HHS Public Access

Author manuscript

Nat Neurosci. Author manuscript; available in PMC 2021 August 20.

Published in final edited form as:

Nat Neurosci. 2019 January ; 22(1): 15–24. doi:10.1038/s41593-018-0284-0.

Stimulus- and goal-oriented frameworks for understanding natural vision

Maxwell H. Turner[1,2,†], **Luis Gonzalo Sanchez Giraldo**[3,†], **Odelia Schwartz**[3,‡], **Fred Rieke**[1,‡,*]

[1]Department of Physiology and Biophysics, University of Washington, Seattle, Washington 98195, USA

[2]Graduate Program in Neuroscience, University of Washington, Seattle, Washington 98195, USA

[3]Department of Computer Science, University of Miami, Coral Gables, FL, 33146, USA

Abstract

Our knowledge of sensory processing has advanced dramatically in the last few decades, but this understanding remains far from complete, especially for stimuli with the large dynamic range and strong temporal and spatial correlations characteristic of natural visual inputs. Here we describe some of the issues that make understanding the encoding of natural image stimuli challenging. We highlight two broad strategies for approaching this problem: a stimulus-oriented framework and a goal-oriented one. Different contexts can call for one or the other framework. Looking forward, recent advances, particularly those based in machine learning, show promise in borrowing key strengths of both frameworks and by doing so illuminating a path to a more comprehensive understanding of the encoding of natural stimuli.

Introduction

The neural circuits that process sensory inputs are shaped by the properties of the stimuli they encounter as well as the behavioral demands of the animal. Because of this, a deep understanding of sensory circuits and the computations they support requires connecting what we know about sensory systems to properties of natural stimuli. In this review, we discuss some of the progress and the challenges in describing the neural encoding of complex stimuli such as those encountered in the real world; related issues extend to many areas beyond neurophysiology. We refer to the encoding of visual scenes as a paradigmatic example, but many of the same issues arise in other sensory modalities.

Progress in studying sensory coding has traditionally relied on parameterized, artificial stimuli designed to isolate and characterize specific circuit mechanisms, such as nonlinearities in the integration of signals across space (reviewed by (Gollisch and Meister, 2010; Schwartz and Rieke, 2011)) or adaptation to changes in particular stimulus properties such as intensity, contrast, or orientation (reviewed by (Demb and Singer, 2015; Graham, 2011; Rieke and Rudd, 2009; Solomon and Kohn, 2014)). These approaches have revealed

*Correspondence: rieke@uw.edu.
†Co-first authors
‡Co-senior authors

the mechanistic basis of many important circuit computations but have not led to a clear understanding of the encoding of natural stimuli.

Two issues make studying the encoding of natural stimuli challenging compared to typical artificial stimuli. First, complex stimuli, such as natural visual inputs, engage a host of interacting circuit mechanisms rather than individual mechanisms in isolation. This complexity is difficult to capture with computational models. For example, many predictive neurobiological models for stimulus-response transformations in the early visual system are based on a common architecture: linear filtering over space and time, followed by a (generally time-dependent) nonlinearity. These models do not generalize well to capture responses to inputs other than those to which they were fit (Carandini et al., 2005; Heitman et al., 2016). Natural images can highlight such failures of generalization (David and Gallant, 2005; Turner and Rieke, 2016; Turner et al., 2018). Incorporating known circuit features, or stacking multiple Linear-Nonlinear layers, can improve generalization in both retina and V1 (David and Gallant, 2005; Maheswaranathan et al., 2017; Mcintosh et al., 2016). Stacked computations are also a central element of Deep Neural Network models for modeling higher cortical areas (Yamins and DiCarlo, 2016), as described in more detail below.

A second challenge inherent in the study of natural stimulus encoding is the complex statistics of natural scenes (reviewed by (Hyvärinen, 2010; Lewicki et al., 2014; Simoncelli and Olshausen, 2001; Zhaoping, 2014)). For example, across different visual scenes and even within a single scene, image statistics (e.g. mean intensity, spatial contrast, and other, higher-order statistics) can vary widely but (fortunately) not randomly (Coen-Cagli et al., 2012; Frazor and Geisler, 2006; Karklin and Lewicki, 2005; Parra et al., 2001; Ruderman and Bialek, 1994). Within a single visual scene, different image features are often strongly correlated, which makes it difficult to relate a neural response to a particular feature of a scene (see (Sharpee et al., 2006) for a computational approach to this issue). One approach to managing this complexity is to develop generative models of natural images that enable a low dimensional representation. Parametric models exist for naturalistic textures (Portilla and Simoncelli, 2000) -- i.e. semi-regular, repeating patterns (see Figure 1) -- and recent advances in machine learning show promise in generating not only textures (Gatys et al., 2015) but non-homogeneous naturalistic images (Arjovsky et al., 2017; Karras et al., 2018); for applications of these approaches see (Freeman et al., 2013; Okazawa et al., 2015; Rust and DiCarlo, 2010).

There is a long history of studying the encoding of natural scenes in neurophysiology experiments (e.g. (Baddeley et al., 1997; Creutzfeldt and Nothdurft, 1978; Smyth et al., 2003; Stanley et al., 1999; Vickers et al., 2001; Vinje and Gallant, 2000)), and recent years have seen this interest expand (e.g., (Sharpee et al., 2006; Theunissen and Elie, 2014; Zwicker et al., 2016) and references therein). Normative (“why”) models can incorporate knowledge about the statistical structure of natural scenes or the task goals and hence add an additional perspective on top of descriptive and mechanistic models (Dayan and Abbott, 2001).

Stimulus- and goal-oriented approaches to natural stimulus encoding

We will focus on two theoretical frameworks that are often appealed to in the study of natural stimulus encoding:

i. A *stimulus-oriented* (or efficient coding) framework, that formalizes the idea that a major goal of sensory processing is to encode as much information as possible about a stimulus using limited resources. This framework identifies transformations of sensory input signals that reduce statistical redundancies present in the natural world. These transformations often rely on general purpose computations that care about the information available in a stimulus and are blind to the specific uses of this information. Complementary approaches based on generative modeling seek to capture the statistical dependencies of natural scenes, and by doing so reveal how they can be reduced. Stimulus-oriented approaches are closely related to unsupervised machine learning, for which learning is based only on properties of the input and does not require a specific goal such as object recognition.

ii. A *goal-oriented* framework, which appeals to the computational or behavioral goal of the circuit or animal. Unlike stimulus-oriented approaches, goal-oriented approaches explicitly treat some features of the stimulus differently than others, and which features are encoded depends on the desired behavioral output or goal. These approaches include recent advances in Deep Convolutional Neural Networks, particularly those based on supervised, discriminative learning from large databases of images with identified and labeled objects.

These two frameworks may appear to be at odds. For instance, a model focused solely on a high-level goal like object recognition will not necessarily reduce redundancies or capture general statistical properties of the stimulus. Conversely, models focusing on reducing redundancies are not likely to explain, at least not explicitly, complex tasks such as object recognition. Historically, stimulus-oriented frameworks have largely been applied to early visual areas and goal-oriented objectives to later cortical areas. But these boundaries are beginning to blur. Indeed, in some cases the two approaches can be seen as complementary. For instance, even well-established visual computations like lateral inhibition can be seen through both lenses: as a mechanism to suppress responses to low spatial frequencies and eliminate some of the redundancies present in natural images (Atick and Redlich, 1992; Srinivasan et al., 1982), or as a way to facilitate the detection of specific features of a scene, namely edges (Marr and Hildreth, 1980). We will discuss some modern computational approaches that may facilitate the merger of these two frameworks, allowing one to inform the other and vice-versa. In particular, deep neural networks provide a promising route for exploring how stimulus- and goal-oriented constraints together shape sensory processing.

Stimulus-oriented approaches to natural vision

An influential hypothesis that undergirds much of the study of natural scene processing is the “efficient coding hypothesis,” first proposed by Barlow (Barlow, 2001, 1961) (see

also (Attneave, 1954)), and influenced by Shannon's earlier work on information theory (Shannon, 1948).

Barlow proposed that an efficient coding scheme should reduce the redundancy of natural inputs, but without loss of the information that is encoded (Barlow, 1961). Redundancy as defined by Barlow is the fraction of the total information carrying capacity of a neuron or neural population that is not used to transmit information about the stimulus. Efficient coding has been contrasted with a "sparse" neural code, where the goal is not defined as redundancy reduction, but rather to produce a sparse representation of natural inputs (Field, 1994).

Redundancy reduction predicts that a single noiseless neuron should distribute its responses uniformly (e.g., subject to a constraint on the maximal firing rate), such that each possible response occurs with equal frequency; to do otherwise would mean that the neuron is not making full use of its dynamic range. Examples of approximately uniformly-distributed sensory representations can be found in a variety of sensory systems (Bhandawat et al., 2007; Laughlin, 1981). Consideration of neural noise can substantially alter predictions of efficient coding because in that case efficiency involves both using a cell's full response range and mitigating the effect of noise (Brinkman et al., 2016; Gjorgjieva et al., 2014; Kastner et al., 2015).

Redundancy reduction in a population of neurons (i.e., multiple channels) relies on removing statistical dependencies among their responses (Barlow, 1961). Reducing redundancy for natural stimuli is particularly challenging because natural visual inputs contain strong (nonlinear) statistical regularities across time and space (for a review, see (Simoncelli and Olshausen, 2001)). We start by describing the application of these ideas in early sensory areas (largely the retina) and then turn to efficient coding in visual cortex.

Efficient coding and second order statistics

Second-order spatial correlations in natural scenes have been a particular focus of efficient coding approaches. Such correlations, on average, obey a power law scaling: the power spectrum of spatial frequencies falls as the inverse of the square of the spatial frequency (Figure 2b) (Field, 1987). This is the result of the scale invariance of natural images -- i.e. many statistical properties are unchanged by magnifying or demagnifying an image (Ruderman and Bialek, 1994). Scale invariance has been suggested to result from the fact that objects can appear at any distance from an observer (Ruderman, 1994).

