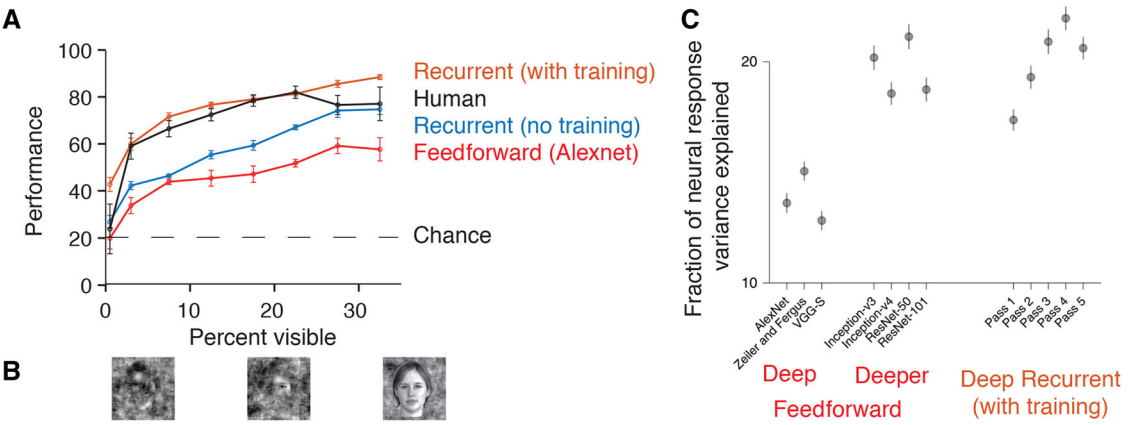


Figure 3. Recurrent networks help visual recognition. (A–B) Recognition performance in a five-way categorization task of partially visible objects for humans (black), layer fc7 in the popular AlexNet neural network (red), AlexNet network embedded with attractor-like horizontal recurrent connectivity in the fc7 layer without any training with occluded objects (blue) or with training (orange). Example objects from limited visibility to full visibility are shown in panel B. Chance performance = 20% (dashed line). Modified from Ref. 48. (C) The fraction of neural response variance explained for neurons in macaque inferior temporal cortex. For images that are difficult to recognize in a rapid feedforward mode, adding more layers to a feedforward network can improve neural variance explained (deeper feedforward networks), but the same effect can be achieved by multiple passes through a shallower network with horizontal recurrent connections (deep recurrent). Modified from Ref. 15.

Consistent with the idea that backward masking interrupts recurrent processing, recent work has shown that the introduction of a rapid mask interferes with the ability to perform visual recognition tasks that require more processing time such as pattern completion,[48] as elaborated in the section entitled "Generalization in visual recognition" (Fig. 3A and 3B).

Consistent with this idea, Eberhardt *et al.* trained classifiers on the outputs of individual layers derived from several representative DCNNs for the categorization of animal versus nonanimal images and found that the accuracy of the classifiers increased as a function of the layers' depth.[24] Interestingly, they found that the correlation between model predictions derived from individual layers versus human participants engaged in the same speeded categorization task peaked at intermediate layers. Because the accuracy of human observers increases monotonically as a function of the response time available to respond, these results suggest that human observers may adjust the depth of visual processing—not through static depth as done in current DCNN architectures—but through time via recurrent processes.

The separation of time scales into a rapid initial feedforward sweep followed by a late recurrent processing mode is of course only an approximation. There is no clear-cut separation between these two modes of operation, and cortical computations are continuous, with varying degrees of the preponderance between feedforward and recurrent computations.[58] Yet, this approximate separation of temporal scales has been useful to conceptualize and understand the sequence of computations that ultimately lead to visual cognition.

Long-range spatial dependencies and perceptual grouping

It has long been assumed that feedback mechanisms play a key role in perceptual grouping (see Refs. 6 and 59–62 for early proposals). Yet, the recent successes of deep convolutional networks for contour detection and image segmentation in seemingly challenging visual tasks (e.g., Refs. 63 and 64) have obfuscated the need for feedback.

To demonstrate the limitations of current feedforward networks for learning long-range spatial dependencies, Linsley *et al.*[65] described a simple visual recognition challenge inspired by cognitive psychology tasks (see Ref. 35 for review) called the "Pathfinder," which involves judging whether there exists a path linking two markers in an image (Fig. 4A). To control for intraclass variability and task difficulty, they systematically varied the length of individual contours in the stimulus set.

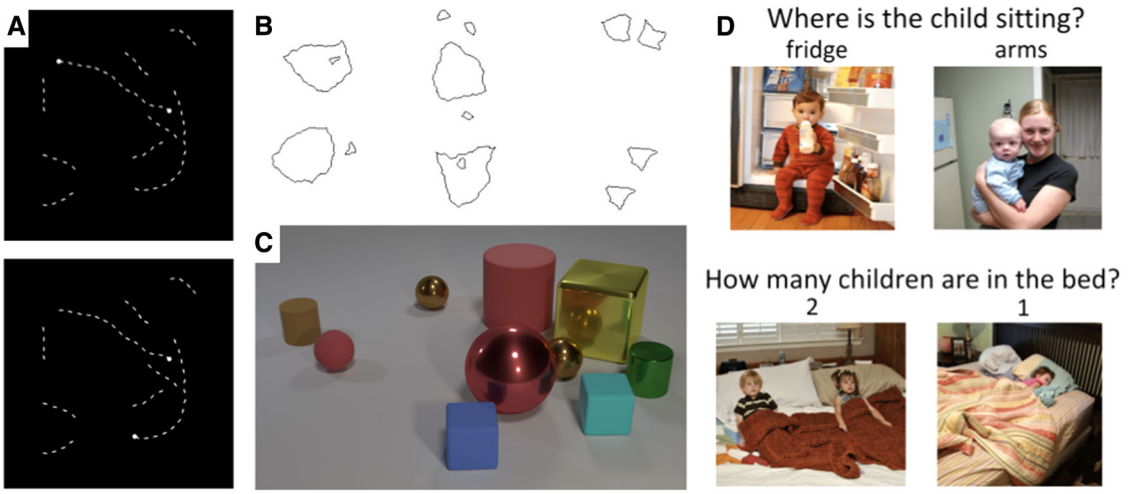


Figure 4. Sample visual reasoning tasks. (A) The Pathfinder challenge whether the task is to evaluate where the two larger white dots are connected or not.[65] (B) Synthetic visual reasoning test.[92] Six examples where the task is to decide whether (left) a small shape is inside (top) or outside (bottom) a larger one, (middle) whether two small shapes fall on the same side (top) or different sides (bottom) of a larger object boundary and (right) whether two shapes are the same (top) or different (bottom). (C) Visual question answering (VQA) on the CLEVR challenge[97] to test aspects of visual reasoning, such as attribute identification, counting, comparison, and logical operations. (D) Sample questions and answers with corresponding images from the VQA challenge.

Increasingly deeper networks were needed to solve this task as the path length increased, which likely reflects the need for receptive fields at the top to contain the entire paths and hence the need for increasingly deep architectures. By contrast, it was found that imbuing neurons with the ability to incorporate context through horizontal connections led to a single-layer highly recurrent neural network that was able to perform on par or better than all tested feedforward hierarchical baselines, despite the fact that these feedforward networks contained orders of magnitude more parameters. This observation provides compelling evidence that some visual tasks, such as contours tracing tasks, are much better suited for recurrent neural circuits. In a follow-up work, Kim *et al.*[66] extended the Pathfinder challenge, which stresses low-level gestalt cues, to a task that they called "cluttered ABC" (cABC), which emphasizes high-level object cues for perceptual grouping. As in the Pathfinder task, in the cABC task, markers are placed either on two different shapes or the same shape. Here, the shapes consist of highly overlapping capitalized English alphabet characters and the task consists of judging whether the two markers fall on the same or different characters. As for the Pathfinder, the authors found that increasing the intraclass variability in cABC strained learning in networks that rely solely on bottom-up processing. Furthermore, a distinct type of feedback resolved the difficulties associated with each challenge: horizontal connections resolved this limitation on tasks, such as Pathfinder, featuring gestalt cues by relying on incremental spatial propagation of activities. Top-down connections rescued learning on tasks, such as cABC, featuring object cues by propagating coarse predictions about the expected pose of the target object. These findings thus disassociate the computational roles of bottom-up, horizontal, and top-down connectivity, and demonstrate how a recurrent network model featuring all these interactions can more flexibly form perceptual groups.

