ANNALS of THE NEW YORK ACADEMY OF SCIENCES

Beyond the feedforward sweep: feedback computations in the visual cortex

Journal:	*Ann NY Acad Sci*
Manuscript ID	Draft
Manuscript Type:	Review
Date Submitted by the Author:	n/a
Complete List of Authors:	Kreiman, Gabriel; Harvard University, Children's Hospital, Harvard Medical School and Center for Brains, Minds, and Machines Serre, Thomas; Brown University, CLPS
Keywords:	feedforward, feedback, deep neural network, visual recognition, visual reasoning

Beyond the feedforward sweep: feedback computations in the visual cortex

Gabriel Kreiman[1] & *Thomas Serre*[2]

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

Corresponding author: Thomas Serre <thomas_serre@brown.edu> and Gabriel Kreiman <gabriel.kreiman@childrens.harvard.edu>

Keywords: deep learning, neural networks, machine vision, visual reasoning, image categorization, incremental grouping, segmentation

Abstract: 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 which highlights the possible role of feedback processes for both visual recognition and beyond. We conclude by discussing important open questions for future research.

Acknowledgments: GK was funded by NSF STC award CCF-1231216. TS was funded by ONR grant #N00014-19-1-2029, CRCNS grant #, and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute. TS and GK wrote the paper jointly. We would like to thank Drew Linsley and Junkyung Kim for their feedback on the manuscript.

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 vs. feedback separation in terms of the connectivity patterns [1,3]. 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 milliseconds after visual onset followed by a late “reentrant” phase carried by feedback processes [4].

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 through an initial feedforward sweep of activity through the visual system and is sufficient for rapid categorization tasks [5,6]. Visual processing can be interrupted after the initial bottom-up phase and, while this interruption may prevent the visual input to reach consciousness [4], the initial computations nonetheless allow the completion of certain visual tasks such as speeded visual recognition [7–9]. 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 [10,11] and active [12] presentations. 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 [5,13]. 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 non-human primates [12,14,15].

Computational models constrained by the anatomy and physiology of the visual cortex (see [16–18] for reviews) account relatively well for this pattern of behavioral responses [5]. These network models process information sequentially – through a bottom-up cascade of filtering, rectification and normalization operations – providing computational evidence for the feedforward hypothesis [18]. 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 [19] 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 [15,20,21] and electrophysiological [22,23] data (but see [24]). These network architectures now achieve accuracy well beyond those of earlier computational models of the visual cortex and are on par with or better than human accuracy during unspeeded image categorization tasks for both object [25] and face [26] 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 [27] for a recent review). The gap between human 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 [28].

We attribute these limitations to the fact that current systems only perform classification – in a processing mode akin to pre-attentive 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 which 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 [29].

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 (1996), visual cognitive tasks can be decomposed into a sequence of simpler elementary operations including e.g., visual search, texture segregation and contour grouping [30]. 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 [31] for review). The challenge is to identify which neural 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 this review is to bring together recent exciting and complementary developments in computational cognitive neuroscience, with behavioral and neurophysiological results as the first step towards 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, i.e., 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 re-used recursively several times to increase the depth of visual processing. Indeed, the first texture discrimination algorithms were recursive [32] and related ideas have also been applied to the recognition of dynamic texture [33]. 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 [34]. 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 which perform better on image categorization tasks with fewer parameters [35,36].

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 **Fig. 2A** requires fewer units, synapses, and overall shorter wiring length than the one in **Fig. 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 5 times the emissions of a single car during its entire lifetime (or about 300 NY-SF flights) [37] (see also [38]).

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 [39]), 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.[1]

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. An image could rapidly traverse through the architecture in **Fig. 2C** to reach a decision stage. This decision stage (perhaps located in the prefrontal cortex), 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 less 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**) [41]. 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 [42] and recurrent networks [43,44].

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 ms. 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 [5,45–47] (but see [48] for a counter-argument). It has been shown that, electrophysiologically, the initial sweep of rapid visually selective signals along the ventral visual cortex is unaffected by backward masking [12].

[1] A related way to achieve flexibility is through bypass routes [40], which allow the architecture to skip some of the processing stages [18], and which may help alleviate the issue of a fixed architecture to some extent (at the expense of adding and training yet more connections).

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 [41], as elaborated upon under the section entitled “Generalization beyond interpolation”.

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 vs. non-animal images and found that the accuracy of the classifiers increased as a function of the layers’ depth [20]. 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 preponderance between feedforward and recurrent computations [49]. 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

To demonstrate the limitations of current feedforward networks for learning long-range spatial dependencies, Linsley et al [50] described a simple visual recognition challenge inspired by cognitive psychology tasks (see [31] for review) called the “Pathfinder” which involves judging whether there exists a path linking two markers in an image (**Fig. 3c**). To control for intra-class variability and task difficulty, they systematically varied the length of individual contours in the stimulus set. 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. In 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 outperform 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 follow-up work, Kim et al [51] extended the Pathfinder challenge, which stresses low-level gestalt cues, to a task which 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 in judging whether the two markers fall on the same or different characters (**Fig 3d**). 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, 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 position 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 [52], object recognition [35,53–55] and super-resolution tasks [60].

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 [56], 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 which is not fully understood (see e.g., [57]).

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). 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 [58] or ImageNet [39] 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 [59] 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, a recurrent neural network may require fewer samples for training [60] (i.e., lower sample complexity). Indeed, Linsley et al [61] 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- vs. 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, etc. 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 [62]. 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 [63]. Evidence that these networks do not generalize in such conditions includes the work by Geirhos et al [64] who showed that modern deep neural 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 [65]. In comparison, they found that recurrent neural networks endowed with horizontal and top-down connections can generalize much better and use fewer training examples [51,61].

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 in the

Introduction. One prominent example is the ability to make inferences from partial information during recognition of heavily occluded objects [66]. 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.

Behavioral, neurophysiological, and computational evidence suggests 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 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 [67,68]. 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 [69,70].

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 deep convolutional networks can help solve the problem of pattern completion [41,71]. Additionally, there is physiological evidence that strongly suggests that top-down signals from prefrontal cortex onto ventral visual cortex play an important role during the recognition of occluded objects [71,72]. It is also known that familiar object shapes have an influence on image segmentation [30,73,74] 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 inferior temporal cortex. Similar to the work on object occlusion, incorporating horizontal connections to bottom-up models can rescue their performance [75]. Recurrent computations are not only relevant for recognition but they can help solve other problems as well. We mentioned earlier the challenges in image segmentation in connectomics with purely feedforward architectures. Linsley et al have shown that recurrent neural networks generalize significantly better to novel volumes without the need to align the various datasets [65].

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, etc. Such visual analyses can be performed without explicit object recognition. A non-exhaustive list of such visual reasoning tasks was proposed in [30] 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 [76]. 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 [77], 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 **Fig. 3a** for 3 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 deep convolutional neural networks as shown by several groups [78–80].