The prevalence of low spatial frequencies in natural images produces correlated responses in nearby cells, leading to a redundant population code. Receptive field surrounds of neurons in retina and LGN decorrelate responses of nearby neurons by suppressing responses to low spatial frequencies. This is sometimes referred to as "whitening" (Atick and Redlich, 1992; Dan et al., 1996) (but see (Franke et al., 2017; Pitkow and Meister, 2012; Vincent and Baddeley, 2003)). Whitening, however, will increase high spatial frequency noise such as that in photoreceptor signals; consideration of noise predicts that the suppressive surround should be minimal or absent when noise is high (for a review, see (Atick, 1992; Zhaoping,

2006)). Similar principles of whitening without amplifying noise have also been proposed in other domains, such as stereo coding in cortex (Li and Atick, 1994).

These applications of efficient coding in the retina do not consider the impact of self-generated movement on stimulus statistics. Eye movements are one example. Human eye movements are characterized by small fixational movements and occasional discrete and rapid saccades (Figure 2a,c). The spatial frequency spectrum of natural images, subject to fixational eye movements, is roughly flat (i.e., whitened) at low spatial frequencies (Kuang et al., 2012) (Figure 2b). Natural inputs that simulate fixational eye movements indeed appear to decorrelate responses in populations of salamander retinal ganglion cells (Segal et al., 2015). This whitening effect does not hold for large and rapid eye movements like saccades (Boi et al., 2017) (see Figure 2b). Thus, Rucci & colleagues (especially (Boi et al., 2017)) suggest that a single cell may use different decorrelation strategies throughout the course of natural stimulation: classical surround-mediated decorrelation or decorrelation via nonlinearities in spike generation (Pitkow and Meister, 2012) immediately following a saccade and eye-movement generated whitening during the later parts of the fixational periods between saccades. Understanding the effects of such self-generated motion on the encoding of natural scenes will require further experiments (e.g., manipulating the statistics of synthetic eye movements in experiments on primate retina).

Efficient coding beyond second order statistics

Much of the classical work on efficient coding considers only second-order statistics and their removal by decorrelation. There is, however, much more to natural images than their spatial frequency spectra. This is evident when viewing artificial stimuli with a “natural” distribution of energy across spatial frequencies but no other statistical constraints; such images look highly unnatural (e.g Figure 3). This raises a concern that coding algorithms focusing on decorrelation may miss essential features of what early visual neurons do.

Statistical independence provides a stronger constraint on efficient coding between channels (i.e. neurons or neuron-like receptive fields) than decorrelation (for a comprehensive review, see book by (Hyvärinen et al., 2009)). Although achieving independence in general is a difficult problem, it can be simplified by considering only linear transformations followed by a point nonlinearity (i.e. a linear-nonlinear approach). Two such approaches applied to natural images (Independent Component Analysis and Sparse Coding) yield filters that qualitatively resemble the oriented and localized structure of receptive fields in primary visual cortex (Bell and Sejnowski, 1997; Olshausen and Field, 1996); for a review, see (Simoncelli and Olshausen, 2001). More recent work shows that optimizing for a form of hard sparseness in which only a limited number of neurons are active can yield a better match to the full variety of cortical receptive fields in macaque (Rehn and Sommer, 2007).

Different channels can also exhibit nonlinear statistical dependencies that cannot be fully removed by linear or linear-nonlinear approaches (see (Eichhorn et al., 2009; Golden et al., 2016; Schwartz and Simoncelli, 2001) and references therein). This has prompted work on reducing statistical dependencies via nonlinear transformations. These approaches have led to more direct comparisons between models derived from scene statistics and nonlinear neural behaviors. One focus in area V1 has been on modeling nonlinear contextual

phenomena, whereby the responses of neurons to a target stimulus are influenced by stimuli that spatially surround the target, or by stimuli that have been observed in the past. Such effects can be modeled by reducing statistical dependencies between filter responses across space or time via a nonlinear computation known as divisive normalization or by other complementary approaches (Coen-Cagli et al., 2012; Karklin and Lewicki, 2009; Lochmann et al., 2012; Rao and Ballard, 1999; Schwartz and Simoncelli, 2001; Spratling, 2010; Zhu and Rozell, 2013). The statistical dependencies between filter responses can also be exploited to build models of V1 complex cells that pool together filters, resulting in invariances to translation and other properties (Hyvärinen and Hoyer, 2000; Karklin and Lewicki, 2009) (for a review see book by (Hyvärinen et al., 2009) and references therein; see also (Berkes and Wiskott, 2005; Cadieu and Olshausen, 2012)). V2 models have been derived by stacking multiple layers of linear-nonlinear transforms to achieve statistical independence, sparseness, or other related stimulus-driven goals (Coen-Cagli and Schwartz, 2013; Hosoya and Hyvarinen, 2015; Lee et al., 2008; Shan and Cottrell, 2013). One can in principle stack many layers to model higher level visual areas, but this has been rather challenging from a stimulus-based scene statistics perspective.

Generative models that capture image statistics can complement efficient coding approaches (Dayan et al., 2003; Hinton and Ghahramani, 1997). Efficient coding approaches seek to transform and manipulate inputs so as to maximize the transfer of information, which can result in statistical independence of the transformed inputs. But learning to generate the statistical dependencies prevalent in natural scenes also shows how to reduce them. To make this more concrete, consider an example in which efficient coding and generative models are complementary. Multiplicative generative models for the nonlinear dependencies in filter responses to images lead immediately to approaches to reduce such dependencies via division (Wainwright and Simoncelli, 2000). Furthermore, these approaches lead to richer models of divisive normalization, predicting that normalization in V1 neurons (e.g., to reduce statistical dependencies) is absent unless center and surround are deemed statistically homogeneous or dependent for a given image (Coen-Cagli et al., 2012) (see also (Li, 1999)). Cagli et al. (Coen-Cagli et al., 2015) used predictions about statistical dependencies in images to fit V1 neural data for large natural image patches that extend beyond the classical receptive field. This resulted in better generalization than some descriptive models of divisive normalization, demonstrating the value of incorporating understanding about statistical dependencies in images.

Goal-oriented approaches to natural vision

Efficient coding predicts that neural processing will maximize the information transmitted about a stimulus without explicitly considering behavioral demands such as the specific tasks required for survival. These behavioral considerations are central to goal-oriented approaches, which view the importance of stimulus structure and circuit mechanisms on coding through the lens of specific behavioral demands. Because many behaviorally-relevant tasks require rich stimuli, goal-oriented approaches are often used to investigate the coding of natural inputs. We first illustrate these issues from studies of the retina and insect behavior, and then turn to their application in cortex.

Retinal ganglion cells support specific behavioral goals

A common observation that supports goal-oriented approaches is high neural selectivity to specific stimulus features to the exclusion of other (equally probable) stimulus features. In an early study of retinal feature selectivity, Lettvin and colleagues interpreted retinal ganglion cell (RGC) types in explicitly ethological terms, famously going so far as to speculate that one class of ganglion cell in the frog retina may be a “bug perceiver” (Lettvin et al., 1959). But the idea that the earliest neurons in the visual system are tuned to highly specific features of the visual world was ahead of its time. Instead, the dominant view of retinal processing for several decades thereafter focused on basic processing, including lateral inhibition (via a center-surround spatial receptive field) and simple forms of luminance adaptation (Masland and Martin, 2007). In this view, the computational heavy lifting to support specific behavioral goals is done in visual areas downstream of the retina and LGN.

A great deal of evidence has now accumulated that retinal computation is more complex (for a review see (Gollisch and Meister, 2010)). A wide variety of “non-standard” RGC computations have been discovered and often explained at the circuit and synaptic level. These include: direction-selectivity, orientation selectivity (Nath and Schwartz, 2016), an omitted stimulus response (Schwartz et al., 2007), and image recurrence sensitivity (Krishnamoorthy et al., 2017). Of specific relevance here, recent work emphasizes intricate specializations of direction-selective circuits for extracting information about the direction of motion, often to the detriment of encoding other visual features (Franke et al., 2016; Zylberberg et al., 2016).

The degree to which retinal neurons are specialized to guide a particular behavior or to perform general-purpose computations predicted by efficient coding may depend on species and on location within the retina. The “complex” computations discussed above (like direction selectivity) have not been observed in primate retina, although many primate RGC types remain unexplored. Further, the fovea and peripheral retina differ dramatically in circuitry (reviewed by (Rodieck, 1998)) and in functional properties (Hecht and Verrijp, 1933; Sinha et al., 2017; Solomon et al., 2002); these differences could indicate a difference in the division of computational labor between retinal and cortical circuits across retinal eccentricity.

Differences like these - across cell types, species, or retinal eccentricity - suggest one way to reconcile stimulus- and goal-oriented frameworks in the retina. Retinal neurons that support a variety of behavioral goals or project to image-forming downstream thalamocortical circuits may show more general purpose computational features in line with efficient coding - these cells act as a common front-end for many downstream feature extractions. To transmit enough information to support a variety of downstream feature extractions, computation in these neurons resembles predictions from efficient coding. Other retinal neurons may violate predictions from efficient coding because they project to areas of the brain that underlie more specialized visually-guided behaviors -- for example, direction selective neurons (Oyster and Barlow, 1967) that project to superior colliculus or the accessory optic system to guide eye movements, or RGCs that control circadian rhythms (for review see (Hughes et al., 2016)).

Lessons from insect vision: behavioral goals shape and constrain visual processing

Goal-oriented approaches have yielded particularly satisfying explanations for complex visual processing in insects. The fly vision community has a long history of examining visual processing as it relates to behaviors like flying (Hausen and Egelhaaf, 1989). Motion processing pathways in several different insects appear tuned to each species' particular flight behaviors (O'Carroll et al., 1996). Some visual neurons in the fly encode visual features directly relevant for flight control, such as optic flow elicited by rotations or translations around and along specific body axes (Krapp and Hengstenberg, 1996; Longden et al., 2017) (see Figure 4). These neurons act as "matched filters" for specific types of optic flow (Franz and Krapp, 2000; Kohn et al., 2018). Optic flow encoding may seem obvious in hindsight, but the local motion receptive fields of these cells would appear quite mysterious if not for the careful consideration of the impact of the fly's own motion on visual inputs.

