ANNUAL REVIEWS

Annual Review of Vision Science

Invariant Recognition Shapes Neural Representations of Visual Input

Andrea Tacchetti, Leyla Isik, and Tomaso A. Poggio

Center for Brains, Minds and Machines, MIT, Cambridge, Massachusetts 02139, USA; email: atacchet@mit.edu, lisik@mit.edu, tp@mit.edu

Annu. Rev. Vis. Sci. 2018. 4:403–22

First published as a Review in Advance on July 27, 2018

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-091517-034103

Keywords

visual representations, invariance, neural decoding, computational neuroscience

Abstract

Recognizing the people, objects, and actions in the world around us is a crucial aspect of human perception that allows us to plan and act in our environment. Remarkably, our proficiency in recognizing semantic categories from visual input is unhindered by transformations that substantially alter their appearance (e.g., changes in lighting or position). The ability to generalize across these complex transformations is a hallmark of human visual intelligence, which has been the focus of wide-ranging investigation in systems and computational neuroscience. However, while the neural machinery of human visual perception has been thoroughly described, the computational principles dictating its functioning remain unknown. Here, we review recent results in brain imaging, neurophysiology, and computational neuroscience in support of the hypothesis that the ability to support the invariant recognition of semantic entities in the visual world shapes which neural representations of sensory input are computed by human visual cortex.

1. INTRODUCTION

Humans effortlessly make sense of the visual world around them. We are able to quickly recognize other people's actions in unfamiliar surroundings and pick someone out in a crowd after a single glance. Remarkably, human performance on these tasks is largely unaffected by those changes in the visual appearance of objects, faces, or action sequences (e.g., changes in illumination, viewpoint, facial expression, and even aging) that do not change their semantic category. The ability to generalize across these complex transformations is a hallmark of human visual intelligence.

The neuroscience of vision has made great strides in describing and characterizing the computations that take place in visual cortex to support human visual perception, both at the level of single units and in whole brain regions. This line of research has revealed that the primate visual system is organized as a hierarchical succession of layers, where invariance to transformations and selectivity to particular stimuli increase at each computational stage (Connor et al. 2007, DiCarlo et al. 2012). Precise descriptions of the preferred stimuli within these layers have shown that visual cortex goes from representing simple oriented lines and edges in its earliest layers (Hubel & Wiesel 1962, Gallant et al. 1993, Ringach 2002) to representing whole categories of objects, across a wide range of transformations, in its most anterior areas (Riesenhuber & Poggio 1999, Serre et al. 2007b, Rust & DiCarlo 2010, DiCarlo et al. 2012).

Despite access to precise characterizations of the neural machinery involved in visual perception (Felleman & Van Essen 1991) and the availability of powerful computational models (Serre et al. 2007a, Kriegeskorte 2015, LeCun et al. 2015), little is known about why visual cortex computes certain representations of visual inputs and not others—or, more precisely, what computational tasks might be relevant to explain and recapitulate its functions. Here, we summarize recent results in vision science and computer vision and organize them as four pieces of evidence in support of the hypothesis that neural representations of visual input are constructed to facilitate the recognition of semantic entities in a manner that is robust to complex transformations.

First, we show that representations that are explicitly designed to support this flavor of recognition can reduce the amount of visual experience required to learn a new task or concept; learning from very few examples is a hallmark trait of human visual intelligence. Second, we summarize recent neurophysiology and brain imaging studies that reveal the presence of robust representations of human actions, human faces, and objects in visual cortex. Third, we review recent studies in computational neuroscience that exposed a positive correlation between how well artificial representations of visual input support the invariant recognition of semantic entities and the degree to which these representations can replicate neural correlation patterns. Finally, we summarize a number of biological predictions that follow from our proposition and show that assuming that invariant recognition shapes neural representations of visual input explains a number of well-known properties of the neural substrate of visual perception in humans and nonhuman primates.

Taken as a whole, these results suggest that supporting recognition tasks, and in particular those that require invariance to complex transformations, is the organizing computational principle that shapes visual cortical representations.