In particular, Kim et al [80] found a clear dichotomy between visual reasoning tasks: While spatial relations appeared to be learnable by feedforward neural networks (DCNNs and their extensions), same-different relations appear to pose a particular strain on these networks. Ultimately, even with 1 million samples available to train the networks for each of the problems, they failed to learn same–different visual relations when stimulus variability made rote memorization difficult. 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 Visual Question Answering known as the Sort-of-CLEVR challenge [81] (**Fig. 4b**) 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 x color attributes). However, when trained on all but one combination of shape x 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 [82–85] (including insects and even newborn ducklings) suggests that feedback mechanisms such as attention, working memory, 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 [86–88] suggests that the processing of spatial relations between pairs of objects in a cluttered scene requires attention, even when individual objects can be detected pre-attentively (but see also [89]). Another brain mechanism implicated in our ability to process visual relations is working memory [90–92]. In particular, imaging studies [90,91] have highlighted the role of working memory in prefrontal and premotor cortices when participants solve Raven's progressive matrices which require both spatial and same-different reasoning.

What is the computational role of attention and working memory in the detection of visual relations? One assumption [88] is that these two mechanisms allow flexible representations of relations to be constructed *dynamically* at run-time via a sequence of attention shifts rather than *statically* by storing visual-relation templates in synaptic weights (as done in feedforward neural networks). Such representations built “on-the-fly” circumvent the combinatorial explosion associated with the storage of templates for all possible relations and objects [93], 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 [94–100], the attentional mechanisms that have been considered remain quite limited in comparison to the rich and diverse array of processes used by the human visual system (see [101] for a review).

One of the prominent types of tasks to study the role of top-down attention in cortical processing is visual search [102]. In a typical scenario, a target object is presented (e.g., Waldo), followed by a search image, and the subject has to freely move the eyes to locate the target. In this type of task, the subject needs to maintain a representation of the target object features in working memory and use knowledge about those features in a top-down fashion to guide active sampling of the image via eye movements.

Recent neurophysiological work has started to provide insights into the neural circuitry involved in visual search [103,104]. Bichot and colleagues trained monkeys to perform a visual search task while recording activity from prefrontal cortex (PFC) and the frontal eye fields (FEF). They found that neurons in 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 frontal cortex involved in visual search led to a significant impairment in the monkey's ability to efficiently find the target [103]. The selective attention signals from PFC are fed back to modulate the responses along the ventral visual stream (reviewed in [104]). 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 ventral visual cortex, and a prefrontal cortex-like module that stores information about the sought target and provides top-down feature-based attentional modulation onto visual cortex [105]. 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 the 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 & 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 [106]. Such feature-based modulation is more efficient when applied at later stages of the visual hierarchy [105,107], which is consistent with physiological observations showing that both spatial and feature-based attention is considerably weaker in early visual cortical areas compared to 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 R-CNN [108], 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 [109] to the Faster R-CNN [110] and the region-based fully convolutional networks (R-FCN) [111] 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 [112], 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 [106]. Consistent with this idea, Eckstein et al [113] have shown that, unlike current architectures for object localization which 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 [105].

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 [114], attention mechanisms help DCNNs learn visual representations that are more similar to those used by human observers [115]. In particular, Linsley et al have shown that it is possible to leverage crowd-sourcing 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 back-propagation [116]. There has been a lot of discussion in the field about the biological plausibility of such back-propagation algorithms [117], [118]. 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 [119]. Moreover, recent work has demonstrated that it may even be possible to perform adequate learning via back-propagation using random feedback weights [120] – at least via matching of the feedback and feedforward synaptic signs without necessarily equating their magnitudes [121]. 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 towards algorithms that can assign and propagate credits in more biologically palatable forms [118,122],[123].

Another widely successful approach to tuning weights is via reinforcement learning [124]. 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 [125]. 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 [126]. Promising results have also been obtained for visual tracking [127,128], face recognition [129], action recognition [130,131], video captioning [132], color enhancement [133] and object detection [134,135].

Another approach to learning structure in the visual world which does not use explicit labeled examples or a teacher that 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 was elegantly introduced in Neuroscience by Rao and Ballard’s with their predictive coding algorithm [136,137]. Predictive coding algorithms have recently re-gained momentum in the context of deep network architectures [138–141]. 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 pre-train networks in order to drastically reduce the number of examples required in subsequent supervised learning steps [142]. 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 [143].

From recognition to synthesis

Much of this review has focused on the dominant paradigm in perception, the so-called *discriminative* approach to vision which casts visual tasks as a classification problem. The alternative, the so-called *generative* approach, which can be traced all the way back to Helmholtz's description of vision as an inverse inference problem is now quickly regaining momentum. This takes on many different incarnations such as *analysis by synthesis* and *inverse graphics* [144–149]. In this framework, the goal for the visual system is to literally invert the generative process which led to the creation of retinal images in order to recover descriptions of all the objects in a scene and their spatial layout as well as estimates of the factors responsible for the generation process beyond image class labels (including shape, appearance, and pose). While these ideas have so far received little direct neuroscience support, our brains exhibit a clear ability to generate mental images and the successes of inverse graphics approaches in computer vision have prompted claims that visual recognition is accompanied by the ability to draw or generate images [150]. Whether such an ability reflects key computations involved in visual recognition or simply a by-product of these computations remains a matter of debate [151,152].

Taken to the extreme, inverse graphics approaches seem inconsistent with neurophysiology. A very basic problem is that there are simply no feedback connections that project back to the retina so there is no physical mechanism by which feedback can generate images with resolutions that match that of the retina. However, there could be coarser implementations through feedback projections to cortical areas as suggested by vision theories where V1 acts as a visual buffer [147,153]. Close your eyes and consider the following question: how many doors are there in your house? To solve this question, subjects report "mentally navigating" through a coarse rendering of their houses. This mental representation lacks the type of details invoked by inverse graphics approaches but still contains some notion of generating an internal image via feedback signals which could not be accounted for by purely bottom-up or even horizontal neural interactions.

A recent highly successful approach in image generation is the introduction of generative adversarial networks (GANs), which consist of two modules: a generator that synthesizes images and a discriminator that tries to discriminate between real and artificial images. By jointly training the two adversary networks, the discriminator becomes increasingly better at detecting "fakes" while the generator improves its forging ability to keep fooling the discriminator. This leads to highly realistic images that can even fool human observers [154,155]. It is hard to conceive how a literal implementation of generator and discriminator network circuits could be instantiated in brains.

Another approach related to inverse graphics which has received a lot of attention within the computer vision community is the capsule networks [156–158] which are extensions of CNNs to enable to explicitly represent structural information. The main idea, which is decades old and can be traced back to Biederman's geons [159,160], is to represent different objects with the same set of basic parts and their relations. In a capsule network, neural "chains" encode object parts and their structural relations through binary links in a way which is independent of the neural interconnections (or synaptic weights). Unlike CNNs where pose information is discarded through (max) pooling operations to build invariant representations and only the presence of features is represented through a single scalar value (the unit activity), capsules "encapsulate" more sophisticated representations related to an object's actual instantiation parameters (including viewpoint) in vector form.

Capsules aim to encode both the probability of an object (or object part) at a given location (as the length of a vector-valued unit) and (as the direction of that vector). Possible object transformations are stored in synaptic weights in a pose matrix and by multiplying the vector output of capsules with this pose matrix, one can encode very rich pose information, e.g, related to the position of an object given the detection of local parts. The weights of these matrices are derived from a dynamic routing algorithm whereby the ability of a lower level capsule to send its input to a higher level capsule is governed by the consistency between the top-level capsule and the low-level prediction. Such routing by agreement allows recovering what parts belong to an object by simply tracing the path of the activations along the hierarchy. So far, initial results were obtained with capsules on toy datasets [156–158,161] but more recent work has shown their potential for image classification on a subset of ImageNet [162] and action recognition datasets [163].