Recent work on mouse directionally-selective RGCs has similarly recast their function in terms of self-generated motion while navigating the environment (Sabbah et al., 2017) (see Figure 4). A long-standing view of directionally-selective RGCs held that they consist of four subtypes, each preferring a cardinal axis of motion (up, down, left, right, each separated by ~90 degrees) and in alignment with the axes of eye movements produced by the four rectus muscles of the eye (Oyster and Barlow, 1967). These RGCs project to the superior colliculus (Gauvain and Murphy, 2015), which further suggests that they are involved in controlling eye movements. While this distribution of preferred directions holds in the mouse central retina, in other regions of the retina the preferred axes of directionally-selective RGCs are not perpendicular and thus do not neatly align with the rectus muscles of the eye. Sabbah & colleagues mapped retinotopic differences in direction-selectivity in relation to extrapersonal visual space and motion by the animal (Figure 4). They found that directionally-selective cells are in fact better thought of as encoding the animal's own "advance/retreat" and "rise/fall" movements than the movement of some external object.

Goal-directed Approaches in Cortex

Goal-directed approaches have also been applied to visual cortex. Geisler and colleagues have promoted the importance of understanding how particular tasks may exploit different properties of natural scenes (Burge and Jaini, 2017; Geisler et al., 2009). They have focused on the representations learned by tasks such as patch identification, foreground identification, retinal speed estimation and binocular disparity. For instance, filters learned for a foreground identification task were oriented either parallel or perpendicular to surface boundaries (Geisler et al., 2009), while filters from an image patch identification task had less discrete orientation preferences and more closely resembled V1 filters. Thus, the representations learned can depend on the visual processing goals imposed on the system.

Deep Neural Networks

Recent years have seen tremendous advances in an area of machine learning known as deep neural networks (DNNs; (Krizhevsky et al., 2012; Lecun et al., 2015)); these advances have driven progress in computer vision and a host of other fields. In deep neural networks, stimuli such as natural images are represented and processed hierarchically,

loosely matched to the hierarchical structure of the brain. These networks come in many different flavors, including those that are trained in an unsupervised manner -- i.e. training in which the network learns to identify and encode statistical structure in the inputs without a specific goal. Here we focus on supervised discriminative networks, which learn to perform specific tasks using labeled training data sets. DNNs have many potential applications; we emphasize their potential to help understand and make predictions about the neural processing of natural images, particularly how the nervous system could achieve invariant object recognition (e.g. to pose, background clutter, and other within class variations).

Architecture and neural circuitry

Deep neural networks consist of a series of connected layers, each of which implements a set of basic computations (Figure 5). The computations in a single layer include linear filtering (convolution), rectification, pooling, and sometimes local response normalization. DNNs can be considered as a hierarchical extension of the linear-nonlinear models often used to empirically describe visual responses. The dimensionality (number of elements) is reduced between successive layers; as a result, effective receptive fields become larger as one progresses along the hierarchy. Thus, individual layers implement computations like those found in descriptive models of neural circuits, and the hierarchical arrangement of layers resembles the organization of visual (and other sensory) pathways.

The parameters governing DNN behavior are not determined by specific low-level computational principles (e.g. reducing statistical dependencies as in efficient coding), but instead emerge by learning to minimize the difference between the DNN output and a desired response corresponding to the DNN goal. For example, DNNs are often trained to categorize a large collection of images into discrete classes based on objects they contain (boats, cars, faces, chairs, etc.). DNNs can also be used in a descriptive (and therefore not goal-oriented) manner by fitting them directly to neural data, rather than training them on a high-level task. One such model, when fit to retinal ganglion cell responses to natural movies, reproduced several of the “complex” retinal computations discussed above. The model did not reproduce these behaviors when fit to white noise stimulation (Maheswaranathan et al., 2018). While neural networks have been around for decades, recent years have seen dramatic improvements in performance due to increases in computer speed and the availability of large data sets (e.g. images with labeled objects) that together make it possible to efficiently train networks with many layers.

Learning from successes and failures of DNNs

DNNs trained on object classification show an intriguing ability to predict the responses of cortical neurons to natural images (for recent reviews, see (Kriegeskorte, 2015; Yamins and DiCarlo, 2016); for other recent work, see (Cadena et al., 2017; Cichy et al., 2016; Pospisil et al., 2016)). This approach has been applied with particular success to processing in the ventral visual pathway, which culminates in neurons in inferotemporal (IT) cortex. Many IT neurons exhibit high feature selectivity -- responding to specific objects and (famously) faces (Young and Yamane, 1992).

The flow of signals from the retina to IT is characterized by the loss of a veridical representation of the retinal image: receptive fields become progressively larger and more complex, invariances to properties like object size and position emerge, and the appropriate space to specify inputs (e.g. inputs that produce similar responses of a given neuron) becomes increasingly difficult to identify. These transformations are challenging to learn using stimulus-based models. DNNs, however, have been more successful. Interrogation of the architecture of DNNs trained on object classification suggests that invariances may arise from the pooling stages of the networks (Fukushima, 1980; Riesenhuber and Poggio, 1999). DNNs show an ability to generalize in two important ways: (1) they are able to classify images of objects not in the original training set, including adjusting their representation of inputs for different tasks through transfer learning (Razavian et al., 2014); and, (2) they capture several aspects of neural responses even though neural data is not used in training. But DNNs are, of course, imperfect. For example, current DNN models fail to capture some aspects of human perception such as insensitivity to perturbations to an image (Szegedy et al., 2013; Ullman et al., 2016). One suggestion is that current DNN architectures operating in rather linear regimes lead to this behavior (Goodfellow et al., 2014), and that more biologically realistic saturating nonlinearities may improve performance (Nayebi and Ganguli, 2017) (although see (Brendel and Bethge, 2017)). DNNs capture some but not all aspects of responses of neurons in mid-cortical layers (Pospisil et al., 2016). In addition, interpreting DNNs can be difficult. Unlike more principled efficient coding approaches in which the form of the computation itself (e.g., divisive normalization or gain control) can be motivated by the computational goal, it is often not clear what feature of the DNN leads to a given level of performance.

Any insights that DNNs trained on high-level tasks like classification provide about how the visual system computes comes from identifying, through learning, key statistical structure in the inputs that is important for performing the specific task used in training. Motivation for such an approach comes from convergent evolution of computations like motion detection in insect and vertebrate visual systems (see above). Given that DNNs are only loosely modeled after visual circuits, a realistic expectation is that they identify the computational capabilities and limitations of specific architectures rather than provide a literal model of how the visual system works. If statistical structure of the inputs, rather than specific hardware constraints, dominates which computational strategies are effective for a given task, we might expect DNNs and neural systems to converge on similar computational algorithms even if the implementations of these algorithms differ due to differences in hardware.

Future directions

Understanding neural computation and coding in the context of naturalistic visual stimuli is a difficult problem. But the wealth of neurophysiological data about the visual system and the emergence of new computational tools for building and fitting models put us in a good position to make progress. Below we highlight a few emerging directions that we believe will help advance understanding. Many of these approaches merge techniques and ideas from the stimulus- and goal-oriented frameworks discussed above.

Identify key circuit mechanisms and integrate into models

A complete understanding of natural visual encoding entails building models that can accurately predict neural responses to natural scenes. We believe that a major reason for the shortcomings of current models is that they lack key architectural and computational features present in biological circuits, and that these features substantially shape neural responses. Certain model abstractions (for example, linearity of the receptive field) may be appropriate under some stimulus conditions but not others. At the same time, simply building models using realistic components is not likely to explain complex computations such as object recognition. Merging DNN techniques with more realistic biological circuitry offers one path forward.

DNNs components and connectivity are typically chosen largely based on the computational efficiency of learning using current optimization tools (e.g. gradient descent). This can lead to architectures that lack key components of neural circuits. Identifying and incorporating biologically-inspired computational motifs will help identify which motifs are important for specific computations -- e.g. computations at different stages of the visual hierarchy. This in turn could lead to direct predictions about the mechanisms operating in the relevant neural circuits.

One indication of the potential benefits of such an approach comes from comparing physiologically-based models of early visual areas (linear-nonlinear models with two forms of local normalization) and layers of the VGG network (which lack normalization): physiological models captured human sensitivity to image perturbations considerably better than DNNs (Berardino et al., 2017). A challenge is our current inability to identify which biological mechanisms are essential for specific computations and which can be abstracted as in linear-nonlinear models. Progress will also require probing the interactions between coactive mechanisms that are likely engaged strongly for complex stimuli such as natural images. A partial list of computational features prominent in neural circuits but under-represented in DNNs includes more sophisticated forms of normalization by stimulus context (for recent work in this direction, see (Giraldo and Schwartz, 2017; Han and Vasconcelos, 2014; Ren et al., 2017)), recurrent connections and adaptation (Mcintosh et al., 2016; Spoerer et al., 2017), and architectures for pooling across neurons (see recent work with descriptive models: (Eickenberg et al., 2012; Pagan et al., 2016; Rowekamp and Sharpee, 2017; Sharpee et al., 2013; Vintch et al., 2015)).

Combine the merits of stimulus- and goal-oriented approaches

DNNs are designed to perform well on the discriminative recognition task at the top level of the network, but this constraint does not uniquely specify the architecture of the other layers. On the other hand, stimulus-oriented approaches provide a principled approach to capture more detailed computations and nonlinearities in early stages of visual processing, including retina and primary visual cortex. But it is not clear if such approaches could capture computations in later stages of the cortical hierarchy.

An important future task is therefore finding better ways to reconcile and integrate the merits of both approaches. For instance, most of the early stages of processing that take place

before primary visual cortex are neglected in current DNNs (an exception is (Mcintosh et al., 2016)). Incorporating these early stages into networks could become a merger point between goal-directed objectives shaping the top levels of the network and stimulus-driven constraints shaping the initial stages of the architecture. Another direction is to incorporate computational motifs derived from stimulus-driven normative approaches (such as the normalization discussed above) into DNNs.