2. INVARIANT REPRESENTATIONS OF VISUAL INPUT

Human visual cortex is organized as a hierarchy of computational layers that transform visual input into a representation of the outside world that is useful to the viewer and supports a variety of perceptual tasks like recognition and navigation (Marr & Nishihara 1978). While much is known about how visual cortex processes its sensory input, why it is that this brain region computes certain representations and not others remains unknown. This review attempts to fill this gap, suggesting that human early visual processing is aimed at constructing robust representations of visual input

that support the recognition of various semantic entities like objects and faces. In this section, we provide a primer on visual representations and formalize our claim. To this end, we abstract the visual system to a network of cortical areas that transform an image $x \in \mathcal{X}$ into a representation

$$\mu = f(x) \qquad 1.$$

that is made available to the rest of the brain for the purpose of planning, navigating, interacting, and, in the most general terms, seeing. Given its central role in visual perception, μ must retain certain information about x, and because $f(\cdot)$ is implemented by neural circuitry, it is subject to numerous constraints (Marr 1982).

First and foremost, $f(\cdot)$ must be available: μ must be a direct function of the visual input x and, crucially, the computations involved in $f(\cdot)$ must be plausibly implemented by neural circuitry. Second, μ must have scope and uniqueness; in other words, μ must be good for something. For example, it is conceivable to define an available $f(\cdot)$ that extracts the average color of the upper-left corner of any input image. However, it is unlikely that an organism with such a useless visual system would escape predators for many consecutive generations. Visual representations must be relevant to their scope—be it recognition, navigation, or planning—and they must be, to some extent, unique to specific stimuli: When the visual input x changes substantially, so does μ (DiCarlo & Cox 2007). Finally, at the other end of that same spectrum, useful representations should be stable with respect to irrelevant perturbations so that the neural response to the sight of our best friend does not change substantially whether she is looking directly at us or slightly to the left.

In this review, we restrict our analysis to those representations that can be computed by purely feed-forward convolutional neural networks (CNNs) (**Figure 1**); $f(x)$, in this case, takes a particular form, and the availability constraint is met. While this restriction confines our discussion to those representations that can be constructed in the first few hundred milliseconds of neural processing in the ventral stream, thereby excluding important neural mechanisms for perception, such as neural feedback and top-down attention, as well as many aspects of human visual intelligence, like navigation and/or spatial relational reasoning, it is appropriate to focus on these early responses, as they account for many crucial aspects of how we as humans perceive and act in the world (Riesenhuber & Poggio 1999, DiCarlo & Cox 2007, Serre et al. 2007a, DiCarlo et al. 2012).

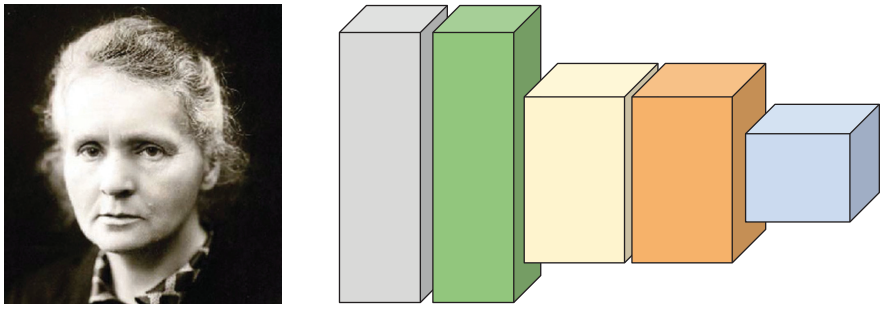


Figure 1

Schematic of a convolutional neural network. The input image x goes through layers of computations $f_l(\cdot)$, and the output of each layer serves as input to the next layer so that $f(\cdot) = f_L \circ \ldots \circ f_1(\cdot)$. Each layer f_l, shown here with a cuboid, is parametrized by a linear map $W_l \in \mathbb{R}^{f_l \times k_{l-1}}$, bias terms $b_l \in \mathbb{R}^{f_l}$, and a nonlinearity function σ_l. The architecture's output $\mu = f(x)$ is a representation of the input image x.