Beyond perceptual grouping, several other computer vision tasks have been shown to benefit from a similar inclusion of recurrent processing, including image generation,[67] object recognition,[39,68–70] and superresolution tasks.[71]

Generalization in visual recognition

To a first approximation, the number of free parameters of a learning algorithm, including neural networks, constrains the sample complexity of the network,[72] that is, the number of training samples

needed to have some reasonable guarantee that the algorithm will be able to generalize to novel examples that were not encountered before. A network with fewer weights may be more *sample efficient* and hence require fewer samples to train although this is not always observed in practice—a phenomenon that is not fully understood (e.g., see Ref. 73).

State-of-the-art deep neural networks include dozens to hundreds of layers of processing (often, they even correspond to ensembles of dozens of networks) corresponding to the equivalent of thousands of processing layers. As a result, these networks contain tens of millions of free parameters. In theory, these algorithms can effortlessly *memorize* millions of training examples. Even entire datasets as large as some of the largest ones currently available, such as CIFAR[74] or ImageNet,[43] could be memorized.

One measure of a network's capacity to memorize training samples is called the *shattering dimension*. The shattering dimension is a measure of the intrinsic degrees of freedom of a neural network. The larger the capacity, the more training examples will be needed for proper generalization from learned to novel data. Initially, the shattering dimension was computed for the perceptron by estimating the number of entirely random patterns that can be classified correctly. A related measure can be computed for real images by shuffling the class labels associated with individual images so as to train the network to learn random associations between individual images and category labels. This idea was used by Recht *et al.*[75] who confirmed that modern deep network architectures could achieve near-perfect training accuracy using random labels. Such high training accuracy for classifying random labels shows that, in principle, neural networks are capable of memorizing millions of individual samples and their class labels without necessarily learning any abstract category information.

With fewer parameters to fit, an RNN may require fewer samples for training[76] (i.e., lower sample complexity). Indeed, Linsley *et al.*[77] have shown that it is possible to reduce the sample complexity of a vision system for contour detection by introducing recurrent connections in state-of-the-art neural networks.

Inherent to the discussion about sample complexity and whether neural networks memorize all their training data is the distinction between interpolation and extrapolation. This dichotomy roughly corresponds to the in- versus out-of-distribution test sample problem in machine learning: the extent to which models can extrapolate to out-of-distribution samples, as opposed to only interpolating to novel samples within the same distribution. Cross-validation is a central tenet in machine learning that guides model evaluation. Cross-validation dictates the separation of training data from test data, but it does not specify how different the training and test data need to be. If there is only a single pixel that distinguishes a training image from a test image, one could still state that there is cross-validation, but the degree of extrapolation is obviously minimal.

Generally, when the test and training data are very similar, an algorithm is tested for its ability to *interpolate*. For example, an algorithm may be trained using images of a chair shown at 90 degrees in-plane rotation and a chair shown at 0 degrees in-plane rotation. The algorithm is afterward tested with an image of the same chair at 45 degrees in-plane rotation. A significantly more impressive feat for a learning algorithm would be to be able to identify a completely different chair, with a different color and texture, in a completely different background, under different illumination conditions, shown from a different 3D angle, and so on. Extrapolation refers to the ability to make adequate responses with out-of-distribution samples.

One prominent feature of our own visual system is its ability to extrapolate to unseen conditions, including views of a novel object not seen during training.[78] Observers are also able to readily identify celebrities from photographs that are blurred even up to leaving only about a hundred pixels or photographs that have been stretched in unnatural never-seen-before conditions.[79] Evidence that modern DCNNs do not generalize in such conditions includes the work by Geirhos *et al.*,[80] who showed that these networks can classify noisy images much better than humans, but they cannot generalize to similar albeit different types of noise. In a similar vein, Linsley *et al.* have shown that the network architectures that exhibit "superhuman" accuracy for the segmentation of neural tissue from serial electron microscopy images when trained and tested on different subsets of the same volume do exhibit a large drop in accuracy when trained and tested on different volumes[81]

(for practical applications, the issue can be alleviated using machine learning methods for "realigning" datasets[82]). By comparison, they found that RNNs endowed with horizontal and top-down connections can generalize much better and use fewer training examples.[66,77]

Solving harder recognition problems with recurrence

There are many visual recognition problems that seem to require additional processing time beyond the mostly feedforward initial wave encompassing ~150 ms described above. One prominent example is the ability to make inferences from partial information during recognition of heavily occluded objects.[83] During natural visual conditions, many objects are partially visible either because they are occluded by other objects in front of them or because of poor illumination or because of unusual viewing angles. Despite such challenging visual conditions, primate visual recognition is quite robust even when up to 90% of the object is occluded, even in the absence of contextual cues, and even when subjects have minimal prior experience with the object in question.[84]

Behavioral, neurophysiological, and computational evidence suggest that purely bottom-up computations are generally insufficient to perform pattern completion of heavily occluded objects. At the behavioral level, recognition of heavily occluded objects takes longer than the recognition of the whole object counterparts. Furthermore, pattern completion performance is impaired by the introduction of a backward mask. These reaction time delays and sensitivity to masking are indicative of the need for additional computations beyond the feedforward sweep. These behavioral measurements are consistent with the latencies reported in neurophysiological recordings during pattern completion. The latency of neurophysiological signals in areas V4 and the inferior temporal (IT) cortex in response to heavily occluded objects is delayed by about 50 ms with respect to the responses of the same circuits to the fully visible objects.[84,85] These behavioral and neurophysiological observations are further corroborated by computational models: state-of-the-art bottom-up models struggle during recognition of heavily occluded objects, unless they are extensively trained with those specific occluded objects.[86,87]

The inadequacy of purely bottom-up signals for pattern completion suggests that the ability to infer the whole from the parts relies on additional horizontal and/or top-down signals. Indeed, computational work has shown that the addition of recurrent computations to DCNNs can help solve the problem of pattern completion.[48,83] Additionally, there is physiological evidence that strongly suggests that top-down signals from the PFC onto the VVC play an important role during the recognition of occluded objects.[83,88] It is also known that familiar object shapes have an influence on image segmentation,[34,89,90] and it is possible that the ability to complete patterns and make inferences from partial information is enhanced by top-down effects on image segmentation.

Occlusion is not the only situation in which visual recognition requires additional computation. Recognition of objects presented under different viewpoints, at extreme scales, or under poor illumination, may require similar computational mechanisms. Consistent with this idea, recent work has shown that the extent to which a given image is hard to recognize by state-of-the-art computational models is also correlated with increased decoding latencies in recordings from the IT cortex. Similar to the work on object occlusion, incorporating horizontal connections to bottom-up models can rescue their performance (Fig. 3C).[15] Recurrent computations are not only relevant for recognition, but they can also help solve other problems. We mentioned earlier the challenges in image segmentation in connectomics with purely feedforward architectures. Linsley *et al.* have shown that RNNs generalize significantly better to novel volumes without the need to align the various datasets.[81]

The role of recurrence beyond recognition

Visual reasoning

Visual cognition entails much more than object recognition and categorization. Observers perform extensive visual analyses in order to plan for their actions or manipulate objects, navigate in their environments, drive, and so on. Such visual analyses can be performed without explicit object recognition. A nonexhaustive list of such visual reasoning tasks was proposed in Ref. 34 by Ullman. For instance, Ullman lists tasks that involve visual judgments as to whether a shape lies inside or outside of a closed curve. Such a task appears to require sophisticated

computations and those computations may be distinct from the ones involved in categorization; for example, pigeons show an impressive capacity for shape classification and recognition, yet they are essentially unable to perform the inside/outside task in a generalizable manner.[91] Another example provided by Ullman involves judging the elongation of ellipse-like figures, whether two black dots lie on a common contour or whether one shape can be moved to another specified location without colliding with any of the other shapes. Such tasks appear artificial, but they are reminiscent of the kinds of visual inference that observers need to solve when "mak[ing] use of visual aids such as diagrams, charts, sketches, and maps, because they draw on the system's natural capacity to manipulate and analyze spatial information, and this ability can be used to help our reasoning and decision processes."