There is currently limited neuroscience evidence for or against the kinds of routing by agreement implemented in capsule networks. It is now clear that synapses play a much more active role in processing information than originally anticipated and there exists evidence for very short-term neural plasticity (at timescales over milliseconds to minutes) to support critical computational functions including routing [164]. In addition, the kind of vector representation implemented in capsule networks require a mechanism for the creation of small neural assemblies. One plausible neural mechanism to implement such dynamic binding of information across neurons is synchronous oscillations (e.g., [168–172]) though these theories are also contested (e.g., [173,174]). Because the degree of synchrony of neuronal spikes affects the output of downstream neurons, synchrony has been postulated to allow for gating of information transmission between network of neurons or whole cortical areas [171,175]. Moreover, the relative timing of neuronal spikes may carry information about the sensory input and the dynamic network state (e.g.,[172,176]), beyond or in addition to what is conveyed by firing rates.

As a proof of concept, Reichert & Serre have shown how aspects of spike timing could be incorporated into deep networks to build richer, more versatile representations [177]. They introduced a neural network formulation of synchrony using the framework of complex numbers and complex-valued neural units. In this framework, units are attributed both a firing rate and a phase, the latter indicating properties of spike timing with respect to some (unspecified) neural oscillations. They showed how this formulation qualitatively captures several aspects thought to be related to neuronal synchrony, including gating of information processing and dynamic binding of distributed object representations. Complex valued-neural networks offer a demonstration that it is at least possible in an architecture that involves bottom-up and top-down inference as in Deep Boltzmann Machines to bind together features that belong to the same objects [177].

Concluding remarks and future directions

A fundamental area of investigation that remains rather enigmatic is how to connect our understanding of visual computations along the ventral visual cortex to high-level cognition. For example, while examining a scene depicting kids playing in the playground, we can interpret the location, the 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 an infinite number of questions about the scene in a flexible manner. This type of general knowledge about the world can be vaguely grouped in the term "common sense", the myriad of facts and knowledge 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 processing on visual cortex 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 **Fig. 1A**, which we uploaded to one of the state-of-the-art systems for image captioning (Microsoft Caption Bot). 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, etc. 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 **Fig. 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 the 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 **Fig. 1A** and **Fig. 1B** 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.

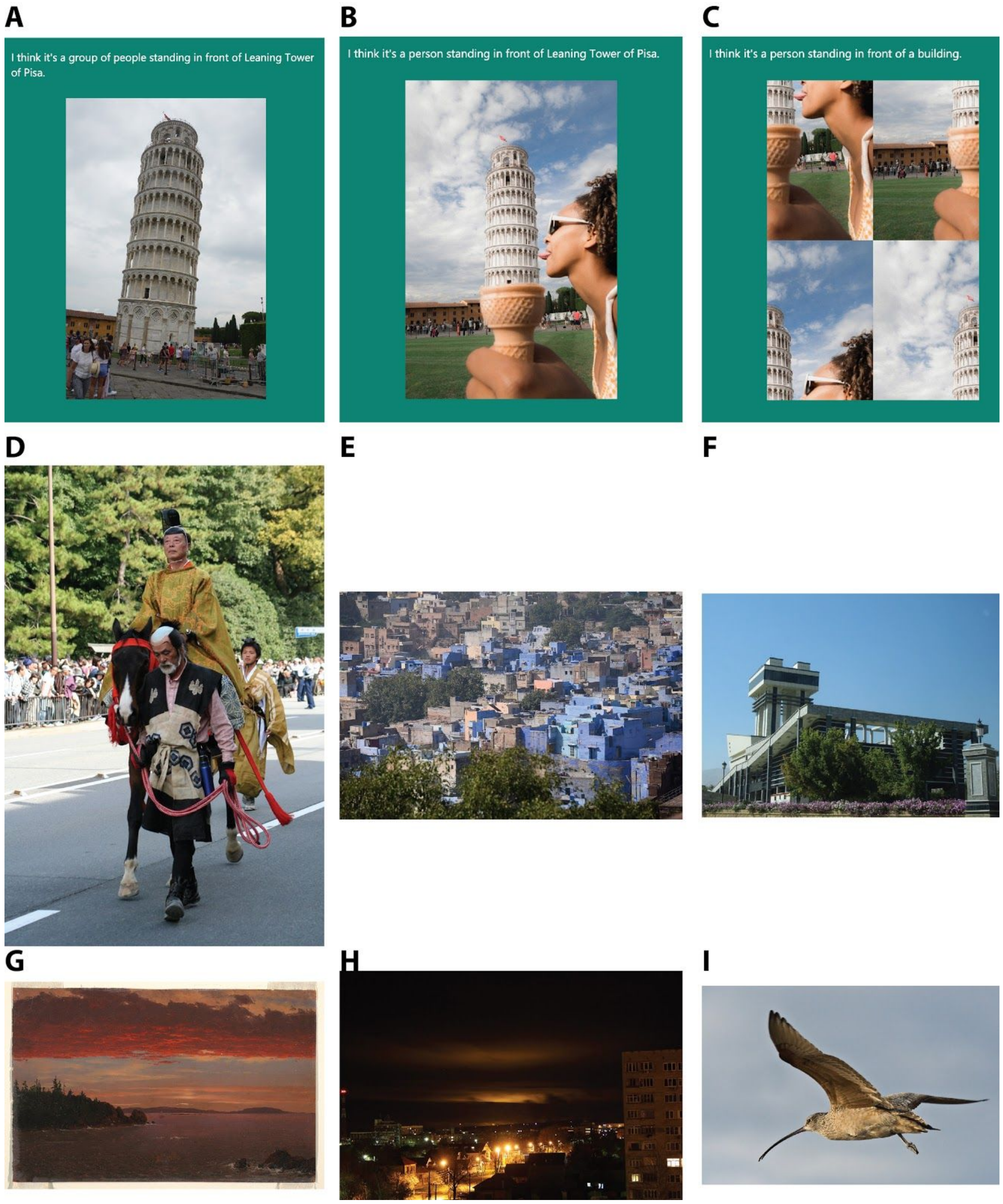
Heroic studies of the initial wave of processing in the visual cortex have led to successful computational-neuroscience models and breakthrough technologies with real-world applications. Here we have argued that the next generation of computational models will focus on the second wave of processing incorporating feedback loops. Modeling short-range interactions within visual cortex and long-range interactions between frontal areas and visual cortex, 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 [28]. Mechanisms in the visual system such as perceptual grouping, attention, and working memory exemplify how the brain learns and handles combinatorial structures in the visual environment with a small amount of experience [178]. However, exactly how attentional and mnemonic mechanisms interact with hierarchical feature representations in the visual cortex 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.

Figures and figure legends

Link to AI

Link to JPG



I think it's a group of people standing in front of Leaning Tower of Pisa.

I think it's a person standing in front of Leaning Tower of Pisa.

I think it's a person standing in front of a building.

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 Caption Bot system (https://www.captionbot.ai/). **(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, 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, a bird flying over a body of water.