CNNs are currently the most successful and accurate artificial models of how visual cortical representations are built. Much like neural processing in human visual cortex, computations within CNNs are organized as a hierarchy of layers, where the output of each layer serves as input to the next layer. Within each layer, all computational units perform the same operation on different spatially and temporally local regions of input—most commonly, either template matching or pooling operations. As the signal travels through the hierarchy, the size of the receptive field processed by each individual unit increases alongside the specificity of the preferred template. Usually, template dictionaries for convolutional layers are either learned by optimizing performance on supervised (LeCun et al. 1989) or unsupervised (Hinton & Salakhutdinov 2006, Mutch & Lowe 2008) tasks. Recently, specific instances of this class of models achieved human-level performance on a number of perceptual tasks (Kriegeskorte 2015, LeCun et al. 2015), like object recognition (Krizhevsky et al. 2012, Sermanet et al. 2013) and face identification (Schroff et al. 2015).

Restricting our analysis to CNN-computable representations allows us to meet the availability constraint illustrated above. Moreover, the notation introduced in this section lets us introduce the specific hypothesis linking the studies summarized in the following sections in terms of a representation's scope and stability.

In this review, we present evidence in support of the following proposition: The particular scope that constrains visual cortex's output μ is that of supporting the recognition of the semantic entities, such as objects, faces, or actions, that populate and animate a visual scene x. Simultaneously, $\mu = f(x)$ must be invariant to identity-preserving changes in x that leave the semantics of the visual scene, such as the identity of a person or the action someone is performing, unaltered (Anselmi et al. 2016a, Poggio & Liao 2017). Formally, let $g \in \mathcal{G}$ be a transformation $g : \mathcal{X} \to \mathcal{X}$ and let $\mathcal{G}$ be the set of transformations that leave the semantics of x unchanged such as rigid translations or changes in illumination ($\mathcal{G}$ may or may not have a group structure). We can then formalize our goal as follows: We present evidence supporting the notion that human visual cortex implements some function $f(\cdot)$ that enjoys the property

$$f(x) = f(x') \Leftrightarrow \exists g \in \mathcal{G} \quad \text{such that} \quad gx = x', \qquad 2.$$

where the forward direction implies that $\mu = f(x)$ is invariant to all transformations $g \in \mathcal{G}$ and the inverse direction ensures both that $f(\cdot)$ is nontrivial and that μ can be readily employed to discriminate between any two images $(x, y) \in \mathcal{X}$ that do not lie on the same transformation orbit (the orbit generated by x, denoted as $\mathcal{O}_x$, is the set $\mathcal{O}_x = \{gx\}, \quad \forall g \in \mathcal{G}$).

3. HISTORY OF INVARIANT REPRESENTATIONS FOR ARTIFICIAL PERCEPTION

Invariant representations for artificial visual perception have a long history and predate CNNs. While these approaches are not strictly related to human vision, it is worth reviewing them here. The obvious starting point is the Fourier power spectrum, which is translation invariant and, when discretized in the frequency domain, can be used to represent images. More recently, the general picture processing operator (Granlund 1978) was introduced as an explicit hierarchical decomposition, where structures rather than uniformities are carried over from one layer to the next, so that the redundant information is lost and, throughout the layers, the representation becomes more and more invariant to the specifics of the input image. The wavelet decomposition (Mallat 1989) and other similar approaches, such as Laplacian and Gaussian pyramids, produce scale- and translation-invariant representations. Steerable pyramids (Simoncelli & Freeman 1995) were attempts to build linear representations with equivariance properties. Finally, the scattering transform (Bruna & Mallat 2011) is a modern take on the wavelet decomposition and has provable

robustness to rigid as well as nonrigid transformations. Finally, CNNs, the class of networks we focus on here, were initially proposed as compact neural-inspired hierarchical architectures that enforced translation invariance over increasingly larger portions of their inputs (Fukushima 1980, LeCun et al. 1989, Poggio & Edelman 1990).

4. THE CASE FOR INVARIANCE

In this section, we summarize the bulk of the evidence supporting our hypothesis that the visual representations constructed in human visual cortex are aimed at supporting invariant recognition. We first review recent theoretical advances linking the properties of invariant representations to the ability to learn new concepts from very few examples, a salient trait of human visual intelligence. Second, we summarize neurophysiology and brain-imaging studies on humans and nonhuman primates that have revealed the presence of neural representations of visual concepts, like objects and actions, that are invariant to changes in viewpoint or position. Third, we review recent results showing that, within the representations of visual input constructed with CNNs, those that better support invariant recognition more closely match neural data. Finally, we show that the adoption of invariant recognition as an overarching organizational principle of visual cortex implies accurate biological predictions.