Some of these tasks were subsequently formalized by Fleuret *et al.* in their Synthetic Visual Reasoning Task,[92] a collection of 23 binary classification problems in which opposing classes differ based on whether or not images obey an abstract rule. All stimuli depict simple, closed, black curves on a white background. Positive and negative examples are shown in Figure 4B for three representative problems. Most importantly, the shapes used in these images are unique without overlap between the training and testing to prevent rote shape memorization and force the learning of abstract rules. The challenge broke the state of the art in computer vision in 2011 right before the deep learning era. Today, the challenge seems to remain significant for modern DCNNs as shown by several groups.[93–95]

In particular, Kim *et al.*[95] found a clear dichotomy between visual reasoning tasks: while spatial relations appeared to be easily learnable by feedforward neural networks (DCNNs and their extensions), same−different relations appear to pose a particular strain on these networks (i.e., they require deeper architectures and significantly more training examples to be learned). Ultimately, even with one million samples available to train the networks for each of the problems, learning same–different visual relations posed a challenge for these architectures when stimulus variability made rote memorization difficult (although see Ref. 96 for evidence that a deeper residual network pretrained on ImageNet could actually fair better). This result is all the more striking as such similarity judgments constitute a major component of IQ tests making them an especially important problem to solve for computer vision systems.

Interestingly, Kim *et al.* suggested that the ability of modern neural networks to solve basic visual reasoning tasks might have been overlooked. They considered a representative challenge used in the visual question answering known as the Sort-of-CLEVR challenge[97] (Fig. 4C) and confirmed that networks appear to learn visual relations when trained and tested on the same sets of shapes (i.e., a fixed combination of shapes × color attributes). However, when trained on all but one combination of shape × color, the neural networks they evaluated did not appear to generalize to the left-out condition, suggesting that they simply memorize the shapes presented during training and do not learn the underlying abstract category rule. Furthermore, Kim *et al.* showed that learning same−different problems became trivial for a feedforward network that is fed with perceptually grouped stimuli.

This demonstration and the comparative success of biological vision in learning visual relations[98–101] (including insects and even newborn ducklings) suggests that feedback mechanisms, such as attention, working memory (WM), and perceptual grouping, may be the key components underlying human-level abstract visual reasoning. There is substantial evidence that visual relation detection in primates depends on recurrent processing that is lacking in standard DCNNs. Indeed, converging evidence[102–104] suggests that the processing of spatial relations between pairs of objects in a cluttered scene requires attention, even when individual objects can be detected preattentively (but see also Ref. 105). Another brain mechanism implicated in our ability to process visual relations is WM.[106–108] In particular, imaging studies[106,107] have highlighted the role of WM in prefrontal and premotor cortices when participants solve Raven's progressive matrices that require both spatial and same−different reasoning.

What is the computational role of attention and WM in the detection of visual relations? One assumption is that these two mechanisms allow flexible representations of relations to be constructed *dynamically* at runtime via a sequence of attention shifts rather than *statically* by storing visual relation templates in synaptic weights (as done in feedforward neural networks).[104,109] Such representations

built "on-the-fly" circumvent the combinatorial explosion associated with the storage of templates for all possible relations and objects,[110] helping to prevent the capacity overload that plagues DCNNs and other feedforward neural networks.

Attention and search

Much of the recent progress in image categorization has been driven by the inclusion of trainable attention modules in state-of-the-art DCNN architectures. While biology is sometimes mentioned as a source of inspiration,[111–117] the attentional mechanisms that have been considered remain quite limited in comparison with the rich and diverse array of processes used by the human visual system (see Ref. 118 for a review).

One of the prominent types of tasks to study the role of top-down attention in cortical processing is visual search.[119] In a typical scenario, a target object is presented (e.g., Waldo), followed by a search image, and the subject has to freely move their eyes to locate the target. In this type of task, the subject needs to maintain a representation of the target object features in WM and use knowledge about those features in a top-down fashion to guide active sampling of the image via eye movements.[120,121]

Recent neurophysiological work has started to provide insights into the neural circuitry involved in visual search.[122,123] Bichot and colleagues trained monkeys to perform a visual search task while recording activity from the PFC and the frontal eye fields (FEFs). They found that neurons in the PFC show a visually selective response upon presentation of the target cue, maintain that information during the delay period, and convey that information to the FEF to direct the next saccade. Furthermore, inactivation of the specific subregions within the frontal cortex involved in visual search led to a significant impairment in the monkey's ability to efficiently find the target.[122] The selective attention signals from the PFC are fed back to modulate the responses along the ventral visual stream (reviewed in Ref. 123). There is a reverse hierarchy in the magnitude of such attentional effects, which are more prominent in higher visual areas and manifest themselves in a clear but largely reduced fashion in early visual areas.

Several computational models have been proposed recently to capture how top-down signals modulate processing of an image and guide eye movements during visual search. Inspired by the neurophysiology of visual search, Zhang and colleagues built a simple architecture consisting of a DCNN, which aims to mimic the extraction of features along the VVC, and a PFC-like module that stores information about the sought target and provides top-down feature-based attentional modulation onto the VC.[124] Combining the bottom-up features with top-down target modulation led to the creation of an attention map that dictates the location of the next saccade in a winner-take-all fashion. The model was able to provide a reasonable approximation to both the number and spatiotemporal sequence of eye movements that humans executed during visual search tasks spanning a wide range of difficulty levels. Both humans and the model were able to locate targets despite large transformations in the target features (i.e., invariantly to object changes) and despite having had no prior experience with the target objects (i.e., in a zero-shot fashion).

Related recent work by Adeli and Zelinsky provided a biologically inspired implementation of biased competition theory, whereby the multiple objects in a display compete with each other for attention and a top-down signal is used to disambiguate and bias this competition in favor of the sought target.[125] Such feature-based modulation is more efficient when applied at later stages of the visual hierarchy,[124,126] which is consistent with physiological observations showing that both spatial and feature-based attention is considerably weaker in early visual cortical areas compared with higher visual cortical areas.

It is instructive to compare these recent advances in modeling visual search with parallel approaches in the computer vision literature. Unlike in the image categorization tasks described earlier, where entire images are associated with a single class label, object localization tasks may require the detection of one or multiple objects and the ability to draw a bounding box around them. Region-based approaches are popular DCNN extensions that achieve state-of-the-art results for object detection and localization. The basic idea behind region-based approaches is to first run a generic object detector over the image, as in the region-based convolutional neural network (R-CNN),[127] to bring down the number of windows to be classified (called the region proposals) to a reasonable

number (from millions for a system scanning the image across all positions and scales to a few thousands). These windows are then classified by a DCNN to yield a class label for each bounding box (including an option to reject the bounding box as containing none of the objects of interest). The approach was improved in a series of papers from the Fast R-CNN[128] to the Faster R-CNN[129] and the region-based fully convolutional networks (R-FCN)[130] by sharing convolutional layers between the region proposal stage and the detection and localization stages—thus allowing the training of a single efficient DCNN for the entire system. Another notable architecture is YOLO (you only look once),[131] which can run with near state-of-the-art accuracy but in real-time for typical image resolutions used in computer vision datasets.

It is worth noting that modern architectures for object localization are not concerned with biological plausibility or computational efficiency. Despite all the aforementioned improvements, searching for a target object in the large image displays would require a very large amount of computational resources. This cost is arguably an evolutionary force behind the biological machinery used to implement eye movements and eccentricity-dependent sampling as done in Ref. 125. Consistent with this idea, Eckstein *et al.*[132] have shown that, unlike current architectures for object localization that scan for objects exhaustively across scales, human search is largely guided by context. As a result, human observers, unlike computer vision systems, will often miss targets when their size is inconsistent with the rest of the scene (even when targets are made larger and more salient and observers fixated the target).