New theoretical and practical approaches that balance stimulus- and goal-oriented approaches provide promising directions. For instance, an approach known as the information bottleneck formalizes the idea of capturing relevant information rather than all information (for recent application to deep learning, see (Shwartz-Ziv and Tishby, 2017)). Another recent approach unifies several definitions of efficient coding and considers the impact of incorporating only stimuli that are predictive about the future on coding (Chalk et al., 2018). Other recent work connects generative (stimulus-oriented) and discriminative (goal-oriented) components in a single model through a shared representation (Kuleshov and Ermon, 2017). This combination has been exploited in ‘semi-supervised’ machine learning, which makes use of scarce labeled data along with unlabeled data, and therefore is a hybrid between supervised and unsupervised approaches. However, this combined stimulus and goal-oriented representation has not been applied to neuroscience and understanding natural vision. Recent theoretical work has also expanded the notion of efficient coding by recasting it as a specific case of Bayesian inference (Park and Pillow, 2017). By using a broader definition of optimality, Bayesian efficient coding allows one to evaluate the efficiency of neural representations in terms of encoding goals beyond simple information maximization.

There is also a need for progress with stimulus-oriented unsupervised learning approaches that exploit the power of DNNs without specialization for a specific goal. Unsupervised learning is considered by many the “holy grail” of learning (for recent examples, see (Ballé et al., 2016) which incorporates multiple levels of divisive normalization; and (Hirayama et al., 2017) which incorporates pooling). It is still unclear whether deep network architectures with unsupervised learning can predict responses of neurons to natural scenes or capture the invariances that characterize higher visual processing.

Train DNNs using multiple, behaviorally-inspired tasks

A DNN trained to perform a particular task can recapitulate some aspects of sensory circuits; for example, the middle layers of an image classification DNN resemble in some respects neurons in intermediate stages of the ventral stream (reviewed by (Yamins and DiCarlo, 2016); (Pospisil et al., 2016)). Presumably these correspondences arise from similarities in both network architecture and task. A real sensory system, however, supports a wide array of tasks or behavioral goals simultaneously. The result is that, especially in early sensory areas, neurons have to process sensory input in a way that supports multiple parallel feature extractions or behavioral goals. Neurons that make up this common biological front end (e.g. photoreceptors or some types of retinal ganglion cells) may therefore align their encoding strategies with efficient coding to support a wide variety of downstream goals. Downstream circuits performing more specialized computations, on the other hand, may not behave according to classical efficient coding principles. This agrees

with our intuition that efficient coding somehow applies more neatly to peripheral sensory systems. Formalizing this intuition requires grappling with several difficult questions: Are there general rules that govern when a stimulus- or goal-oriented perspective is more appropriate? At what point does a sensory pathway stop simply efficiently packaging information and start "doing" something with that information?

Multi-task DNNs offer one approach for exploring how shared circuitry could support multiple tasks (Scholte et al., 2017). Indeed, such networks trained for speech and music classification naturally divide into separate pathways, and the level at which that split occurs can affect the performance of the network on these two tasks (Kell et al., 2018). An interesting question is whether constraining networks by multiple mid-level tasks (as in (Chengxu Zhuang, 2018)) can provide a more general-purpose representation resembling that predicted by efficient encoding. A major impediment to developing multi-task DNNs is the limited availability of datasets that could be used to train such networks (e.g. ImageNet, which consists of a collection of labeled objects, is the dominant dataset used for vision-related applications).

Acknowledgments

We thank Holger Krapp, Dean Pospisil, and Jonathon Shlens for helpful feedback on an earlier version of this review. Holger Krapp very generously provided the data and schematic shown in Fig. 4a,b. This work was supported by NIH grants F31-EY026288 (M.H.T.), EY11850 (F.R.), and a National Science Foundation Grant 1715475 (O.S.).

References cited

Arjovsky M, Chintala S, and Bottou L (2017). Wasserstein GAN. ArXiv Prepr. arXiv:1701.07875v3.

Atick JJ (1992). Could information theory provide an ecological theory of sensory processing? Netw. Comput. Neural Syst 3, 213–251.

Atick JJ, and Redlich AN (1992). What Does the Retina Know About Natural Scenes? Neural Comput. 210, 196–210.

Attneave F (1954). Some informational apsects of visual perception. Psychol. Rev 3.

Baddeley R, Abbott LF, Booth MCA, Sengpiel F, Freeman T, Wakeman EA, and Rolls ET (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proc. R. Soc. London B 264, 1775–1783.

Ballé J, Laparra V, and Simoncelli EP (2016). End-to-end Optimized Image Compression. ArXiv Prepr. arXiv:1611.01704v3.

Barlow H (2001). Redundancy reduction revisited. Netw. Comput. Neural Syst 12, 241–253.

Barlow HB (1961). Possible principles underlying the transformations of sensory messages. Sens. Commun 6, 217–234.

Bell AJ, and Sejnowski TJ (1997). The "independent components'"of natural scenes are edge filters." Vision Res. 37, 3327–3338. [PubMed: 9425547]

Berardino A, Ballé J, Laparra V, and Simoncelli EP (2017). Eigen-Distortions of Hierarchical Representations. ArXiv Prepr. arXiv:1710.02266v3.

Berkes P, and Wiskott L (2005). Slow feature analysis yields a rich repertoire of complex cell properties. J. Vis 5, 9–9.

Bhandawat V, Olsen SR, Gouwens NW, Schlief ML, and Wilson RI (2007). Sensory processing in the Drosophila antennal lobe increases reliability and separability of ensemble odor representations. Nat. Neurosci 10, 1474–1482. [PubMed: 17922008]

Boi M, Poletti M, Victor JD, and Rucci M (2017). Consequences of the Oculomotor Cycle for the Dynamics of Perception. Curr. Biol 27, 1268–1277. [PubMed: 28434862]

Brendel W, and Bethge M (2017). Comment on “Biologically inspired protection of deep networks from adversarial attacks.” ArXiv Prepr. arXiv:1704.01547v1.

Brinkman BAW, Weber AI, Rieke F, and Shea-Brown E (2016). How Do Efficient Coding Strategies Depend on Origins of Noise in Neural Circuits? PLoS Comput. Biol 12, 1–34.

Burge J, and Jaini P (2017). Accuracy Maximization Analysis for Sensory-Perceptual Tasks: Computational Improvements, Filter Robustness, and Coding Advantages for Scaled Additive Noise.

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, and Ecker AS (2017). Deep convolutional models improve predictions of macaque V1 responses to natural images. BioRxiv Prepr. 201764.

Cadieu CF, and Olshausen BA (2012). Learning intermediate-level representations of form and motion from natural movies. Neural Comput. 24, 827–866. [PubMed: 22168556]

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, and Rust NC (2005). Do We Know What the Early Visual System Does? J. Neurosci 25, 10577–10597. [PubMed: 16291931]

Chalk M, Marre O, and Tkačik G (2018). Toward a unified theory of efficient, predictive, and sparse coding. Proc. Natl. Acad. Sci 115, 186–191. [PubMed: 29259111]

Chengxu Zhuang DY (2018). Using multiple optimization tasks to improve deep neural network models of higher ventral cortex. In COSYNE Abstracts, p.

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

Coen-Cagli R, and Schwartz O (2013). The impact on midlevel vision of statistically optimal divisive normalization in V1. J. Vis 13, 1–20.

Coen-Cagli R, Dayan P, and Schwartz O (2012). Cortical surround interactions and perceptual salience via natural scene statistics. PLoS Comput. Biol 8.

Coen-Cagli R, Kohn A, and Schwartz O (2015). Flexible gating of contextual influences in natural vision. Nat. Neurosci 18, 1648–1655. [PubMed: 26436902]

Creutzfeldt OD, and Nothdurft HC (1978). Representation of complex visual stimuli in the brain. Naturwissenschaften 65, 307–318. [PubMed: 673010]

Dan Y, Atick JJ, and Reid RC (1996). Efficient Coding of Natural Scenes in the Lateral Geniculate Nucleus: Experimental Test of a Computational Theory. J. Neurosci 16, 3351–3362. [PubMed: 8627371]

David SV, and Gallant JL (2005). Predicting neuronal responses during natural vision. Netw. Comput. Neural Syst 16, 239–260.

Dayan P, and Abbott LF (2001). Theoretical Neuroscience (Cambridge, MA: MIT Press).

Dayan P, Sahani M, and Deback G (2003). Adaptation and Unsupervised Learning. Adv. Neural Inf. Process. Syst 15 237–244.

Demb JB, and Singer JH (2015). Functional Circuitry of the Retina. Annu. Rev. Vis. Sci 1, 263–289. [PubMed: 28532365]

Eichhorn J, Sinz F, and Bethge M (2009). Natural Image Coding in V1: How Much Use Is Orientation Selectivity? PLoS Comput. Biol 5.

Eickenberg M, Rowekamp RJ, Kouh M, and Sharpee TO (2012). Characterizing responses of translation-invariant neurons to natural stimuli: Maximally informative invariant dimensions. Neural Comput. 24, 2384–2421. [PubMed: 22734487]

Field DJ (1987). Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A 4, 2379–2394. [PubMed: 3430225]

Field DJ (1994). What Is the Goal of Sensory Coding? Neural Comput. 6, 559–601.

Franke F, Fiscella M, Sevelev M, Roska B, Hierlemann A, and Azeredo da Silveira R (2016). Structures of Neural Correlation and How They Favor Coding. Neuron 89, 409–422. [PubMed: 26796692]

Franke K, Berens P, Schubert T, Bethge M, Euler T, and Baden T (2017). Balanced excitation and inhibition decorrelates visual feature representation in the mammalian inner retina. Nature 542, 439–444. [PubMed: 28178238]

Franz MO, and Krapp HG (2000). Wide-field, motion-sensitive neurons and matched filters for optic flow fields. Biol. Cybern 83, 185–197. [PubMed: 11007295]

Frazor RA, and Geisler WS (2006). Local luminance and contrast in natural images. Vision Res. 46, 1585–1598. [PubMed: 16403546]

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, and Movshon JA (2013). A functional and perceptual signature of the second visual area in primates. Nat. Neurosci 16, 974–981. [PubMed: 23685719]

Fukushima K (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern 36, 193–202. [PubMed: 7370364]

Gatys LA, Ecker AS, and Bethge M (2015). Texture Synthesis Using Convolutional Neural Networks. Adv. Neural Inf. Process. Syst 262–270.