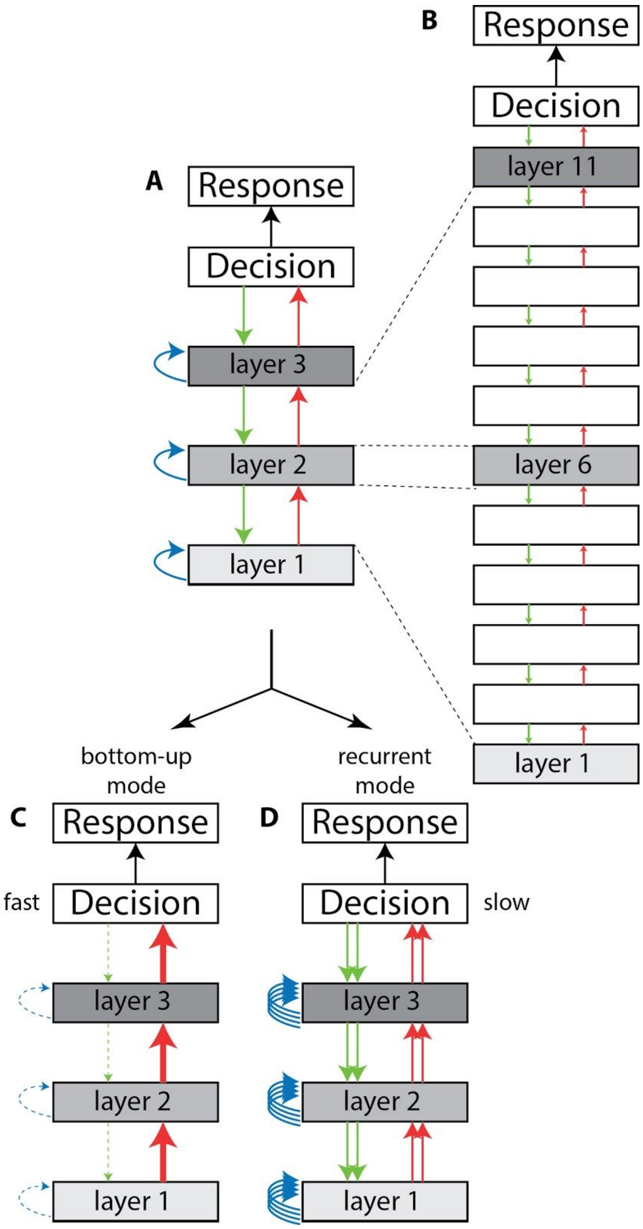


Figure 2. Recurrent networks show greater parameter efficiency and computational flexibility. **A.** Schematic illustration of a 3-layer network showing bottom-up connections (red), horizontal recurrent connections (blue), and top-down connections (green). The top layer sends signals to a decision mechanism that evaluates how confident the network is about the solution and decides whether to emit a response or send 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 4 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**).

Link to AI

Link to JPG

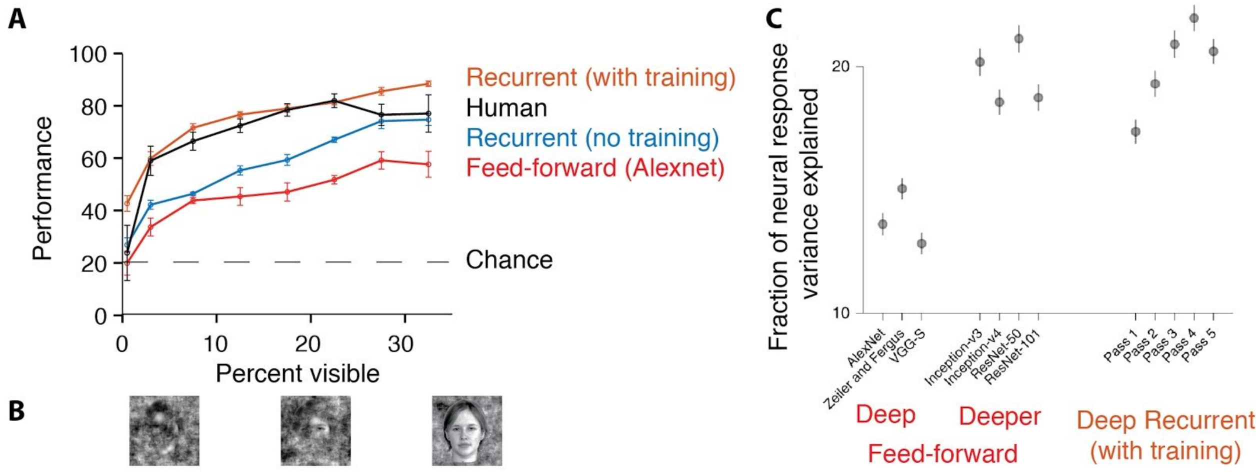


Figure 3. Recurrent networks help visual recognition. **A-B**. Recognition performance in a 5-way categorization task of partially visible objects for humans (black), layer fc7 in Alexnet (red), Alexnet network embedded with attractor-like horizontal recurrent connectivity in the fc7 layer without any training with occluded objects (blue) or with training (orange). Example objects from limited visibility to full visibility are shown in part **B**. Chance performance = 20% (dashed line). Modified from [41]. **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 [75].

Link to AI

Link to JPG

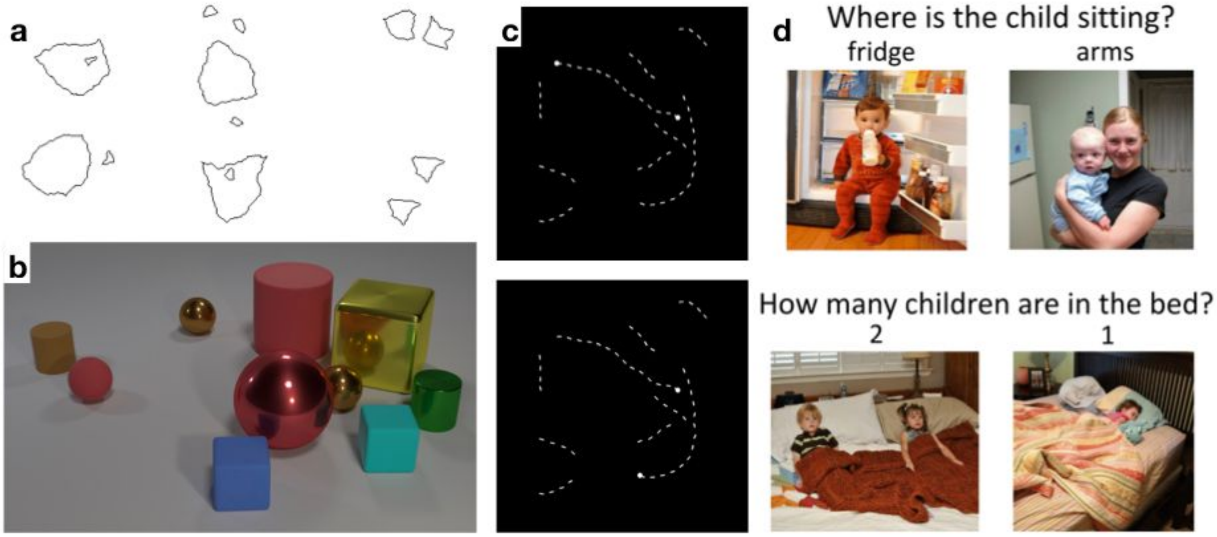

a
b
c
d
Where is the child sitting?
fridge
arms
How many children are in the bed?
2
1

Figure 4. Sample visual reasoning tasks. a) Synthetic visual reasoning test [77]. Six examples where the task is to decide whether a small shape is inside or outside a larger one. b) Visual question answering on the CLEVR challenge [81] to test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations. c) The pathfinder challenge where the task is to evaluate where the two larger white dots are connected or not [50]. d) Sample questions and answers with corresponding images from the Visual Question Answering (VQA) challenge [179].