Another remarkable distinction between computer vision object detection algorithms and biologically inspired models is that the former requires extensive training with the sought targets. A state-of-the-art algorithm for object detection, such as YOLO, can only look for the types of objects that it was trained on. Nothing more, nothing less. In stark contrast, Zhang *et al.* show that their model can rapidly find target objects after a single exposure to them.[124]

Nonetheless, it has been shown that, while the visual representations learned by DCNNs without attention bear little overlap with those used by human observers for visual recognition,[133] attention mechanisms help DCNNs learn visual representations that are more similar to those used by human observers.[134] In particular, Linsley *et al.* have shown that it is possible to leverage crowdsourcing methods to identify image features that are diagnostic for human recognition and to leverage that knowledge to cue DCNNs to attend to these regions during training for image categorization. As a result, DCNNs learn visual representations that are significantly more similar to those used by human observers in addition to DCNNs that generalize better to novel images (Fig. 5).

Learning and plasticity

At the core of modern deep learning is the need to adjust the large number of tunable weight parameters present in the network. For the most part, successes in vision have relied on supervised learning approaches whereby weights are adjusted via the presentation of labeled examples so as to minimize the classification error on the training data. One of the most widely used algorithms for this type of training is backpropagation.[135] There has been a lot of discussion in the field about the biological plausibility of such backpropagation algorithms.[136,137] There has been a recent spur of interest in the design of more biologically plausible learning algorithms for training neural networks.

An important criticism of the backpropagation algorithm has been the need for "symmetric" connectivity with feedback connections matching the weights of their corresponding feedforward counterparts (the weight transport problem). While the extent of such symmetry—or lack thereof—in cortical networks remains to be quantified, algorithms have been described that provide simple and biologically plausible learning mechanisms for feedback synaptic weights to adapt so as to match feedforward ones.[138] Moreover, recent work has demonstrated that it may even be possible to perform adequate learning via backpropagation using random feedback weights[139]—at least via matching of the feedback and feedforward synaptic signs without necessarily equating their magnitudes.[140] Another important limitation concerns the mechanisms of credit assignment during learning, including the propagation of gradients, the timing of credit allocations, and even the mere origin of such credit signals. Here again, there has been significant progress toward algorithms that can assign

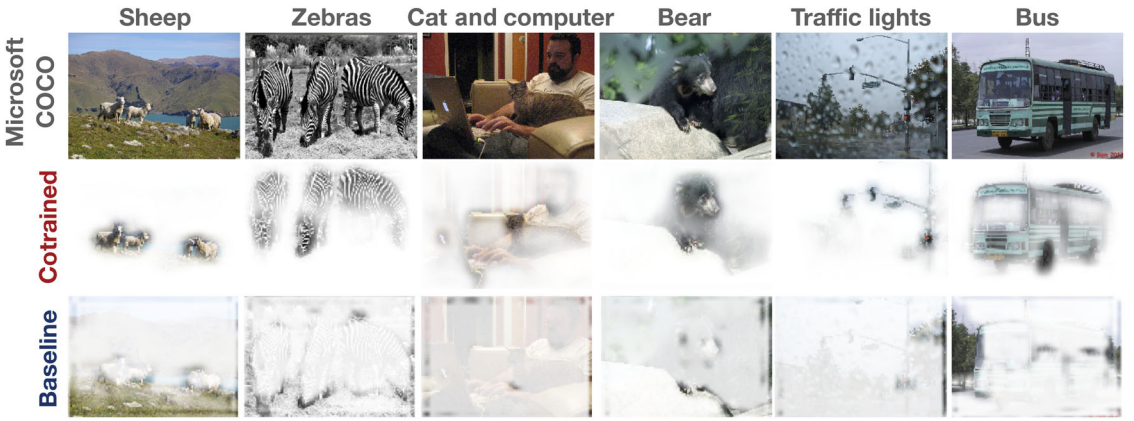


Figure 5. Learning what and where to attend. The top row depicts representative images from the Microsoft Common Objects in Context (COCO) dataset depicting object categories also present in ILSVRC12 (which was used for training the system). In the middle row, each of these images is shown with the transparency set to the attention map it yielded in the attention network by Linsley *et al.*[167] trained with human supervision (see the text for details). Visible features were attended to by the model, and transparent features were ignored. Animal parts like faces and tails are typically emphasized, whereas vehicle parts like windows and windshields are not. Cotraining the attention network with human supervision yields better classification accuracy on ImageNet, as well as learned feature representations that are more human-like. The system also generalizes from the ImageNet to the Microsoft COCO dataset (shown here) despite significant changes in the objects' scale. The bottom row shows the same visualization using attention maps from the same architecture trained without human supervision, which has distributed and less interpretable attention. Image credit: Drew Linsley. Adapted with permission.

and propagate credits in more biologically palatable forms.[137,141,142]

Another widely successful approach to tuning weights is via reinforcement learning.[143] Reinforcement learning algorithms have demonstrated seemingly magical performance in tasks, such as learning how to play games like chess, Go, or different types of video games, even beating world champions.[144] One can only dream about the potential of reinforcement learning approaches to learning vision, but there has not been much progress on their implementation yet. Initial work has already demonstrated the benefits of combining reinforcement learning with RNNs to play Atari® games.[145] Promising results have also been obtained for visual tracking,[146,147] face recognition,[148] action recognition,[149,150] video captioning,[151] color enhancement,[152] and object detection.[153,154]

Another approach to learning structure in the visual world, which does not use explicit labeled examples or a teacher and provides direct rewards/punishment for specific actions, is based on the intuition that predicting what will happen next may be an important principle of computation in the brain. This idea is at the core of several theories, including the adaptive resonance theory[155,156] and closely related predictive coding algorithms.[157,158] Predictive coding algorithms have recently regained momentum in the context of deep network architectures.[159–162] Common to many of these models is the notion that feedback signals provide a prediction of what will transpire next while the feedforward signals convey an error, or difference, between those predictions and the incoming inputs.

Predictive signals carried by top-down connections can provide a powerful and highly efficient mechanism to learn structure in the world because they do not require the type of expensive and abundant guidance from a teacher as in traditional supervised learning methods. In fact, many of these predictive algorithms have been trained using unlabeled videos, of which there is no shortage of for the computer science community, and it is particularly easy to conceive that infants also have almost unlimited access to this type of input during development. In the computer science literature, using prediction as a learning signal in video sequences is generally grouped under the term *self-supervised learning*, and there is intense work in trying to use this type of approach to pretrain networks in

order to drastically reduce the number of examples required in subsequent supervised learning steps.[163] It is particularly intriguing that predictive networks trained with random natural videos (e.g., videos of cars navigating in a city) can automatically develop units that resemble fundamental properties of cortical computation and perception.[164]

Concluding remarks and future directions

We have brought together recent complementary developments in computational cognitive neuroscience, with behavioral and neurophysiological results, as a critical step toward a unifying theory for how our visual system integrates bottom-up sensory inputs with top-down mnemonic and cognitive processes. In particular, we have highlighted the limitations of state-of-the-art feedforward neural network architectures to solve visual reasoning tasks beyond image categorization. Recent computer vision work toward the development of RNNs constitutes an initial first step toward addressing some of the abovementioned shortcomings. While our understanding of feedback processes in the visual system remains relatively limited, it is our hope that recent developments in computer vision may start to provide computational-level hypotheses for linking feedback processes with visual functions.

A fundamental area of the investigation that remains rather enigmatic is how to connect our understanding of visual computations along the VVC to high-level cognition. For example, while examining a scene depicting kids playing in the playground, we can interpret the location, actions, what is behind what, how different people interact with each other, we understand what those strange structures in the playground are—even if they may be heavily occluded and even if we have never seen them before, we can easily infer why the swing is in a given position, we can guess a kid's intentions by following their gaze, we can predict the trajectory of a ball even from a static snapshot, and we can generally answer a near-infinite number of questions about the scene in a flexible manner. This type of general knowledge about the world can be vaguely construed as "common sense," reflecting the understanding that humans have about their environment. How this information is stored in the brain and the mechanisms by which it provides top-down modulation of bottom-up sensory processing in the VC remains as enigmatic as ever and will probably constitute an area of active research in the upcoming years.