Gauvain G, and Murphy GJ (2015). Projection-Specific Characteristics of Retinal Input to the Brain. J. Neurosci 35, 6575–6583. [PubMed: 25904807]

Geisler WS, Najemnik J, and Ing AD (2009). Optimal stimulus encoders for natural tasks. J. Vis 9, 17–17.

Giraldo LGS, and Schwartz O (2017). Flexible normalization in deep convolutional neural networks. In COSYNE Abstracts, p.

Gjorgjieva J, Sompolinsky H, and Meister M (2014). Benefits of pathway splitting in sensory coding. J. Neurosci 34, 12127–12144. [PubMed: 25186757]

Golden JR, Vilankar KP, Wu MCK, and Field DJ (2016). Conjectures regarding the nonlinear geometry of visual neurons. Vision Res. 120, 74–92. [PubMed: 26902730]

Gollisch T, and Meister M (2010). Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. Neuron 65, 150–164. [PubMed: 20152123]

Goodfellow IJ, Shlens J, and Szegedy C (2014). Explaining and Harnessing Adversarial Examples. ArXiv Prepr. arXiv:1412.6572v3.

Graham NV (2011). Beyond multiple pattern analyzers modeled as linear filters (as classical V1 simple cells): Useful additions of the last 25 years. Vision Res. 51, 1397–1430. [PubMed: 21329718]

Han S, and Vasconcelos N (2014). Object recognition with hierarchical discriminant saliency networks. Front. Comput. Neurosci 8, 1–20. [PubMed: 24550816]

Hausen K, and Egelhaaf M (1989). Neural mechanisms of visual course Control in Insects. In Facets of Vision, Stavenga DG, and Hardie RC, eds. (Springer London), pp. 391–424.

Hecht S, and Verrijp C (1933). Intermittent Stimulation By Light III. The relation between intensity and critical fusion frequency for different retinal locations. J. Gen. Physiol 251.

Heitman A, Brackbill N, Greschner M, Sher A, Litke AM, and Chichilnisky EJ (2016). Testing pseudo-linear models of responses to natural scenes in primate retina. BioRxiv.

Hinton GE, and Ghahramani Z (1997). Generative models for discovering sparse distributed representations. Philos. Trans. R. Soc. Lond. B. Biol. Sci 352, 1177–1190. [PubMed: 9304685]

Hirayama J, Hyvärinen A, and Kawanabe M (2017). SPLICE: Fully Tractable Hierarchical Extension of ICA with Pooling. Proc. 34th Int. Conf. Mach. Learn 70, 1491–1500.

Hosoya H, and Hyvarinen A (2015). A Hierarchical Statistical Model of Natural Images Explains Tuning Properties in V2. J. Neurosci 35, 10412–10428. [PubMed: 26203137]

Hughes S, Jagannath A, Rodgers J, Hankins MW, Peirson SN, and Foster RG (2016). Signalling by melanopsin (OPN4) expressing photosensitive retinal ganglion cells. Eye 30, 247–254. [PubMed: 26768919]

Hyvärinen A (2010). Statistical models of natural images and cortical visual representation. Top. Cogn. Sci 2, 251–264. [PubMed: 25163788]

Hyvärinen A, and Hoyer P (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural Comput. 12, 1705–1720. [PubMed: 10935923]

Hyvärinen A, Hurri J, and Hoyer PO (2009). Natural Image Statistics-A Probabilistic Approach to Early Computational Vision.

Karklin Y, and Lewicki MS (2005). A Hierarchical Bayesian Model for Learning Nonlinear Statistical Regularities in Nonstationary Natural Signals. Neural Comput. 17, 397–423. [PubMed: 15720773]

Karklin Y, and Lewicki MS (2009). Emergence of complex cell properties by learning to generalize in natural scenes. Nature 457, 83–86. [PubMed: 19020501]

Karras T, Aila T, Laine S, and Lehtinen J (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. ICLR 2018.

Kastner DB, Baccus S. a., and Sharpee TO (2015). Critical and maximally informative encoding between neural populations in the retina. Pnas 112, 2533–2538. [PubMed: 25675497]

Kell AJE, Yamins DLK, Shook EN, Norman-haignere SV, Mcdermott JH, Kell AJE, Yamins DLK, Shook EN, and Norman-haignere SV (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 98, 1–15. [PubMed: 29621482]

Kohn JR, Heath SL, and Behnia R (2018). Eyes Matched to the Prize : The State of Matched Filters in Insect Visual Circuits. Front. Neural Circuits 12, 26. [PubMed: 29670512]

Krapp HG, and Hengstenberg R (1996). Estimation of self-motion by optic flow processing in single visual interneurons. Nature 384, 463–466. [PubMed: 8945473]

Kriegeskorte N (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annu. Rev. Vis. Sci 1, 417–446. [PubMed: 28532370]

Krishnamoorthy V, Weick M, and Gollisch T (2017). Sensitivity to image recurrence across eye-movement-like image transitions through local serial inhibition in the retina. Elife e22431. [PubMed: 28230526]

Krizhevsky A, Sutskever I, and Hinton GE (2012). ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst 1–9.

Kuang X, Poletti M, Victor JD, and Rucci M (2012). Temporal encoding of spatial information during active visual fixation. Curr. Biol 22, 510–514. [PubMed: 22342751]

Kuleshov V, and Ermon S (2017). Deep Hybrid Models: Bridging Discriminative and Generative Approaches. Uncertain. Ai

Laughlin S (1981). A simple coding procedure enhances a neuron's information capacity. Zeitschrift Fur Naturforsch. - Sect. C J. Biosci 36, 910–912.

Lecun Y, Bengio Y, and Hinton G (2015). Deep learning. Nature 521, 436–444. [PubMed: 26017442]

Lee H, Ekanadham C, and Ng AY (2008). Sparse deep belief net model for visual area V2. Adv. Neural Inf. Process. Syst 20 873–880.

Lettvin JY, Maturana HR, McCulloch WS, and Pitts WH (1959). What the Frog's Eye Tells the Frog's brain. Proc. Natl. Acad. Sci 1940–1951.

Lewicki MS, Olshausen BA, Surlykke A, and Moss CF (2014). Scene analysis in the natural environment. Front. Psychol 5, 1–21. [PubMed: 24474945]

Li Z (1999). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. Proc. Natl. Acad. Sci 96, 10530–10535. [PubMed: 10468643]

Li Z, and Atick JJ (1994). Efficient stereo coding in the multiscale representation. Netw. Comput. Neural Syst 5, 157–174.

Van Der Linde I, Rajashekar U, Bovik AC, and Cormack LK (2009). DOVES: a database of visual eye movements. Spat. Vis 22, 161–177. [PubMed: 19228456]

Lochmann T, Ernst UA, and Deneve S (2012). Perceptual Inference Predicts Contextual Modulations of Sensory Responses. J. Neurosci 32, 4179–4195. [PubMed: 22442081]

Longden KD, Wicklein M, Hardcastle BJ, Huston SJ, and Krapp HG (2017). Spike Burst Coding of Translatory Optic Flow and Depth from Motion in the Fly Visual System. Curr. Biol 27, 3225–3236.e3. [PubMed: 29056452]

Maheswaranathan N, Baccus SA, and Ganguli S (2017). Inferring hidden structure in multilayered neural circuits. BioRxiv.

Maheswaranathan N, McIntosh LT, Kastner DB, Melander J, Brezovec L, Nayebi A, Wang J, Ganguli S, and Baccus SA (2018). Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. BioRxiv.

Marr D, and Hildreth E (1980). Theory of edge detection. Proc. R. Soc. London B 207, 187–217. [PubMed: 6102765]

Masland RH, and Martin PR (2007). The unsolved mystery of vision. Curr. Biol 17, 577–582.

Mcintosh LT, Maheswaranathan N, Nayebi A, Ganguli S, and Baccus SA (2016). Deep Learning Models of the Retinal Response to Natural Scenes. Adv. Neural Inf. Process. Syst 30, 1–9.

Nath A, and Schwartz GW (2016). Cardinal Orientation Selectivity Is Represented by Two Distinct Ganglion Cell Types in Mouse Retina. J. Neurosci 36, 3208–3221. [PubMed: 26985031]

Nayebi A, and Ganguli S (2017). Biologically inspired protection of deep networks from adversarial attacks. ArXiv Prepr. arXiv:1703.09202v1.

O'Carroll DC, Bidwell NJ, Laughlin SB, and Warrant EJ (1996). Insect motion detectors matched to visual ecology. Nature 382, 63–66. [PubMed: 21638927]

Okazawa G, Tajima S, and Komatsu H (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc. Natl. Acad. Sci 112, E351–E360. [PubMed: 25535362]

Olshausen BA, and Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609. [PubMed: 8637596]

Oyster CW, and Barlow HB (1967). Direction-selective units in rabbit retina: distribution of preferred directions. Science (80-.). 155, 841–842.

Pagan M, Simoncelli EP, and Rust NC (2016). Neural Quadratic Discriminant Analysis: Nonlinear Decoding with V1-Like Computation. Neural Comput. 28, 2291–2319. [PubMed: 27626960]

Park IM, and Pillow JW (2017). Bayesian Efficient Coding. BioRxiv 178418.

Parra L, Spence C, and Sajda P (2001). Higher-order statistical properties arising from the non-stationarity of natural signals. Adv. Neural Inf. Process. Syst 786–792.

Pitkow X, and Meister M (2012). Decorrelation and efficient coding by retinal ganglion cells. Nat. Neurosci 15, 628–635. [PubMed: 22406548]

Portilla J, and Simoncelli EP (2000). Parametric texture model based on joint statistics of complex wavelet coefficients. Int. J. Comput. Vis 40, 49–71.