4.1. Invariance and Sample Complexity

Modern computer vision systems achieve human-level performance on a number of perceptual tasks (Deng et al. 2009, Krizhevsky et al. 2012) and are successfully used in numerous applications. However, these architectures must be trained with millions of labeled examples to achieve acceptable levels of accuracy. Humans, by contrast, can learn to perform complex visual tasks, like picking someone out in a crowd or recognizing a new object in a different pose, by looking at a single image. This wide separation in sample complexity, the number of supervised examples required by a learning system to achieve a certain performance level, is the core divide between human and artificial perception (Lake et al. 2016).

Recently, a series of studies (Anselmi et al. 2016b, Poggio & Anselmi 2016) has focused on the computation of invariant representations as an important factor for explaining the remarkably low sample complexity of human perception. These studies analyzed a specific, biologically plausible algorithm for computing representations that are selective and invariant to group transformations. In the case of two-dimensional vision, perfect invariance can be achieved with respect to planar transformations of an image: translation, scaling, and rotation in the image plane.

Invariance to transformations can be built in in a system and even learned from visual experience. As we review below, it has been proven that such invariance can yield significant decrease in the sample complexity of learning. The open question is whether visual cortex computes and exploits invariant representations. We conjecture it does.

The intuition that invariance can help reduce the complexity of learning is straightforward. Recognition—in other words, both identification (e.g., of a specific car relative to other cars) and categorization (e.g., distinguishing between cars and airplanes)—would seem indeed much easier if the images of objects were rectified with respect to all transformations or, equivalently, if the image representation itself were invariant. The case of identification is obvious since the difficulty in recognizing exactly the same object (e.g., an individual face) is due only to transformations. In the case of categorization, consider the suggestive evidence from the classification task in **Figure 2**. The figure shows that if an ideal preprocessing module (an oracle) factors out all transformations in images of many different cars and airplanes, providing rectified images with respect to viewpoint,

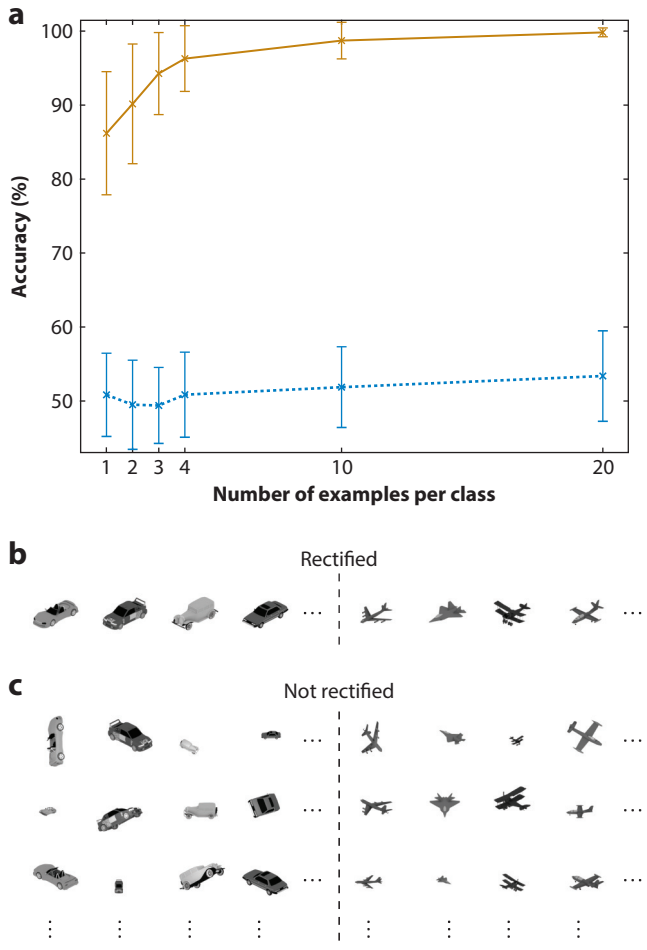


Figure 2

Sample complexity for the task of discriminating images of cars and images of airplanes from their raw pixels. (*a*) Performance of a nearest-neighbor classifier (correlation distance) as a function of the number of training examples. Error bars are ±1 standard deviation computed over 100 random train/test splits. The solid line indicates a rectified task. Shown is a classifier performance where all images (both train and test sets) were rectified. The dashed line indicates the use of unrectified images. (*b*) Examples of rectified images. (*c*) Examples of images of cars and airplanes used for this experiment. Figure adapted from Poggio & Anselmi (2016), with permission from MIT Press.