Link to Keynote

Link to JPG

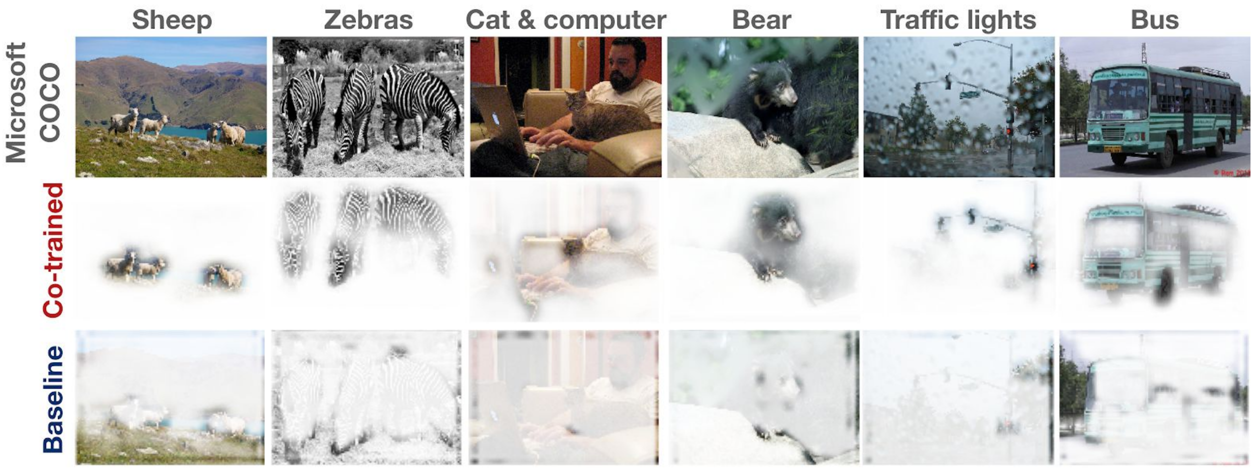


Figure 5. Learning what and where to attend. The top row depicts representative images from the Microsoft COCO dataset depicting object categories also present in ILSVRC12 (which was used for training the system). In the middle row, each of these images is shown with the transparency set to the attention map it yielded in the attention network by Linsley et al [180]trained with human supervision (see 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. Co-training 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.

Link to PDF

Link to PNG

References

1. Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex. 1991;1: 1–47.

2. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U. Complex networks: Structure and dynamics. Phys Rep. 2006;424: 175–308.

3. Markov NT, Ercsey-Ravasz MM, Ribeiro Gomes AR, Lamy C, Magrou L, Vezoli J, et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cereb Cortex. 2014;24: 17–36.

4. Lamme VA, Roelfsema PR. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci. 2000;23: 571–579.

5. Serre T, Oliva A, Poggio T. A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci U S A. 2007;104: 6424–6429.

6. VanRullen R. The power of the feed-forward sweep. Adv Cogn Psychol. 2007;3: 167–176.

7. Biederman I, Rabinowitz JC, Glass AL. On the information extracted from a glance at a scene. J Exp Psychol. 1974;103: 597–600.

8. Potter MC. Meaning in visual search. Science. 1975. pp. 565–566.

9. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature. 1996;381: 520–522.

10. Hung CP, Kreiman G, Poggio T, Dicarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. Science. 2005;2164: 863–866.

11. Liu H, Madsen JR, Agam Y, Kreiman G. Timing, Timing, Timing: Fast Decoding of Object Information from Intracranial Field Potentials in Human Visual Cortex. Neuron. 2009;62: 281–290.

12. Cauchoix M, Crouzet SM, Fize D, Serre T. Fast ventral stream neural activity enables rapid visual categorization. Neuroimage. 2016;125: 280–290.

13. VanRullen R, Thorpe SJ. Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. Perception. 2001;30: 655–668.

14. Fize D, Cauchoix M, Fabre-Thorpe M. Humans and monkeys share visual representations. Proc Natl Acad Sci U S A. 2011;108: 7635–7640.

15. Rajalingham R, Schmidt K, DiCarlo JJ. Comparison of Object Recognition Behavior in Human and Monkey. J Neurosci. 2015;35: 12127–12136.

16. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999;2.

17. Serre T. Hierarchical models of the visual system. Encyclopedia of computational neuroscience. Springer; 2015; Available: https://link.springer.com/content/pdf/10.1007/978-1-4614-6675-8_345.pdf

18. Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T. A quantitative theory of immediate visual recognition. Prog Brain Res. 2007;165: 33–56.

19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;521: 436–444.

20. Eberhardt S, Cader JG, Serre T. How Deep is the Feature Analysis underlying Rapid Visual Categorization? In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. Advances in Neural Information Processing Systems 29. Curran Associates, Inc.; 2016. pp. 1100–1108.

21. Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. Sci Rep. 2016;6: 32672.

22. Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol. Public Library of Science; 2014;10: e1003963.

23. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A. 2014;111: 8619–8624.

24. Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. 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. 2018;38: 7255–7269.

25. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Computer Vision and Pattern Recognition. 2016. Available: http://arxiv.org/abs/1512.03385

26. Kemelmacher-Shlizerman I, Seitz SM, Miller D, Brossard E. The megaface benchmark: 1 million faces for recognition at scale. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 4873–4882.

27. Serre T. Deep learning: The good, the bad and the uggly. Annual review of visual neuroscience (in press). 2019;

28. Geman D, Geman S, Hallonquist N, Younes L. Visual Turing test for computer vision systems. Proc Natl Acad Sci U S A. 2015;112: 3618–3623.

29. Shepard RN, Metzler J. Mental rotation of three-dimensional objects. Science. 1971. pp. 701–703.

30. Ullman S. High-level vision: Object recognition and visual cognition. Cambridge, MA: The MIT Press; 1996.

31. Roelfsema PR, Lamme VA, Spekreijse H. The implementation of visual routines. Vision Res. 2000;40: 1385–1411.

32. Malik J, Perona P. Preattentive texture discrimination with early vision mechanisms. J Opt Soc Am A. 1990;7: 923–932.

33. Hadji I, Wildes RP. A spatiotemporal oriented energy network for dynamic texture recognition. Proceedings of the IEEE International Conference on Computer Vision. 2017. pp. 3066–3074.

34. Bruna J, Mallat S. Invariant scattering convolution networks. IEEE Trans Pattern Anal Mach Intell. 2013;35: 1872–1886.

35. Liao Q, Poggio T. Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex [Internet]. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1604.03640

36. Guo Q, Yu Z, Wu Y, Liang D, Qin H, Yan J. Dynamic Recursive Neural Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. pp. 5147–5156.

37. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. Annual Meeting of the Association for Computational Linguistics. 2019. Available: http://arxiv.org/abs/1906.02243

38. Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI [Internet]. arXiv [cs.CY]. 2019. Available: http://arxiv.org/abs/1907.10597

39. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge [Internet]. International Journal of Computer Vision. 2015. pp. 211–252. doi:10.1007/s11263-015-0816-y

40. Nakamura H, Gattass R, Desimone R, Ungerleider LG. The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. J Neurosci. 1993;13: 3681–3691.

41. Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Caro JO, et al. Recurrent computations for visual pattern completion [Internet]. Proceedings of the National Academy of Sciences. 2018. pp. 8835–8840. doi:10.1073/pnas.1719397115

42. Srivastava RK, Greff K, Schmidhuber J. Training Very Deep Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. pp. 2377–2385.

43. Graves A. Adaptive Computation Time for Recurrent Neural Networks [Internet]. arXiv [cs.NE]. 2016. Available: http://arxiv.org/abs/1603.08983

44. Zilly JG, Srivastava RK, Koutník J, Schmidhuber J. Recurrent Highway Networks. Proceedings of the 34th International Conference on Machine Learning - Volume 70. Sydney, NSW, Australia: JMLR.org; 2017. pp. 4189–4198.

45. Lamme VA, Zipser K, Spekreijse H. Masking interrupts figure-ground signals in V1. J Cogn Neurosci. 2002;14: 1044–1053.

46. Breitmeyer B, Ogmen H. Visual Masking: Time slices through conscious and unconscious vision. Oxford Psychology Series; 2006.

47. Fahrenfort JJ, Scholte HS, Lamme VAF. Masking disrupts reentrant processing in human visual cortex. J Cogn Neurosci. 2007;19: 1488–1497.

48. Macknik SL, Martinez-conde S. The role of feedback in visual masking and visual processing. Advances. 2007;3: 125–153.

49. Hegdé J, Felleman DJ. Reappraising the functional implications of the primate visual anatomical hierarchy. Neuroscientist. 2007;13: 416–421.

50. Linsley D, Kim JK, Veerabadran V, Windolf C, Serre T. Learning long-range spatial dependencies with horizontal gated recurrent units. Neural Information Processing Systems (NIPS). 2018. Available: https://nips.cc/Conferences/2018/Schedule?showEvent=11042

51. Kim J, Linsley D, Thakkar K, Serre T. Disentangling neural mechanisms for perceptual grouping [Internet]. arXiv [cs.CV]. 2019. Available: http://arxiv.org/abs/1906.01558

52. Van Den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel Recurrent Neural Networks. Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. New York, NY, USA: JMLR.org; 2016. pp. 1747–1756.

53. O'Reilly RC, Wyatte D, Herd S, Mingus B, Jilk DJ. Recurrent Processing during Object Recognition. Front Psychol. 2013;4: 1–14.

54. Liang M, Hu X. Recurrent convolutional neural network for object recognition. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society; 2015. pp. 3367–3375.

55. Zamir AR, Wu T, Sun L, Shen WB, Shi BE, Malik J, et al. Feedback Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 1808–1817.

56. Anthony M, Bartlett PL. Neural Network Learning: Theoretical Foundations. Cambridge University Press; 2009.

57. Neyshabur B, Bhojanapalli S, McAllester D, Srebro N. Exploring Generalization in Deep Learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. USA: Curran Associates Inc.; 2017. pp. 5949–5958.

58. Krizhevsky A, Hinton G, Others. Learning multiple layers of features from tiny images [Internet]. Citeseer; 2009. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf

59. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization [Internet]. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1611.03530

60. Akpinar N-J, Kratzwald B, Feuerriegel S. Sample Complexity Bounds for Recurrent Neural Networks with Application to Combinatorial Graph Problems [Internet]. arXiv [stat.ML]. 2019. Available: http://arxiv.org/abs/1901.10289

61. Linsley D, Kim J, Serre T. Sample-efficient image segmentation through recurrence [Internet]. arXiv [cs.CV]. 2018. Available: http://arxiv.org/abs/1811.11356

62. Biederman I, Gerhardstein PC. Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance [Internet]. Journal of Experimental Psychology: Human Perception and Performance. 1993. pp. 1162–1182. doi:10.1037//0096-1523.19.6.1162

63. Sinha P. Recognizing complex patterns. Nat Neurosci. 2002;5 Suppl: 1093–1097.

64. Geirhos R, Temme CRM, Rauber J, Schütt HH, Bethge M, Wichmann FA. Generalisation in humans and deep neural networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31. Curran Associates, Inc.; 2018. pp. 7549–7561.

65. Linsley D, Kim J, Berson D, Serre T. Robust neural circuit reconstruction from serial electron microscopy with convolutional recurrent networks [Internet]. arXiv [cs.CV]. 2018. Available: http://arxiv.org/abs/1811.11356

66. Tang H, Kreiman G. Recognition of Occluded Objects. In: Zhao Q, editor. Computational and Cognitive Neuroscience of Vision. Singapore: Springer Singapore; 2017. pp. 41–58.

67. El-Shamayleh Y, Fyall AM, Pasupathy A. The role of visual area V4 in the discrimination of partially occluded shapes. Journal of. Soc Neuroscience; 2014; Available: http://www.jneurosci.org/content/34/25/8570.short

68. Tang H, Buia C, Madhavan R, Crone NE, Madsen JR, Anderson WS, et al. Spatiotemporal Dynamics Underlying Object Completion in Human Ventral Visual Cortex [Internet]. Neuron. 2014. pp. 736–748. doi:10.1016/j.neuron.2014.06.017

69. Rosenfeld A, Zemel R, Tsotsos JK. The Elephant in the Room [Internet]. arXiv [cs.CV]. 2018. Available: http://arxiv.org/abs/1808.03305

70. Wang J, Zhang Z, Xie C, Zhou Y, Premachandran V, Zhu J, et al. Visual concepts and compositional voting. Annals of Mathematical Sciences and Applications. 2018;3: 151–188.

71. Wyatte D, Jilk DJ, O'Reilly RC. Early recurrent feedback facilitates visual object recognition under challenging conditions. Front Psychol. 2014;5: 674.

72. Fyall AM, El-Shamayleh Y, Choi H, Shea-Brown E, Pasupathy A. Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. Elife. 2017;6. doi:10.7554/eLife.25784

73. Peterson MA, Harvey EM, Weidenbacher HJ. Shape recognition contributions to figure-ground reversal: which route counts? J Exp Psychol Hum Percept Perform. 1991;17: 1075–1089.

74. Vecera SP, Farah MJ. Is visual image segmentation a bottom-up or an interactive process? Percept Psychophys. 1997;59: 1280–1296.

75. Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior [Internet]. Nature Neuroscience. 2019. doi:10.1038/s41593-019-0392-5

76. Herrnstein RJ, Vaughan W Jr, Mumford DB, Kosslyn SM. Teaching pigeons an abstract relational rule: insideness. Percept Psychophys. 1989;46: 56–64.

77. Fleuret F, Li T, Dubout C, Wampler EK, Yantis S, Geman D. Comparing machines and humans on a visual categorization test. Proc Natl Acad Sci U S A. 2011;108: 17621–17625.