Pospisil D, Pasupathy A, and Bair W (2016). Comparing the brain's representation of shape to that of a deep convolutional neural network. Proc. 9th EAI Int. Conf. Bio-Inspired Inf. Commun. Technol. (Formerly BIONETICS) 516–523.

Rao RPN, and Ballard DH (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci 2, 79–87. [PubMed: 10195184]

Razavian AS, Azizpour H, Sullivan J, and Carlsson S (2014). CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. CVPR2014 Work. 806–813.

Rehn M, and Sommer FT (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. J. Comput. Neurosci 22, 135–146. [PubMed: 17053994]

Ren M, Liao R, Urtasun R, Sinz FH, and Zemel RS (2017). Normalizing the Normalizers: Comparing and Extending Network Normalization Schemes. ICLR 2017.

Rieke F, and Rudd ME (2009). The Challenges Natural Images Pose for Visual Adaptation. Neuron 64, 605–616. [PubMed: 20005818]

Riesenhuber M, and Poggio T (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci 2, 1019–1025. [PubMed: 10526343]

Rodieck RW (1998). The First Steps in Seeing.

Rowekamp RJ, and Sharpee TO (2017). Cross-orientation suppression in visual area V2. Nat. Commun 8, 1–9. [PubMed: 28232747]

Rucci M, and Victor JD (2015). The unsteady eye: An information-processing stage, not a bug. Trends Neurosci. 38, 195–206. [PubMed: 25698649]

Ruderman DL (1994). The statistics of natural images. Netw. Comput. Neural Syst 5, 517–548.

Ruderman DL, and Bialek W (1994). Statistics of natural images: Scaling in the woods. Adv. Neural Inf. Process. Syst 73, 551–558.

Rust NC, and DiCarlo JJ (2010). Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. J. Neurosci 30, 12978–12995. [PubMed: 20881116]

Sabbah S, Gemmer JA, Bhatia-Lin A, Manoff G, Castro G, Siegel JK, Jeffery N, and Berson DM (2017). A retinal code for motion along the gravitational and body axes. Nature.

Scholte HS, Losch MM, Ramakrishnan K, de Haan EHF, and Bohte SM (2017). Visual pathways from the perspective of cost functions and deep learning. BioRxiv 146472.

Schwartz GW, and Rieke F (2011). Perspectives on: Information and coding in mammalian sensory physiology: Nonlinear spatial encoding by retinal ganglion cells: when 1 + 1 != 2. J. Gen. Physiol 138, 283–290. [PubMed: 21875977]

Schwartz O, and Simoncelli EP (2001). Natural signal statistics and sensory gain control. Nat. Neurosci 4, 819–825. [PubMed: 11477428]

Schwartz G, Harris R, Shrom D, and Berry MJ (2007). Detection and prediction of periodic patterns by the retina. Nat. Neurosci 10, 552–554. [PubMed: 17450138]

Segal IY, Giladi C, Gedalin M, Rucci M, Ben-Tov M, Kushinsky Y, Mokeichev A, and Segev R (2015). Decorrelation of retinal response to natural scenes by fixational eye movements. Proc. Natl. Acad. Sci 112, 3110–3115. [PubMed: 25713370]

Shan H, and Cottrell G (2013). Efficient Visual Coding: From Retina To V2. ArXiv Prepr.

Shannon CE (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J 27, 379–423.

Sharpee TO, Sugihara H, Kurgansky AV, Rebrik SP, Stryker MP, and Miller KD (2006). Adaptive filtering enhances information transmission in visual cortex. Nature 439, 936–942. [PubMed: 16495990]

Sharpee TO, Kouh M, and Reynolds JH (2013). Trade-off between curvature tuning and position invariance in visual area V4. Proc. Natl. Acad. Sci 110, 11618–11623. [PubMed: 23798444]

Shwartz-Ziv R, and Tishby N (2017). Opening the Black Box of Deep Neural Networks via Information. ArXiv Prepr. arXiv:1703.00810v3.

Simoncelli EP, and Olshausen BA (2001). Natural image statistics and neural representation. Annu. Rev. Neurosci 24, 1193–1216. [PubMed: 11520932]

Sinha R, Hoon M, Baudin J, Okawa H, Wong ROL, and Rieke F (2017). Cellular and Circuit Mechanisms Shaping the Perceptual Properties of the Primate Fovea. Cell 168, 413–426.e12. [PubMed: 28129540]

Smyth D, Willmore B, Baker GE, Thompson ID, and Tolhurst DJ (2003). The Receptive-Field Organization of Simple Cells in Primary Visual Cortex of Ferrets under Natural Scene Stimulation. J. Neurosci 23, 4746–4759. [PubMed: 12805314]

Solomon SG, and Kohn A (2014). Moving sensory adaptation beyond suppressive effects in single neurons. Curr. Biol 24, R1012–R1022. [PubMed: 25442850]

Solomon SG, Martin PR, White AJR, Rüttiger L, and Lee BB (2002). Modulation sensitivity of ganglion cells in peripheral retina of macaque. Vision Res. 42, 2893–2898. [PubMed: 12450500]

Spoerer CJ, McClure P, and Kriegeskorte N (2017). Recurrent convolutional neural networks: A better model of biological object recognition. Front. Psychol 8, 1–14. [PubMed: 28197108]

Spratling MW (2010). Predictive Coding as a Model of Response Properties in Cortical Area V1. J. Neurosci 30, 3531–3543. [PubMed: 20203213]

Srinivasan MV, Laughlin SB, and Dubs A (1982). Predictive coding: a fresh view of inhibition in the retina. Proc. R. Soc. London B 216, 427–459. [PubMed: 6129637]

Stanley GB, Li FF, and Dan Y (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. J. Neurosci 19, 8036–8042. [PubMed: 10479703]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, and Fergus R (2013). Intriguing properties of neural networks. ArXiv Prepr. arXiv:1312.6199v4.

Theunissen FE, and Elie JE (2014). Neural processing of natural sounds. Nat. Rev. Neurosci 15, 355–366. [PubMed: 24840800]

Thomsom MGA (1999). Visual coding and the phase structure of natural scenes. Netw. Comput. Neural Syst 10, 123–132.

Turner MH, and Rieke F (2016). Synaptic Rectification Controls Nonlinear Spatial Integration of Natural Visual Inputs. Neuron 90, 1257–1271. [PubMed: 27263968]

Turner MH, Schwartz GW, and Rieke F (2018). Receptive field center-surround interactions mediate context-dependent spatial contrast encoding in the retina. BioRxiv 252148.

Ullman S, Dorfman N, and Harari D (2016). Discovering 'containment': from infants to machines. ArXiv Prepr. arXiv:1610.09625v1.

Vickers NJ, Christensen TA, Baker TC, and Hildebrand JG (2001). Odour-plume dynamics influence file brain's olfactory code. Nature 410, 466–470. [PubMed: 11260713]

Vincent BT, and Baddeley RJ (2003). Synaptic energy efficiency in retinal processing. Vision Res.

Vinje WE, and Gallant JL (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. Science (80-.). 287, 1273–1276.

Vintch B, Movshon JA, and Simoncelli EP (2015). A Convolutional Subunit Model for Neuronal Responses in Macaque V1. J. Neurosci 35, 14829–14841. [PubMed: 26538653]

Wainwright MJ, and Simoncelli EP (2000). Scale mixtures of Gaussians and the statistics of natural images. Adv. Neural Inf. Process. Syst 12, 855–861.

Yamins DLK, and DiCarlo JJ (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci 19, 356–365. [PubMed: 26906502]

Young M, and Yamane S (1992). Sparse population coding of faces in the inferotemporal cortex. Science (80-.). 256, 1327–1331.

Zhaoping L (2006). Theoretical understanding of the early visual processes by data compression and data selection.

Zhaoping L (2014). Understanding Vision: theory, models, and data.

Zhu M, and Rozell CJ (2013). Visual Nonclassical Receptive Field Effects Emerge from Sparse Coding in a Dynamical System. PLoS Comput. Biol 9, 1–15.

Zwicker D, Murugan A, and Brenner MP (2016). Receptor arrays optimized for natural odor statistics. Proc. Natl. Acad. Sci

Zylberberg J, Cafaro J, Turner MH, Shea-Brown E, and Rieke F (2016). Direction-Selective Circuits Shape Noise to Ensure a Precise Population Code. Neuron 89, 369–383. [PubMed: 26796691]