illumination, position, and scale, the problem of categorizing cars versus airplanes becomes easy: It can be done accurately with very few labeled examples. In this case, good performance was obtained from a single training image of each class, using a simple classifier. In other words, the sample complexity of the problem seems to be very low. We argue in this review that the ventral stream in visual cortex tries to approximate such an oracle, providing a quasi-invariant signature for images.

A qualitative argument involves estimating the cardinality of the universe of possible images generated by different viewpoints—such as variations in scale, position, and rotation in three dimensions (3D)—versus true intraclass variability (e.g., different types of cars). Let us try to

estimate whether the cardinality of the universe of possible images generated by an object originates more from intraclass variability (e.g., different types of dogs) or more from the range of possible viewpoints, including scale, position, and rotation in 3D. Assuming a granularity of a few minutes of arc in terms of resolution and a visual field of, say, 10°, one would get 10^3–10^5 different images of the same object from x, y translations, another factor of 10^3–10^5 from rotations in depth, a factor of 10–10^2 from rotations in the image plane, and another factor of 10–10^2 from scaling. This gives on the order of 10^8–10^{14} distinguishable images for a single object. However, how many different distinguishable (for humans) types of dogs exist within the dog category? It is unlikely that there are more than, say, 10^2–10^3. From this point of view, it is a much greater win to be able to factor out the geometric transformations than the intracategory differences.

Notice that for any representation that is invariant to all transformations $g \in \mathcal{G}$ and selective for $h \in \mathcal{G}'$, there may be a dual representation that is invariant to all transformations $h \in \mathcal{G}'$ but selective for $g \in \mathcal{G}$. In general, they are both needed for different tasks and both can be computed by a CNN module with different pooling strategies. In general, the circuits computing them share a good deal of overlap.

4.1.1. Invariance reduces sample complexity of learning. In a machine learning context, invariance to image translations, for instance, can be built up trivially by memorizing examples of the specific object in different positions. Human vision, by contrast, is clearly invariant for novel objects: People do not have any problem recognizing, in a distance-invariant way, a face seen only once. It is rather intuitive that representations of images that are invariant to transformations such as scaling, illumination, and pose, just to mention a few, should allow supervised learning from many fewer examples.

A proof of the conjecture for the special case of translation or scale or rotation is provided by Anselmi et al. (2016b) and Poggio & Anselmi (2016). For images defined on a grid of pixels, the result (in the case of group transformations such as translation) can be proved using well-known relations between covering numbers and sample complexity.

4.1.1.1. Sample complexity. Sample complexity is the number of examples needed for the estimate of a target function to be within a given error rate. In the example of the number of airplanes or cars (**Figure 2**), we trained the linear classifier to perform the recognition task with a certain precision.

4.1.1.2. Sample complexity for translation invariance. Consider a space of images of dimensions $p \times p$ that may appear in any position within a window of size $rp \times rp$. The natural image representation yields a sample complexity (for a linear classifier) of order $m_{\text{image}} = O(r^2 p^2)$; the invariant representation yields a sample complexity of order

$$m_{\text{inv}} = O(p^2). \qquad 3.$$

This simple observation says that, in the case of translation group, an invariant representation can decrease the sample complexity—that is, the number of supervised examples necessary for a certain level of accuracy in classification. A heuristic rule is that the sample complexity gain is in the order of the number of virtual examples generated by the action of the group on a single image (Niyogi et al. 1998).