78. Ellis K, Solar-Lezama A, Tenenbaum J. Unsupervised Learning by Program Synthesis. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. pp. 973–981.

79. Stabinger S, Rodríguez-Sánchez A, Piater J. 25 Years of CNNs: Can We Compare to Human Abstraction Capabilities? Artificial Neural Networks and Machine Learning – ICANN 2016. Springer International Publishing; 2016. pp. 380–387.

80. Kim JK, Ricci M, Serre T. Not-So-CLEVR: learning same–different relations strains feedforward neural networks. Interface Focus theme issue on "Understanding images in biological and computer vision." 2018;

81. Johnson J, Hariharan B, d. Maaten L v., Fei-Fei L, Zitnick CL, Girshick R. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 1988–1997.

82. Donderi DC, Zelnicker D. Parallel processing in visual same-different decisions. Percept Psychophys. 1969;5: 197–200.

83. Giurfa M, Zhang S, Jenett A, Menzel R, Srinivasan MV. The concepts of "sameness" and "difference" in an insect. Nature. 2001;410: 930–933.

84. Wasserman EA, Castro L, Freeman JH. Same-different categorization in rats. Learn Mem. 2012;19: 142–145.

85. Martinho A, Kacelnik A. Ducklings imprint on the relational concept of "same or different." Science. American Association for the Advancement of Science; 2016;353: 286–288.

86. Logan GD. Spatial attention and the apprehension of spatial relations. J Exp Psychol Hum Percept Perform. 1994;20: 1015–1036.

87. Rosielle LJ, Crabb BT, Cooper EE. Attentional coding of categorical relations in scene perception: evidence from the flicker paradigm. Psychon Bull Rev. 2002;9: 319–326.

88. Franconeri SL, Scimeca JM, Roth JC, Helseth S a., Kahn LE. Flexible visual processing of spatial relationships. Cognition. Elsevier B.V.; 2012;122: 210–227.

89. Hayworth KJ, Lescroart MD, Biederman I. Neural encoding of relative position. J Exp Psychol Hum Percept Perform. 2011;37: 1032–1050.

90. Kroger JK, Sabb FW, Fales CL, Bookheimer SY, Cohen MS, Holyoak KJ. Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. Cereb Cortex. 2002;12: 477–485.

91. Golde M, von Cramon DY, Schubotz RI. Differential role of anterior prefrontal and premotor cortex in the processing of relational information. Neuroimage. 2010;49: 2890–2900.

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

93. Riesenhuber M, Poggio T. Are Cortical Models Really Bound by the “ Binding Problem ”? Neuron. 1999;24: 87–93.

94. Stollenga M, Masci J, Gomez F, Schmidhuber J. Deep Networks with Internal Selective Attention through Feedback Connections. arXiv preprint arXiv: …. 2014; 13.

95. Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent Models of Visual Attention. Advances in Neural Information Processing Systems 27. 2014;27: 1–9.

96. Cao C, Liu X, Yang Y, Yu Y, Wang J, Wang Z, et al. Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks. 2015 IEEE International Conference on Computer Vision (ICCV). 2015. pp. 2956–2964.

97. You Q, Jin H, Wang Z, Fang C, Luo J. Image Captioning with Semantic Attention. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. pp. 4651–4659.

98. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, et al. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 6298–6306.

99. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, et al. Residual Attention Network for Image Classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 6450–6458.

100. Biparva M, Tsotsos J. STNet: Selective Tuning of Convolutional Networks for Object Localization. The IEEE International Conference on Computer Vision (ICCV). 2017. Available: http://openaccess.thecvf.com/content_ICCV_2017_workshops/papers/w40/Biparva_STNet_Selective_Tuning_ICCV_2017_paper.pdf

101. Itti L, Rees G, Tsotsos JK. Neurobiology of attention. Academic Press; 2005.

102. Wolfe JM, Gray W. Guided search 4.0. Integrated models of cognitive systems. 2007; 99–119.

103. Bichot NP, Heard MT, DeGennaro EM, Desimone R. A Source for Feature-Based

Attention in the Prefrontal Cortex. Neuron. 2015;88: 832–844.

104. Moore T, Zirnsak M. Neural Mechanisms of Selective Visual Attention [Internet]. Annual Review of Psychology. 2017. pp. 47–72. doi:10.1146/annurev-psych-122414-033400

105. Zhang M, Feng J, Ma KT, Lim JH, Zhao Q, Kreiman G. Finding any Waldo with zero-shot invariant and efficient visual search [Internet]. Nature Communications. 2018. doi:10.1038/s41467-018-06217-x

106. Adeli H, Zelinsky G. Deep-BCN: Deep Networks Meet Biased Competition to Create a Brain-Inspired Model of Attention Control [Internet]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018. doi:10.1109/cvprw.2018.00259

107. Lindsay GW, Miller KD. How biological attention mechanisms improve task performance in a large-scale visual system model. Elife. 2018;7. doi:10.7554/eLife.38105

108. Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014. pp. 580–587.

109. Girshick R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision. 2015. pp. 1440–1448.

110. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. pp. 91–99.

111. Dai J, Li Y, He K, Sun J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. Advances in Neural Information Processing Systems 29. Curran Associates, Inc.; 2016. pp. 379–387.

112. Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 6517–6525.

113. Eckstein MP, Koehler K, Welbourne LE, Akbas E. Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. Curr Biol. 2017;27: 2827–2832.e3.

114. Linsley D, Eberhardt S, Sharma T, Gupta P, Serre T. What are the visual features underlying human versus machine vision? IEEE ICCV Workshop on the Mutual Benefit of Cognitive and Computer Vision. 2017. Available: http://openaccess.thecvf.com/content_ICCV_2017_workshops/papers/w40/Linsley_What_Are_the_ICCV_2017_paper.pdf

115. Linsley D, Shiebler D, Eberhardt S, Serre T. Learning what and where to attend [Internet]. In ICLR. 2019. Available: https://openreview.net/pdf?id=BJgLg3R9KQ

116. Rumelhart DE, Hinton GE, McClelland JL. A general framework for parallel distributed processing. Mit Press Computational Models Of Cognition And Perception …. 1986;

Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=11243779389490373524

117. Crick F. The recent excitement about neural networks. Nature. 1989;337: 129–132.

118. Bengio Y, Lee D-H, Bornschein J, Mesnard T, Lin Z. Towards Biologically Plausible Deep Learning [Internet]. arXiv [cs.LG]. 2015. Available: http://arxiv.org/abs/1502.04156

119. Burbank KS, Kreiman G. Depression-Biased Reverse Plasticity Rule Is Required for Stable Learning at Top-Down Connections [Internet]. PLoS Computational Biology. 2012. p. e1002393. doi:10.1371/journal.pcbi.1002393

120. Lillicrap TP, Cownden D, Tweed DB, Akerman CJ. Random synaptic feedback weights support error backpropagation for deep learning. Nat Commun. 2016;7: 13276.

121. Liao Q, Leibo JZ, Poggio T. How important is weight symmetry in backpropagation? Thirtieth AAAI Conference on Artificial Intelligence. 2016. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/12325

122. Guerguiev J, Lillicrap TP, Richards BA. Towards deep learning with segregated dendrites. Elife. 2017;6. doi:10.7554/eLife.22901

123. Miconi T. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. Elife. 2017;6. doi:10.7554/eLife.20899

124. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. MIT Press; 2018.

125. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529: 484–489.

126. Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps. 2015 AAAI Fall Symposium Series. aaai.org; 2015; Available: https://www.aaai.org/ocs/index.php/FSS/FSS15/paper/viewPaper/11673

127. Ren L, Yuan X, Lu J, Yang M, Zhou J. Deep reinforcement learning with iterative shift for visual tracking. Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 684–700.

128. Guo M, Lu J, Zhou J. Dual-agent deep reinforcement learning for deformable face tracking. Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 768–783.

129. Rao Y, Lu J, Zhou J. Attention-aware deep reinforcement learning for video face recognition. Proceedings of the IEEE International Conference on Computer Vision. 2017. pp. 3931–3940.

130. Tang Y, Tian Y, Lu J, Li P, Zhou J. Deep progressive reinforcement learning for skeleton-based action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp. 5323–5332.

131. Chen L, Lu J, Song Z, Zhou J. Part-activated deep reinforcement learning for action prediction. Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 421–436.

132. Wang X, Chen W, Wu J, Wang Y-F, Yang Wang W. Video captioning via hierarchical reinforcement learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp. 4213–4222.

133. Park J, Lee J-Y, Yoo D, So Kweon I. Distort-and-recover: Color enhancement using deep reinforcement learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp. 5928–5936.

134. Kong X, Xin B, Wang Y, Hua G. Collaborative deep reinforcement learning for joint object search. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. pp. 1695–1704.

135. Rao Y, Lin D, Lu J, Zhou J. Learning globally optimized object detector via policy gradient. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp. 6190–6198.

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

137. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. Neuron. 2012;76: 695–711.

138. Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:160508104. arxiv.org; 2016; Available: http://arxiv.org/abs/1605.08104

139. Vondrick C, Pirsiavash H, Torralba A. Anticipating Visual Representations from Unlabeled Video [Internet]. arXiv [cs.CV]. 2015. Available: http://arxiv.org/abs/1504.08023

140. O'Reilly RC, Wyatte DR, Rohrlich J. Deep Predictive Learning: A Comprehensive Model of Three Visual Streams [Internet]. arXiv [q-bio.NC]. 2017. Available: http://arxiv.org/abs/1709.04654

141. Wen H, Han K, Shi J, Zhang Y, Culurciello E, Liu Z. Deep Predictive Coding Network for Object Recognition [Internet]. arXiv [cs.CV]. 2018. Available: http://arxiv.org/abs/1802.04762

142. van den Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding [Internet]. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1807.03748

143. Lotter W, Kreiman G, Cox D. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception [Internet]. arXiv [q-bio.NC]. 2018. Available: http://arxiv.org/abs/1805.10734

144. Grenander U. Pattern Synthesis: Lectures in Pattern Theory Volume 1. Springer, New York, NY; 1976.

145. Grenander U. Pattern Analysis: Lectures in Pattern Theory Volume II. Springer, New York, NY; 1978.

146. Yuille A, Kersten D. Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci. 2006;10: 301–308.

147. Lee TS, Mumford D, Romero R, Lamme VA. The role of the primary visual cortex in higher level vision. Vision Res. 1998;38: 2429–2454.

148. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis. 2003;20: 1434–1448.

149. Olshausen BA. Perception as an Inference Problem. In: Gazzaniga VM, Mangun R, editors. The Cognitive Neurosciences. MIT Press. 2013.

150. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;350: 1332–1338.

151. Finke RA. Theories Relating Mental Imagery to Perception. Psychol Bull. 1985;98: 236–259.

152. Behrmann M, Winocur G, Moscovitch M. Dissociation between mental imagery and object recognition in a brain-damaged patient. Nature. 1992;359: 636–637.

153. Kosslyn SM. Image and Mind. Harvard University Press; 1980.

154. Brock A, Donahue J, Simonyan K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations. 2019.

155. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Computer Vision and Pattern Recognition Conference. 2019; Available: http://arxiv.org/abs/1812.04948

156. Hinton GE, Krizhevsky A, Wang SD. Transforming Auto-Encoders. Artificial Neural Networks and Machine Learning – ICANN 2011. Springer Berlin Heidelberg; 2011. pp. 44–51.

157. Sabour S, Frosst N, Hinton GE. Dynamic Routing Between Capsules [Internet]. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1710.09829

158. Hinton GE, Sabour S, Frosst N. Matrix capsules with EM routing [Internet]. 2018. Available: https://openreview.net/pdf?id=HJWLfGWRb

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

160. Hummel JE, Biederman I. Dynamic binding in a neural network for shape recognition. 1992. pp. 480–517.

161. Lenssen JE, Fey M, Libuschewski P. Group Equivariant Capsule Networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in

Neural Information Processing Systems 31. Curran Associates, Inc.; 2018. pp. 8844–8853.

162. Li H, Guo X, DaiWanli Ouyang B, Wang X. Neural network encapsulation. Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 252–267.

163. Duarte K, Rawat Y, Shah M. VideoCapsuleNet: A Simplified Network for Action Detection. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31. Curran Associates, Inc.; 2018. pp. 7610–7619.

164. Abbott LF, Regehr WG. Synaptic computation. Nature. 2004. pp. 796–803.

165. Gilbert CD, Li W, Pie V, Piech V. Contour Saliency in Primary Visual Cortex. Neuron. 2006;50: 951–962.

166. Pooresmaeili A, Poort J, Roelfsema PR. Simultaneous selection by object-based attention in visual and frontal cortex. Proc Natl Acad Sci U S A. 2014;111: 6467–6472.

167. Roelfsema PR, Spekreijse H. The representation of erroneously perceived stimuli in the primary visual cortex. Neuron. 2001;31: 853–863.

168. von der Malsburg C. The Correlation Theory of Brain Function [Internet]. 1981. Available: http://cogprints.org/1380/5/vdM_correlation.ps

169. Crick F. Function of the thalamic reticular complex: the searchlight hypothesis. Proc Natl Acad Sci U S A. 1984;81: 4586–4590.

170. Singer W, Gray CM. Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci. 1995;18: 555–586.

171. Fries P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends Cogn Sci. 2005;9: 474–480.

172. Stanley GB. Reading and writing the neural code. Nat Neurosci. 2013;16: 259–263.

173. Shadlen MN, Movshon JA. Synchrony unbound: A critical evaluation of the temporal binding hypothesis. Neuron. 1999;24: 67–77.

174. Ray S, Maunsell JHR. Differences in gamma frequencies across visual cortex restrict their possible use in computation. Neuron. Elsevier Inc.; 2010;67: 885–896.

175. Benchenane K, Tiesinga PH, Battaglia FP. Oscillations in the prefrontal cortex: a gateway to memory and attention. Curr Opin Neurobiol. 2011;21: 475–485.

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

177. Reichert DP, Serre T. Neuronal synchrony in complex-valued deep networks. arXiv preprint arXiv:13126115. arxiv.org; 2013; Available: http://arxiv.org/abs/1312.6115

178. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. How to grow a mind: statistics,

structure, and abstraction. Science. 2011;331: 1279–1285.

179. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. pp. 6904–6913.

180. Linsley D, S E, Shiebler D, Serre T. Learning what and where to attend. International Conference on Learning Representations. 2019.