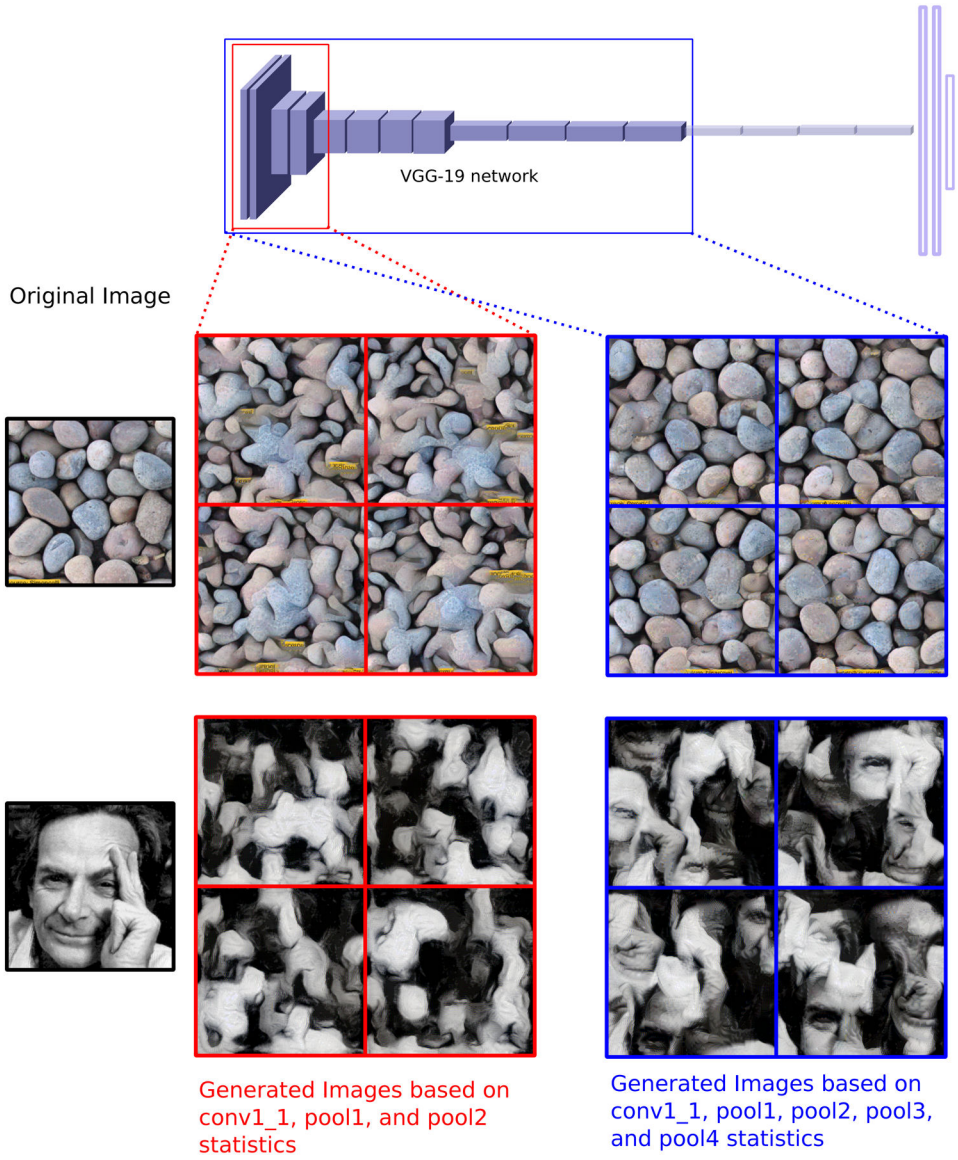

Figure 1: Texture Synthesis based on Deep Convolutional Neural Networks.

The activations of different layers of a DNN trained for object recognition can be employed to capture statistics of textures beyond second order (Gatys et al., 2015). Texture synthesis is accomplished by numerical optimization of the pixel values of an image that matches the statistics of a reference image (Original Image enclosed in black). Statistics can be obtained from activation values at different stages of the deep DNN. Images enclosed in red are synthesized by considering only activations from the first and second pooling stages of the DNN, whereas images enclosed in blue include the third and fourth pooling stages in their statistics. In the case of the inhomogeneous images (bottom row) the texture generation tiles local features in scrambled places that will match the activation statistics that have been averaged over space.

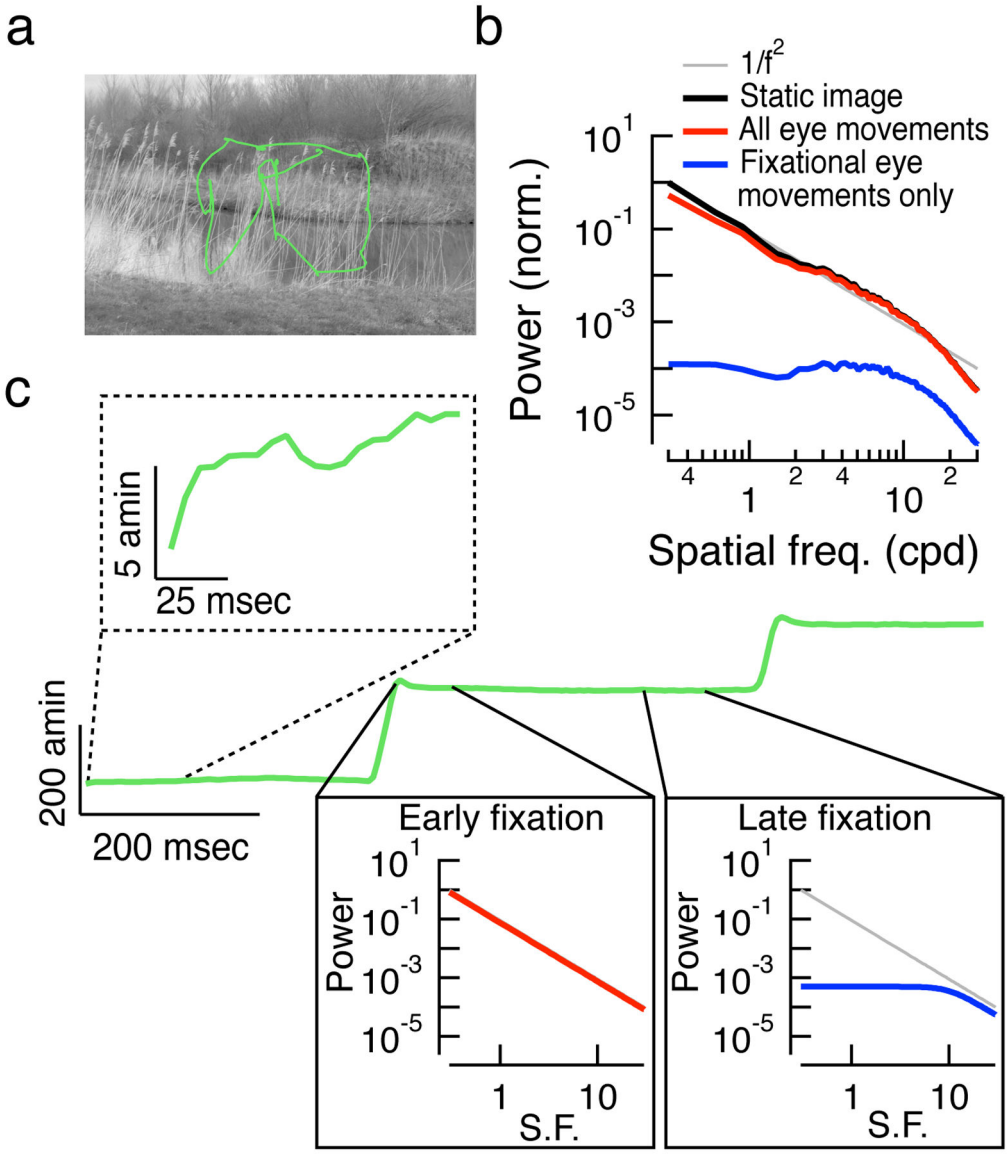

Figure 2: Efficient coding strategies rely on self-generated movement.

(a) A natural image and measured human eye movement trajectory from (Van Der Linde et al., 2009). An observer will explore a scene using large, ballistic changes in fixation called saccades. In the time between saccades, observers make much smaller, involuntary eye movements called fixational eye movements (for review, see (Rucci and Victor, 2015)). (b) Using these eye movement data, we can reconstruct the time-varying image on the retina into a naturalistic movie stimulus. We summed the Fourier spatial power spectra of each frame of this movie, resulting in a roughly $1/f^2$ power law scaling, which is characteristic of static natural images (black trace). Following the analysis in (Kuang et al., 2012), we then measured spatial power spectra for the dynamic component of the natural movie. To produce these spatial power spectra, we computed the spatiotemporal power spectrum of a movie and summed over all *non-zero temporal frequencies*. Fixational eye movements simply shift much of the power, except that at the lowest spatial frequencies, to higher

temporal frequencies. The removal of the temporal DC component of the movie thus selectively removes low spatial frequency content, and the result is a whitened spatial power spectrum (Fig. 2b, blue trace). Importantly, this result relies on fixational eye movements and *not* saccades. When saccades are included in the natural movie stimulus, considerable low spatial frequency content is still present at nonzero temporal frequencies, so whitening does not occur (Fig. 2b, red trace). (c) The position (in one dimension) of the eye as a function of time is shown by the green trace. Examining the eye position at a finer time scale (dashed inset) reveals smaller fixational eye movements. Boi et al. (Boi et al., 2017) suggested that during a saccade, the dynamic spatial frequency content of natural images follows the familiar $1/f^2$ power law scaling (left inset, red trace). As the fixation proceeds, the retinal input is whitened (right inset, blue trace). Between saccades (when the image is relatively stable), any low spatial frequency content is present mostly in the temporal DC component of the input. In other words, the large-scale spatial structure isn't changing very much within a single fixation. The whitening effect of fixational eye movements will depend on how completely (and how quickly) a visual neuron adapts to the (mostly static) low spatial frequency content imposed by each new fixation.

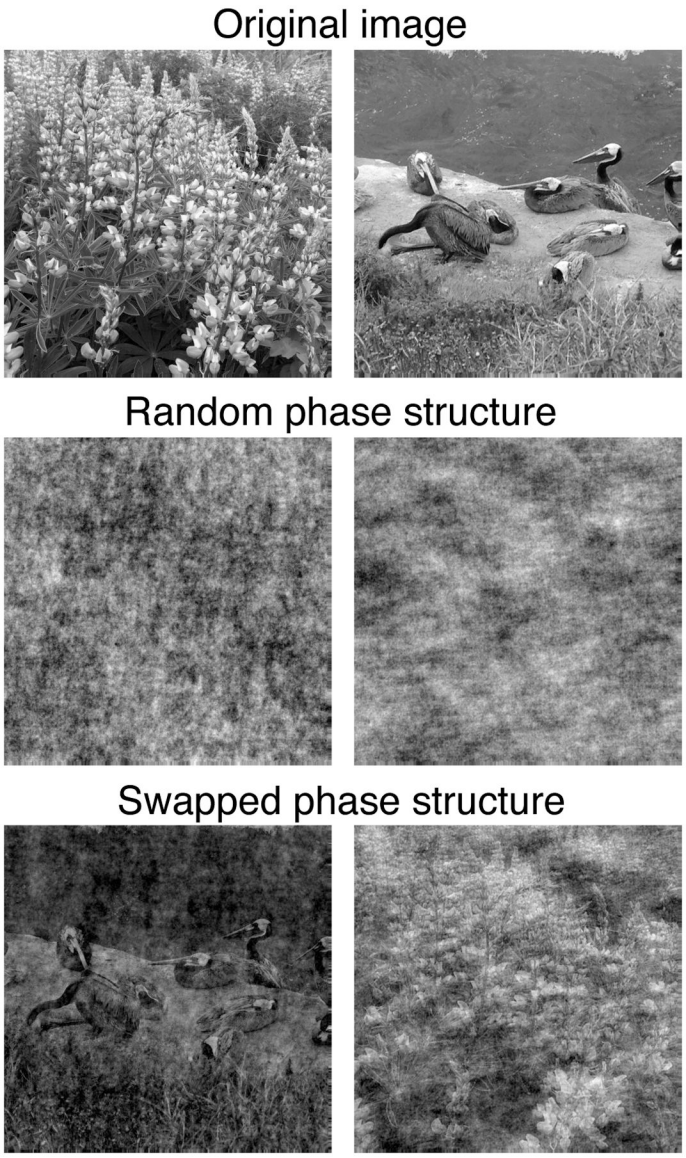

Figure 3: Beyond-pairwise statistics contribute to complex structure in natural images.
Top row: Two grayscale natural images. Middle row: The natural images above with randomized phase spectra. Both of these images have the roughly $1/f^2$ spatial power spectrum characteristic of natural images, yet appear quite unnatural. Bottom row: The natural images with their phase spectra swapped, such that the image on the left now has the phase spectrum of the original image on the right, and vice-versa. See (Simoncelli and Olshausen, 2001; Thomsom, 1999).