4.2. Invariant Neural Representations

The work outlined in the previous paragraph highlighted a strong connection between recognition performance, especially in low sample regimes, and invariant representations. These results can

easily be extended to visual cortex, one of whose primary goals is factoring out identity-preserving transformations. Invariant visual representations have been extensively reviewed in both the human (Grill-Spector & Weiner 2014) and nonhuman primate (DiCarlo et al. 2012) ventral streams. In this subsection, we briefly discuss key neurophysiology and brain imaging studies, in the context of the hypothesis that invariant recognition shapes which neural representations human and nonhuman primate visual cortices compute. In particular, we review more recent work looking beyond object recognition in the domains of invariant face and action recognition.

Single-unit physiological recording has painted a clear picture of the ventral stream of visual cortex. This network of brain areas is organized as a series of anatomically contained regions that transform representations of lines and edges in its earliest layer (V1) (Hubel & Wiesel 1962) to representations of complex shapes in the top layer [inferior temporal cortex (IT)] (Desimone et al. 1984, Gross & Schonen 1992). More recently, population recording and decoding methods have allowed researchers to test the generalization properties of neural codes (Hung et al. 2005) and shown directly a gradual buildup of invariant representations that develop between primary visual cortex, V4, and IT (Rust & DiCarlo 2010).

Visual cortex's hierarchy has been most comprehensively described in the primate face patch network (Tsao et al. 2006). Monkey IT contains a network of several patches of cells that show strong face selectivity. Within these patches, the earliest/most posterior ones contain cells that show a preference for low-level face properties, such as specific face views. The most anterior patch, in contrast, contains representations that are invariant to complex, nonaffine transformations such as changes in viewpoint. Interestingly, by recording throughout the face patch network, Freiwald & Tsao (2010) were able to gain insight into how these invariant representations are constructed. Cells in the middle face patches contained representations that were mirror symmetric, a logical midpoint between view specificity and full-view invariance, providing important insights into how invariance arises in IT (see **Figure 3**).

A similar hierarchical progression of selectivity and invariance has been described in the human ventral pathway (Grill-Spector & Weiner 2014). Moreover, high-temporal-resolution noninvasive and invasive methods have revealed a hierarchy that extends not only in space but also in time. In fact, greater degrees of invariance to transformations progress between 60 and 150 ms (Liu et al. 2009, Carlson et al. 2013, Isik et al. 2014). These findings provide important insight into markers for full semantic, invariant representations that have been revealed for faces and animals at 150 ms (Thorpe et al. 1996).

Beyond static objects and faces, representations develop from simple moving lines and patterns in macaque MT (Rust et al. 2006) to neural recordings in macaque superior temporal sulcus that show selectivity for specific action sequences invariant to changes in actor (Singer & Sheinberg 2010). Recently, through human magnetoencephalography (MEG) decoding, it has been shown that neural recordings can be used to discriminate the video stimuli that elicited them on the basis of action content, across changes in actor and 3D viewpoint (Isik et al. 2017).

Taken together, these results provide compelling evidence that invariant representations of the semantic entities that populate the visual world can be measured in human as well as nonhuman primate brains.

4.3. Invariant Recognition Shapes Neural Representations of Visual Input

The results just summarized show that neural representations of visual input that can efficiently support invariant discriminative tasks are present in the primate brain and can be readily measured with current neuroimaging and neurophysiology techniques. These findings are crucial to build