Perhaps one of the paradigmatic examples of exciting progress, which at the same time illustrates how far we still have to go, is the problem of image captioning. Consider the example image in Figure 1A that we uploaded to one of the state-of-the-art systems for image captioning (Microsoft CaptionBot). The system correctly determined that there is a group of people. Captioning systems tend to be pretty good at detecting people, in part because it is likely that a large fraction of the training data contain people. The system astutely infers that the people are standing, not a trivial feat. Perhaps, there are lots of features that show that the picture is outdoors and there is an imperfect but strong correlation between outdoor pictures and people standing. Furthermore, the system correctly recognizes the leaning Tower of Pisa. There is probably an enormous corpus of photographs with "Tower of Pisa" labels for training and the vast majority of those pictures are probably circumscribed to a relatively small number of well-described angles, sizes, colors, and so on. It is perhaps possible but not very common to find an image of the Tower of Pisa upside down, with each level painted in a different color and with a black background instead of the blue sky (a quick search in Google images yields images with some, but not all, of those features). Recognizing major landmarks from conventional angles is probably a relatively easy task. The system not only achieves all of these recognition feats, but it also produces a grammatically correct sentence. All of these are quite remarkable achievements that go well beyond where image captioning was a decade ago.

Yet, that is as far as the algorithms go. Consider the example in Figure 1B. Here again, the algorithm correctly infers that there is a person, detects the Tower of Pisa and even conjectures, probably correctly, that the person is standing. But the algorithm misses some of the essential aspects of the image. It fails to detect an ice cream cone, the hand holding the cone, and other background elements. The system fails to notice that the cone is particularly well aligned with the base of the Tower of Pisa, nor does it appreciate that the Tower of Pisa appears to be the ice cream. And the system does not understand that the girl is holding the cone and sticking her tongue to lick the ice cream. Frustratingly,

scrambling the image yields a similar caption (Fig. 1C), even though the scrambled version lacks the critical gist of what is happening in the image. In this case, the algorithm was not even able to detect the scrambled Tower of Pisa. The captions for Figure 1A and B are very similar, despite the fact that those images evoke rather different reactions in human observers. This example illustrates some of the fundamental challenges ahead to bring in feedback signals that can incorporate our common sense knowledge about the world in the interpretation of a visual scene.

Heroic studies of the initial wave of processing in the VC have led to successful computational-neuroscience models and breakthrough technologies with real-world applications. Here, we have argued that the next generation of computational models will focus on the second wave of processing incorporating feedback loops. Modeling short-range interactions within the VC and long-range interactions between the frontal areas and VC promises an even wider and more radical transformation whereby common sense knowledge, prior experience, language, and symbolic reasoning can be systematically and rigorously integrated with incoming visual signals to create richer models that are capable of general intelligence in more complex and generalizable tasks.

Humans can effortlessly construct an unbounded set of structured descriptions about their visual world.[32] Mechanisms in the visual system, such as perceptual grouping, attention, and WM, exemplify how the brain learns and handles combinatorial structures in the visual environment with a small amount of experience.[165] However, exactly how attentional and mnemonic mechanisms interact with hierarchical feature representations in the VC is not well understood. Given the vast superiority of humans over modern computers in their ability to solve seemingly simple visual reasoning tasks, we see the exploration of these cortical mechanisms as a crucial step in our computational understanding of visual reasoning.

Acknowledgments

G.K. was funded by NSF STC Award CCF-1231216. T.S. was funded by ONR Grant #N00014-19-1-2029, CRCNS Grant #IIS-1912280, and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute Grant #ANR-19-PI3A-0004. We would like to thank Drew Linsley and Junkyung Kim for their feedback on the manuscript.

Author contributions

T.S. and G.K. wrote the paper jointly.

Competing interests

T.S. serves as a scientific advisor for Vium, Inc.

References

1. Felleman, D.J. & D.C. Van Essen. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1:** 1–47.
2. Boccaletti, S., V. Latora, Y. Moreno, *et al.* 2006. Complex networks: structure and dynamics. *Phys. Rep.* **424:** 175–308.
3. Salin, P.A. & J. Bullier. 1995. Corticocortical connections in the visual system: structure and function. *Physiol. Rev.* **75:** 107–154.
4. Markov, N.T., M.M. Ercsey-Ravasz, A.R. Ribeiro Gomes, *et al.* 2014. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cereb. Cortex* **24:** 17–36.
5. Lamme, V.A. & P.R. Roelfsema. 2000. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23:** 571–579.
6. Bullier, J. 2001. Integrated model of visual processing. *Brain Res. Brain Res. Rev.* **36:** 96–107.
7. Serre, T., A. Oliva & T. Poggio. 2007. A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. USA* **104:** 6424–6429.
8. VanRullen, R. 2007. The power of the feed-forward sweep. *Adv. Cogn. Psychol.* **3:** 167–176.
9. Biederman, I., J.C. Rabinowitz & A.L. Glass. 1974. On the information extracted from a glance at a scene. *J. Exp. Psychol.* **103:** 597–600.
10. Potter, M.C. 1975. Meaning in visual search. *Science* **187:** 565–566.
11. Thorpe, S., D. Fize & C. Marlot. 1996. Speed of processing in the human visual system. *Nature* **381:** 520–522.
12. Hung, C.P., G. Kreiman, T. Poggio & J.J. Dicarlo. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* **2164:** 863–866.
13. Liu, H., J.R. Madsen, Y. Agam & G. Kreiman. 2009. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* **62:** 281–290.
14. Cauchoix, M., S.M. Crouzet, D. Fize & T. Serre. 2016. Fast ventral stream neural activity enables rapid visual categorization. *Neuroimage* **125:** 280–290.
15. Kar, K., J. Kubilius, K. Schmidt, *et al.* 2019. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22:** 974–983.
16. VanRullen, R. & S.J. Thorpe. 2001. Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception* **30:** 655–668.