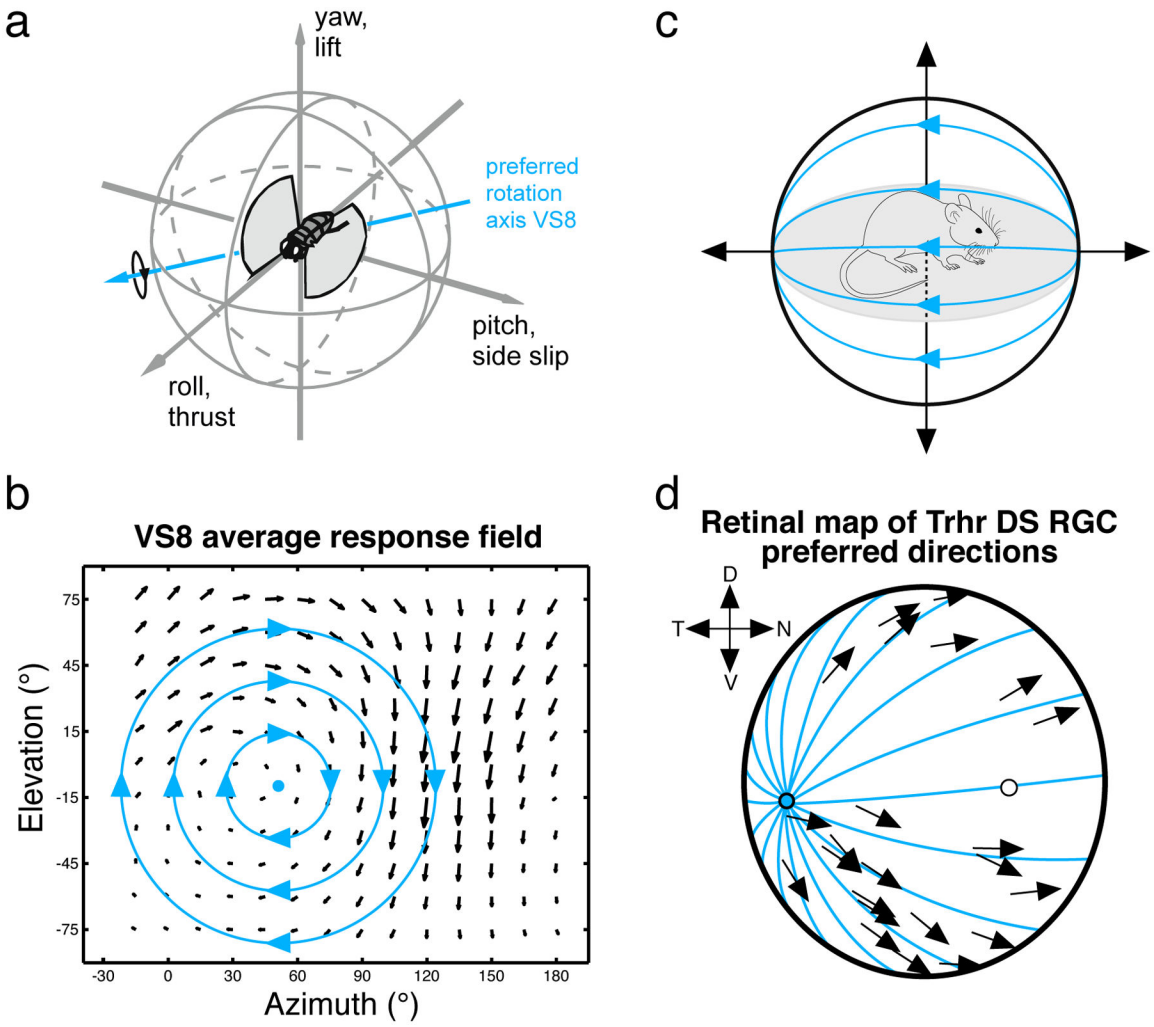

Figure 4: Motion sensitive neurons encode self-movement across the animal kingdom.

(a) Schematic showing a fly in flight. (b) Local motion receptive field of the VS8 neuron in the blowfly Calliphora. The direction of each arrow indicates the local preferred direction, and the length of each arrow indicates the cell's motion sensitivity. This local motion receptive field corresponds to the optic flow pattern that would result from a rotation of the animal. The rotation axis around which the fly would need to turn to maximally activate this neuron is indicated in (a). Data & schematic provided by Holger Krapp. (c) Schematic showing a mouse ambulating in a forward direction. The resulting visual input is an optic flow pattern emanating from a singularity directly ahead of the animal (blue lines). (d) Direction preferences of a population of DS RGCs in mouse retina are overlaid on the retinal surface. Forward motion optic flow moves outward from a point in the retina (blue lines). The direction preferences of this cell type roughly align with the optic flow lines that result from forward motion. Other DS RGC types similarly respond to optic flow resulting from other directions of motion of the animal. Data redrawn from (Sabbah et al., 2017).

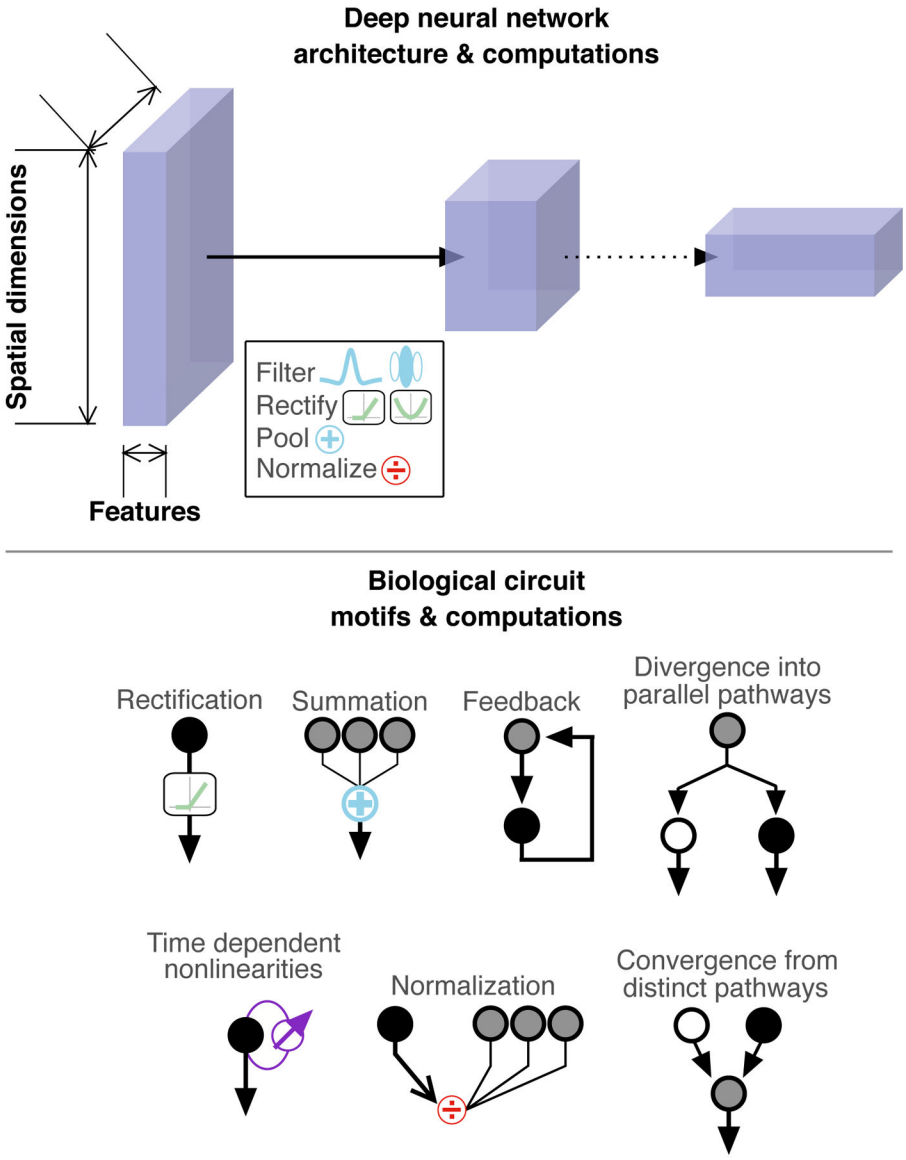

Figure 5: Deep neural networks reflect some, but not all, architectural and computational motifs found in neural circuits.

Top: Deep neural networks are composed of multiple, connected layers. Several basic computations are performed within each layer. Bottom: examples of common circuit motifs and computations observed in neural circuits. Some of these examples are well-represented by many DNNs (e.g. pooling / filtering), others can be included in DNNs but their precise nature & location are not necessarily well reflected (e.g. rectification or normalization), and still others are excluded from most DNNs (e.g. time-dependent nonlinearities).