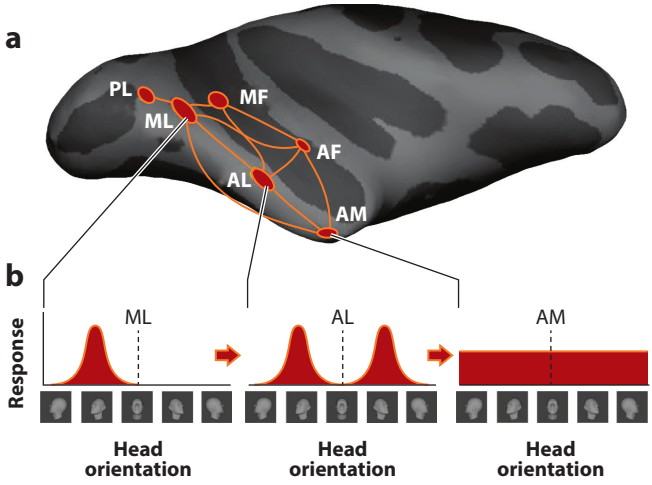


Figure 3

(*a*) Side view of computer-inflated macaque cortex with six areas of face-selective cortex (*red*) in the temporal lobe, together with a connectivity graph (*orange*) (Tsao & Livingstone 2008). Face areas are named on the basis of their anatomical location (AF, anterior fundus; AL, anterior lateral; AM, anterior medial; MF, middle fundus; ML, middle lateral; PL, posterior lateral) and have been found to be directly connected to one another to form a face-processing network (Moeller et al. 2008). Recordings from three face areas—ML, AL, and AM—during presentations of faces at different head orientations revealed qualitatively different tuning properties, schematized in panel *b*. (*b*) Prototypical ML neurons are tuned to head orientation (e.g., a left profile, as shown). A prototypical neuron in AL, when tuned to one profile view, is tuned to the mirror-symmetric profile view as well. And a typical neuron in AM is only weakly tuned to head orientation. Because of this increasing invariance to in-depth rotation, increasing invariance to size and position (not shown), and increased average response latencies from ML to AL to AM, it is thought that the main AL properties, including mirror symmetry, have to be understood as transformations of ML representations and the main AM properties as transformations of AL representations (Freiwald & Tsao 2010). Figure adapted from Leibo et al. (2016), with permission from Elsevier.

the argument that discriminative tasks, and especially those that require invariance to complex transformations, drive the neural computations carried out in human and nonhuman primate visual cortex. However, they fall short of establishing a direct link between the computational principles that drive neural representations and their biological substrate. To fill this gap, recent studies reported that, within the CNN class, models that better support invariant discriminative tasks also produce representations of visual stimuli that better match those implied by neural recordings (Khaligh-Razavi & Kriegeskorte 2014, Yamins et al. 2014, Tacchetti et al. 2017a). This observation demonstrates a connection between a representation's ability to support invariant recognition and its fidelity in replicating neural data. Importantly, these studies confirm this effect across various recognition domains (e.g., objects and actions), different sets of stimuli (e.g., static images and videos), and diverse neural recording techniques [e.g., MEG, neurophysiology, and fMRI (functional MRI)] and utilize different figures of merit to quantify the match between artificial and biological representations (e.g., preferred stimulus ranking and representational similarity analysis). Here, we provide a summary of these studies.

Yamins et al. (2014) recorded IT neural responses of awake and behaving monkeys while they were fixating static images. The image stimulus set, which contained single objects floating on naturalistic backgrounds, was designed specifically to test strong tolerance to object viewpoint