17. Fize, D., M. Cauchoix & M. Fabre-Thorpe. 2011. Humans and monkeys share visual representations. *Proc. Natl. Acad. Sci. USA* **108:** 7635–7640.
18. Rajalingham, R., K. Schmidt & J.J. DiCarlo. 2015. Comparison of object recognition behavior in human and monkey. *J. Neurosci.* **35:** 12127–12136.
19. 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.
20. Riesenhuber, M. & T. Poggio. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2:** 1019–1025.
21. Serre, T. 2015. Hierarchical models of the visual system. *Encyclopedia of Computational Neuroscience*. https://doi.org/10.1007/978-1-4614-6675-8_345.
22. Serre, T., G. Kreiman, M. Kouh, *et al.* 2007. A quantitative theory of immediate visual recognition. *Prog. Brain Res.* **165:** 33–56.
23. LeCun, Y., Y. Bengio & G. Hinton. 2015. Deep learning. *Nature* **521:** 436–444.
24. Eberhardt, S., J.G. Cader & T. Serre. 2016. How deep is the feature analysis underlying rapid visual categorization? In *Advances in Neural Information Processing Systems*. D.D. Lee, M. Sugiyama, U.V. Luxburg, *et al*., Eds.: 1100–1108. Red Hook, NY: Curran Associates, Inc.
25. Kheradpisheh, S.R., M. Ghodrati, M. Ganjtabesh & T. Masquelier. 2016. Deep networks can resemble human feed-forward vision in invariant object recognition. *Sci. Rep.* **6:** 32672.
26. Cadieu, C.F., H. Hong, D.L.K. Yamins, *et al*. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10:** e1003963.
27. Yamins, D.L.K., H. Hong, C.F. Cadieu, *et al.* 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111:** 8619–8624.
28. Rajalingham, R., E.B. Issa, P. Bashivan, *et al.* 2018. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38:** 7255–7269.
29. He, K., X. Zhang, S. Ren & J. Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*. https://arxiv.org/abs/1512.03385.
30. Kemelmacher-Shlizerman, I., S.M. Seitz, D. Miller & E. Brossard. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4873–4882.
31. Serre, T. 2019. Deep learning: the good, the bad and the ugly. *Annu. Rev. Vis. Neurosci*. **5:** 399–426.
32. Geman, D., S. Geman, N. Hallonquist & L. Younes. 2015. Visual Turing test for computer vision systems. *Proc. Natl. Acad. Sci. USA* **112:** 3618–3623.
33. Shepard, R.N. & J. Metzler. 1971. Mental rotation of three-dimensional objects. *Science* **171:** 701–703.
34. Ullman, S. 1996. *High-Level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: MIT Press.
35. Roelfsema, P.R., V.A. Lamme & H. Spekreijse. 2000. The implementation of visual routines. *Vision Res.* **40:** 1385–1411.
36. Malik, J. & P. Perona. 1990. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A* **7:** 923–932.
37. Hadji, I. & R.P. Wildes. 1990. A spatiotemporal oriented energy network for dynamic texture recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2017, pp. 3066–3074.
38. Bruna, J. & S. Mallat. 2013. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **35:** 1872–1886.
39. Liao, Q. & T. Poggio. 2016. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. https://arxiv.org/abs/1604.03640.
40. Guo, Q., Z. Yu, Y. Wu, *et al.* 2019. Dynamic recursive neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5147–5156.
41. Strubell, E., A. Ganesh & A. McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Annual Meeting of the Association for Computational Linguistics*.
42. Schwartz, R., J. Dodge, N.A. Smith & O. Etzioni. 2019. Green AI. https://arxiv.org/abs/1907.10597.
43. Russakovsky, O., J. Deng, H. Su, *et al*. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis*. **115:** 211–252.
44. Nakamura, H., R. Gattass, R. Desimone & L.G. Ungerleider. 1993. The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. *J. Neurosci.* **13:** 3681–3691.
45. Hochstein, S. & M. Ahissar. 2002. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* **36:** 791–804.
46. Tsotsos, J.K., A.J. Rodríguez-Sánchez, A.L. Rothenstein & E. Simine. 2008. The different stages of visual recognition need different attentional binding strategies. *Brain Res.* **1225:** 119–132.
47. Hegdé, J. 2008. Time course of visual perception: coarse-to-fine processing and beyond. *Prog. Neurobiol.* **84:** 405–439.
48. Tang, H., M. Schrimpf, W. Lotter, *et al*. 2018. Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci. USA* **115:** 8835–8840.
49. Srivastava, R.K., K. Greff & J. Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems*. C. Cortes, N.D. Lawrence, D.D. Lee, *et al*., Eds.: 2377–2385. Red Hook, NY: Curran Associates, Inc.
50. Graves, A. 2016. Adaptive computation time for recurrent neural networks. https://arxiv.org/abs/1603.08983.
51. Zilly, J.G., R.K. Srivastava, J. Koutník & J. Schmidhuber. 2017. Recurrent highway networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, Sydney, NSW, JMLR.org, 4189–4198.
52. Di Lollo, V., J.T. Enns & R.A. Rensink. 2000. Competition for consciousness among visual events: the psychophysics of reentrant visual processes. *J. Exp. Psychol. Gen.* **129:** 481–507.

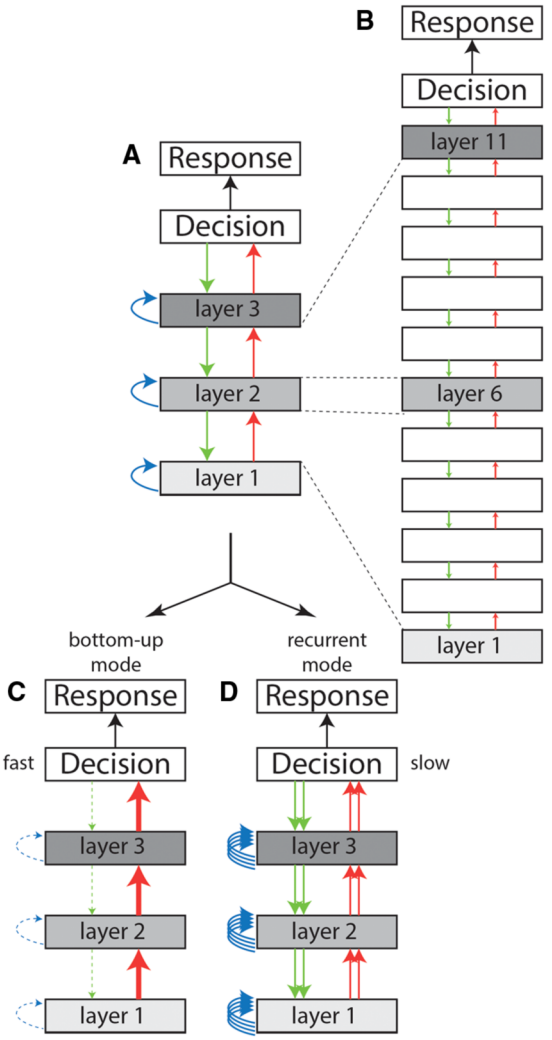
53. Lamme, V.A., K. Zipser & H. Spekreijse. 2002. Masking interrupts figure-ground signals in V1. *J. Cogn. Neurosci.* **14:** 1044–1053.
54. Breitmeyer, B. & H. Ogmen. 2006. *Visual Masking: Time Slices through Conscious and Unconscious Vision*. Oxford Psychology Series. Oxford: Oxford University Press.
55. Fahrenfort, J.J., H.S. Scholte & V.A.F. Lamme. 2007. Masking disrupts reentrant processing in human visual cortex. *J. Cogn. Neurosci.* **19:** 1488–1497.
56. Di Lollo, V. 2007. Iterative reentrant processing: a conceptual framework for perception and cognition (the blinding problem? No worries, mate). *Tutorials in Visual Cognition*. Vol. 2010, pp. 9–42.
57. Macknik, S.L. & S. Martinez-conde. 2007. The role of feedback in visual masking and visual processing. *Adv. Cogn. Psychol.* **3:** 125–153.
58. Hegdé, J. & D.J. Felleman. 2007. Reappraising the functional implications of the primate visual anatomical hierarchy. *Neuroscientist* **13:** 416–421.
59. Fukushima, K. 1987. Neural network model for selective attention in visual pattern recognition and associative recall. *Appl. Opt.* **26:** 4985.
60. Grossberg, S., E. Mingolla & W.D. Ross. 1997. Visual brain and visual perception: how does the cortex do perceptual grouping? *Trends Neurosci.* **20:** 106–111.
61. Gilbert, C.D. & W. Li. 2013. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14:** 350–363.
62. Li, Z. 2002. A saliency map in primary visual cortex. *Trends Cogn. Sci.* **6:** 9–16.
63. He, J., S. Zhang, M. Yang, *et al.* 2002. Bi-directional cascade network for perceptual edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2019, pp. 3828–3837.
64. Lee, K., J. Zung, P. Li, *et al.* 2017. Superhuman accuracy on the SNEMI3D connectomics challenge. https://arxiv.org/abs/1706.00120.
65. Linsley, D., J.K. Kim, V. Veerabadran, *et al.* 2018. Learning long-range spatial dependencies with horizontal gated recurrent units. In *Neural Information Processing Systems (NIPS)*. Accessed September 22, 2018. https://nips.cc/Conferences/2018/Schedule?showEvent=11042.
66. Kim, J., D. Linsley, K. Thakkar & T. Serre. 2020. Disentangling neural mechanisms for perceptual grouping. In *Proceedings of the International Conference on Learning Representations* (ICLR), Accepted.
67. Van Den Oord, A., N. Kalchbrenner & K. Kavukcuoglu. 2016. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, New York, NY, 1747–1756.
68. O'Reilly, R.C., D. Wyatte, S. Herd, *et al.* 2013. Recurrent processing during object recognition. *Front. Psychol.* **4:** 1–14.
69. Liang, M. & X. Hu. 2015. Recurrent convolutional neural network for object recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 3367–3375.
70. Zamir, A.R., T. Wu, L. Sun, *et al.* 2017. Feedback networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1808–1817.
71. Kim, J., J. Kwon Lee & K. Mu Lee. 2016. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1637–1645.
72. Anthony, M. & P.L. Bartlett. 2009. *Neural Network Learning: Theoretical Foundations*. Cambridge: Cambridge University Press.
73. Neyshabur, B., S. Bhojanapalli, D. McAllester & N. Srebro. 2017. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, USA, Curran Associates Inc., 5949–5958.
74. Krizhevsky, A. & G. Hinton. 2009. Learning multiple layers of features from tiny images. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf.
75. Zhang, C., S. Bengio, M. Hardt, *et al.* 2016. Understanding deep learning requires rethinking generalization. https://arxiv.org/abs/1611.03530.
76. Akpinar, N.-J., B. Kratzwald & S. Feuerriegel. 2019. Sample complexity bounds for recurrent neural networks with application to combinatorial graph problems. https://arxiv.org/abs/1901.10289.
77. Linsley, D., J. Kim & T. Serre. 2020. Recurrent neural circuits for contours detection. In *Proceedings of the International Conference on Learning Representations* (ICLR), Accepted, 2020.
78. Biederman, I. & P.C. Gerhardstein. 1993. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *J. Exp. Psychol. Hum. Percept. Perform.* **19:** 1162–1182.
79. Sinha, P. 2002. Recognizing complex patterns. *Nat. Neurosci.* **5**(Suppl.): 1093–1097.
80. Geirhos, R., C.R.M. Temme, J. Rauber, *et al.* 2018. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*. S. Bengio, H. Wallach, H. Larochelle, *et al.*, Eds.: 7549–7561. Red Hook, NY: Curran Associates, Inc.
81. Linsley, D., J. Kim, D. Berson & T. Serre. 2018. Robust neural circuit reconstruction from serial electron microscopy with convolutional recurrent networks. https://arxiv.org/abs/1811.11356.
82. Januszewski, M. & V. Jain. 2019. Segmentation-enhanced CycleGAN. bioRxiv. https://doi.org/10.1101/548081.
83. Wyatte, D., D.J. Jilk & R.C. O'Reilly. 2014. Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front. Psychol.* **5:** 674.
84. Tang, H., C. Buia, R. Madhavan, *et al.* 2014. Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* **83:** 736–748.
85. El-Shamayleh, Y., A.M. Fyall & A. Pasupathy. 2014. The role of visual area V4 in the discrimination of partially occluded shapes. *J. Neurosci.* **34:** 8570–8584.
86. Rosenfeld, A., R. Zemel & J.K. Tsotsos. 2018. The elephant in the room. https://arxiv.org/abs/1808.03305.

87. Wang, J., Z. Zhang, C. Xie, *et al*. 2018. Visual concepts and compositional voting. *Ann. Math. Sci. Appl*. **3:** 151–188.
88. Fyall, A.M., Y. El-Shamayleh, H. Choi, *et al.* 2017. Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. *eLife* **6:** e25784.
89. Peterson, M.A., E.M. Harvey & H.J. Weidenbacher. 1991. Shape recognition contributions to figure-ground reversal: which route counts? *J. Exp. Psychol. Hum. Percept. Perform.* **17:** 1075–1089.
90. Vecera, S.P. & M.J. Farah. 1997. Is visual image segmentation a bottom-up or an interactive process? *Percept. Psychophys.* **59:** 1280–1296.
91. Herrnstein, R.J., W. Vaughan, Jr, D.B. Mumford & S.M. Kosslyn. 1989. Teaching pigeons an abstract relational rule: insideness. *Percept. Psychophys.* **46:** 56–64.
92. Fleuret, F., T. Li, C. Dubout, *et al.* 2011. Comparing machines and humans on a visual categorization test. *Proc. Natl. Acad. Sci. USA* **108:** 17621–17625.
93. Ellis, K., A. Solar-Lezama & J. Tenenbaum. 2015. Unsupervised learning by program synthesis. In *Advances in Neural Information Processing Systems*. C. Cortes, N.D. Lawrence, D.D. Lee, *et al*., Eds.: 973–981. Red Hook, NY: Curran Associates, Inc.
94. Stabinger, S., A. Rodríguez-Sánchez & J. Piater. 2016. 25 Years of CNNs: can we compare to human abstraction capabilities? In *Artificial Neural Networks and Machine Learning – ICANN 2016*, 380–387, Springer International Publishing.
95. Kim, J.K., M. Ricci & T. Serre. 2018. Not-So-CLEVR: learning same–different relations strains feedforward neural networks. *Interface Focus*. https://doi.org/10.1098/rsfs.2018.0011.
96. Borowski, J., C.M. Funke, K. Stosio, *et al.* 2019. The notorious difficulty of comparing human and machine perception. In *2019 Conference on Cognitive Computational Neuroscience*.
97. Johnson, J., B. Hariharan, L. van der Maaten, *et al.* 2017. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2017, pp. 1988–1997.
98. Donderi, D.C. & D. Zelnicker. 1969. Parallel processing in visual same–different decisions. *Percept. Psychophys.* **5:** 197–200.
99. Giurfa, M., S. Zhang, A. Jenett, *et al.* 2001. The concepts of "sameness" and "difference" in an insect. *Nature* **410:** 930–933.
100. Wasserman, E.A., L. Castro & J.H. Freeman. 2012. Same–different categorization in rats. *Learn. Mem.* **19:** 142–145.
101. Martinho, A. & A. Kacelnik. 2016. Ducklings imprint on the relational concept of "same or different." *Science* **353:** 286–288.
102. Logan, G.D. 1994. Spatial attention and the apprehension of spatial relations. *J. Exp. Psychol. Hum. Percept. Perform.* **20:** 1015–1036.
103. Rosielle, L.J., B.T. Crabb & E.E. Cooper. 2002. Attentional coding of categorical relations in scene perception: evidence from the flicker paradigm. *Psychon. Bull. Rev.* **9:** 319–326.
104. Franconeri, S.L., J.M. Scimeca, J.C. Roth, *et al*. 2012. Flexible visual processing of spatial relationships. *Cognition* **122:** 210–227.
105. Hayworth, K.J., M.D. Lescroart & I. Biederman. 2011. Neural encoding of relative position. *J. Exp. Psychol. Hum. Percept. Perform.* **37:** 1032–1050.
106. Kroger, J.K., F.W. Sabb, C.L. Fales, *et al.* 2002. Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cereb. Cortex* **12:** 477–485.
107. Golde, M., D.Y. von Cramon & R.I. Schubotz. 2010. Differential role of anterior prefrontal and premotor cortex in the processing of relational information. *Neuroimage* **49:** 2890–2900.
108. Clevenger, P.E. & J.E. Hummel. 2014. Working memory for relations among objects. *Atten. Percept. Psychophys*. **76:** 1933–1953.
109. Tsotsos, J.K. 2011. *A Computational Perspective on Visual Attention*. Cambridge, MA: MIT Press.
110. Riesenhuber, M. & T. Poggio. 1999. Are cortical models really bound by the "binding problem"? *Neuron* **24:** 87–93.
111. Stollenga, M., J. Masci, F. Gomez & J. Schmidhuber. 2014. Deep networks with internal selective attention through feedback connections. arXiv:1407.3068[cs.CV].
112. Mnih, V., N. Heess, A. Graves & K. Kavukcuoglu. 2014. Recurrent models of visual attention. *Adv. Neural Inform. Process. Syst.* **27:** 1–9.
113. Cao, C., X. Liu, Y. Yang, *et al*. 2015. Look and think twice: capturing top-down visual attention with feedback convolutional neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2956–2964.
114. You, Q., H. Jin, Z. Wang, *et al.* 2016. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4651–4659.
115. Chen, L., H. Zhang, J. Xiao, *et al.* 2017. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6298–6306.
116. Wang, F., M. Jiang, C. Qian, *et al*. 2017. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6450–6458.
117. Biparva, M. & J. Tsotsos. 2017. STNet: selective tuning of convolutional networks for object localization. In *The IEEE International Conference on Computer Vision (ICCV)*. Venice, 2017. Vol.**2**: 2715–2723.
118. Itti, L., G. Rees & J.K. Tsotsos. 2005. *Neurobiology of Attention*. Cambridge, MA: Academic Press.
119. Wolfe, J.M. 2007. Guided search 4.0: current progress with a model of visual search. In *Series on Cognitive Models and Architectures. Integrated Models of Cognitive Systems*. W.D. Gray, Ed.: 99–119. Oxford: Oxford University Press.
120. Hamker, F. 2005. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Comput. Vis. Image Underst*. **100:** 64–106.
121. Hamker, F.H. 2005. The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal

cortex, and areas V4, IT for attention and eye movement. *Cereb. Cortex* **15:** 431–447.
122. Bichot, N.P., M.T. Heard, E.M. DeGennaro & R. Desimone. 2015. A source for feature-based attention in the prefrontal cortex. *Neuron* **88:** 832–844.
123. Moore, T. & M. Zirnsak. 2017. Neural mechanisms of selective visual attention. *Annu. Rev. Psychol.* **68:** 47–72.
124. Zhang, M., J. Feng, K.T. Ma, *et al.* 2018. Finding any Waldo with zero-shot invariant and efficient visual search. *Nat. Commun.* **9**. https://doi.org/10.1038/s41467-018-06217-x.
125. Adeli, H. & G. Zelinsky. 2018. Deep-BCN: deep networks meet biased competition to create a brain-inspired model of attention control. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
126. Lindsay, G.W. & K.D. Miller. 2018. How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife* **7**. https://doi.org/10.7554/eLife.38105.
127. Girshick, R., J. Donahue, T. Darrell & J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.
128. Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.
129. Ren, S., K. He, R. Girshick & J. Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. C. Cortes, N.D. Lawrence, D.D. Lee, *et al*., Eds.: 91–99. Red Hook, NY:Curran Associates, Inc.
130. Dai, J., Y. Li, K. He & J. Sun. 2016. R-FCN: object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*. D.D. Lee, M. Sugiyama, U.V. Luxburg, *et al*., Eds.: 379–387. Red Hook, NY: Curran Associates, Inc.
131. Redmon, J. & A. Farhadi. 2017. YOLO9000: better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525.
132. Eckstein, M.P., K. Koehler, L.E. Welbourne & E. Akbas. 2017. Humans, but not deep neural networks, often miss giant targets in scenes. *Curr. Biol.* **27:** 2827–2832.e3.
133. Linsley, D., S. Eberhardt, T. Sharma, *et al.* 2017. What are the visual features underlying human versus machine vision? In *IEEE ICCV Workshop on the Mutual Benefit of Cognitive and Computer Vision*.
134. Linsley, D., D. Shiebler, S. Eberhardt & T. Serre. 2019. Learning what and where to attend. *In ICLR*. Accessed January 4, 2019. https://openreview.net/pdf?id=BJgLg3R9KQ.
135. Rumelhart, D.E., G.E. Hinton & J.L. McClelland. 1986. A general framework for parallel distributed processing. In *Parallel Distributed Processing, Volume 1. Explorations in the Microstructure of Cognition: Foundations*. D.E. Rumelhart, G.E. Hinton & J.L. McClellan, Eds.: 45–76. Cambridge, MA: MIT Press.
136. Crick, F. 1989. The recent excitement about neural networks. *Nature* **337:** 129–132.
137. Bengio, Y., D.-H. Lee, J. Bornschein, *et al.* 2015. Towards biologically plausible deep learning. https://arxiv.org/abs/1502.04156.
138. Burbank, K.S. & G. Kreiman. 2012. Depression-biased reverse plasticity rule is required for stable learning at top-down connections. *PLoS Comput. Biol.* **8:** e1002393.
139. Lillicrap, T.P., D. Cownden, D.B. Tweed & C.J. Akerman. 2016. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* **7:** 13276.
140. Liao, Q., J.Z. Leibo & T. Poggio. 2016. How important is weight symmetry in backpropagation? In *13th AAAI Conference on Artificial Intelligence*.
141. Guerguiev, J., T.P. Lillicrap & B.A. Richards. 2017. Towards deep learning with segregated dendrites. *eLife* **6:** e22901.
142. Miconi, T. 2017. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife* **6:** e20899.
143. Sutton, R.S. & A.G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
144. Silver, D., A. Huang, C.J. Maddison, *et al*. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* **529:** 484–489.
145. Hausknecht, M. & P. Stone. 2015. Deep recurrent Q-learning for partially observable MDPs. *2015 AAAI Fall Symposium Series*, pp. 29–37.
146. Ren, L., X. Yuan, J. Lu, *et al.* 2015. Deep reinforcement learning with iterative shift for visual tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 2018, pp. 684–700.
147. Guo, M., J. Lu & J. Zhou. 2018. Dual-agent deep reinforcement learning for deformable face tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 768–783.
148. Rao, Y., J. Lu & J. Zhou. 2017. Attention-aware deep reinforcement learning for video face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3931–3940.
149. Tang, Y., Y. Tian, J. Lu, *et al.* 2018. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5323–5332.
150. Chen, L., J. Lu, Z. Song & J. Zhou. 2018. Part-activated deep reinforcement learning for action prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 421–436.
151. Wang, X., W. Chen, J. Wu, *et al.* 2018. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4213–4222.
152. Park, J., J.-Y. Lee, D. Yoo & I. So Kweon. 2018. Distort-and-recover: color enhancement using deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5928–5936.
153. Kong, X., B. Xin, Y. Wang & G. Hua. 2018. Collaborative deep reinforcement learning for joint object search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2017, pp. 1695–1704.
154. Rao, Y., D. Lin, J. Lu & J. Zhou. 2018. Learning globally optimized object detector via policy gradient. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, 6190–6198.
155. Pollen, D.A. 1999. On the neural correlates of visual perception. *Cereb. Cortex* **9:** 4.
156. Grossberg, S. 2013. Adaptive resonance theory. *Scholarpedia* **8:** 1569.
157. Rao, R.P. & D.H. Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2:** 79–87.
158. Bastos, A.M., W.M. Usrey, R.A. Adams, *et al.* 2012. Canonical microcircuits for predictive coding. *Neuron* **76:** 695–711.
159. Lotter, W., G. Kreiman & D. Cox. 2016. Deep predictive coding networks for video prediction and unsupervised learning. https://arxiv.org/abs/1605.08104.
160. Vondrick, C., H. Pirsiavash & A. Torralba. 2015. Anticipating visual representations from unlabeled video. https://arxiv.org/abs/1504.08023.
161. O'Reilly, R.C., D.R. Wyatte & J. Rohrlich. 2017. Deep predictive learning: a comprehensive model of three visual streams. https://arxiv.org/abs/1709.04654.
162. Wen, H., K. Han, J. Shi, *et al.* 2018. Deep predictive coding network for object recognition. https://arxiv.org/abs/1802.04762.
163. van den Oord, A., Y. Li & O. Vinyals. 2018. Representation learning with contrastive predictive coding. https://arxiv.org/abs/1807.03748.
164. Lotter, W., G. Kreiman & D. Cox. 2018. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. https://arxiv.org/abs/1805.10734.
165. Tenenbaum, J.B., C. Kemp, T.L. Griffiths & N.D. Goodman. 2011. How to grow a mind: statistics, structure, and abstraction. *Science* **331:** 1279–1285.
166. Goyal, Y., T. Khot, D. Summers-Stay, *et al.* 2017. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913.
167. Linsley, D.S.E., D. Shiebler & T. Serre. 2019. Learning what and where to attend. In *International Conference on Learning Representations*.

Graphical Abstract & Image



The goal of our review is to bring together recent exciting and complementary developments in computational cognitive neuroscience, with behavioral and neurophysiological results as the first step toward a unifying theory for how our visual system integrates bottom-up sensory inputs with top-down mnemonic and cognitive processes.