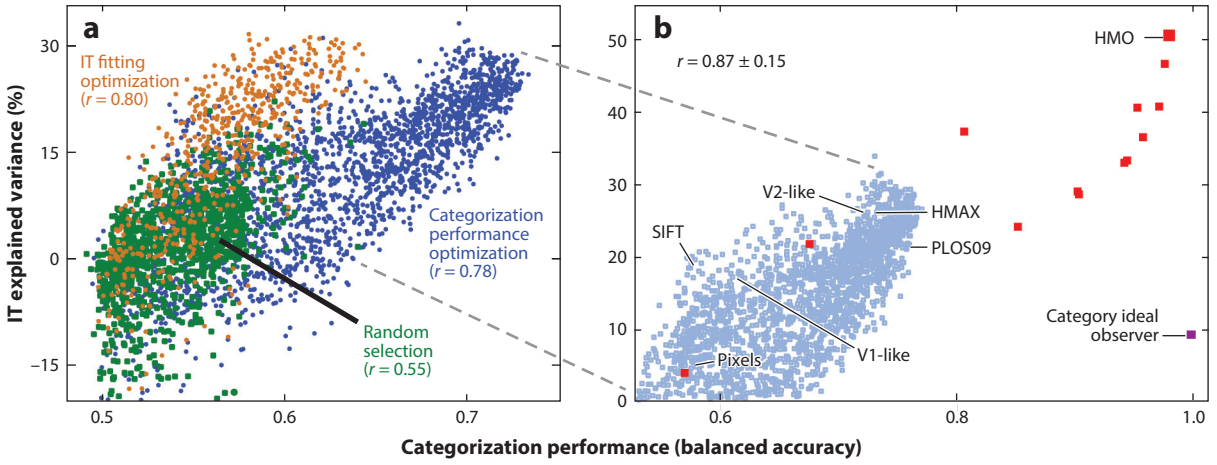


Figure 4

Performance–IT predictivity correlation. (*a*) Object-categorization performance versus IT neural explained-variance percentage (IT predictivity) for convolutional neural network models in three independent high-throughput computational experiments (each point is a distinct neural network architecture). The *x*-axis shows performance (balanced accuracy, chance is 0.5) of the model output features on a high-variation categorization task; the *y*-axis shows the median single-site IT explained-variance percentage ($n = 168$ sites) of that model. Each dot corresponds to a distinct model selected from a large family of convolutional neural network architectures. Models were selected by random draws from parameter space (*green dots*), object-categorization-performance optimization (*blue dots*), or explicit IT predictivity optimization (*orange dots*). (*b*) A high-performing neural network was identified, in pursuing the correlation identified in panel *a*, that matches human performance on a range of recognition tasks, the HMO model. The object-categorization performance versus IT neural predictivity correlation extends across a variety of models exhibiting a wide range of performance levels. Black circles include controls and published models; red squares are models produced during the HMO optimization procedure. The category ideal observer (*purple square*) lies significantly off the main trend but is not an actual image-computable model. The *r*-value is computed over red and black points. For reference, light blue circles indicate performance-optimized models (*blue dots*) from panel *a*. Figure adapted from Yamins et al. (2014), with permission. Abbreviations: HMO, hierarchical modular optimization; IT, inferior temporal cortex; SIFT, scale invariant feature transform.

variation. Subsequently, a CNN model was trained to solve an object classification task on a similar stimulus set. It was found that units in the top layer of the CNN model so trained exhibited selectivity to specific object categories (e.g., a certain unit would produce a stronger output for images containing cars than for any other category) and a large degree of tolerance to viewpoint or instance transformations (i.e., the strong responses would be observed regardless of the specific make and model or specific pose of the car in the image). Interestingly, neural recordings from IT sites exhibited similar characteristics and, by aligning artificial and biological units according to their preferred stimuli, it was shown that units in the top layers of the CNN model accurately predicted the response of IT neurons elicited by new, unseen images. Finally, by randomly sampling model hyperparameters to obtain different model instances, the authors were able to establish a positive correlation between a model's categorization performance and its ability to predict IT neural recordings (see **Figure 4**).

Similar findings were reported using monkey IT neurophysiology data as well as human fMRI recordings (Khaligh-Razavi & Kriegeskorte 2014). Specifically, the authors utilized 27 different artificial systems and constructed, for each model, a representational similarity matrix of a fixed set of stimuli. Likewise, they used recordings of the neural activity these stimuli elicited to construct neural dissimilarity matrices for both humans and monkeys. Finally, they used the correlation between the matrix constructed using each model and the two generated using neural data to

establish a measure of agreement between artificial and neural representations. They reported that, across the 27 models they considered, those representations that better supported categorization between animate and inanimate objects (an invariant discriminative task) also better matched neural correlation patterns.

Recently, these results were extended to human perception of action sequences (Tacchetti et al. 2017a). The authors constructed four instances of spatiotemporal CNN (ST-CNN) models to extract artificial representations of action sequences[1] and, crucially, designed these models to exhibit varying degrees of invariance to changes in viewpoint. In particular, they used a purely convolutional model (Giese & Poggio 2003, Jhuang et al. 2007), two models with memory-based pooling units (Leibo et al. 2011, Anselmi et al. 2016a), and one model with convolutional templates learned by optimizing performance on an action recognition task (LeCun et al. 2015). These models were used to extract representations of a fixed set of action videos. The same videos were shown to human subjects while their brain activity was being recorded with an MEG scanner. Each of the artificial representations was used to compute a dissimilarity matrix that was compared to one constructed with human neural recordings, using a representational similarity analysis score (Kriegeskorte et al. 2008). These results show that, within the ST-CNN class and across the model modifications considered, those representations that better support invariant action recognition also produce a similarity structure that better matches neural correlation patterns. These findings effectively show that the effect described in this section does not concern only primate perception of objects and animals in static images but extends to human perception of others' actions in dynamic sequences.

These results provide a direct link between performance on discriminative tasks, especially those that require generalization across complex transformations, and the representation of visual input that primate visual cortex computes. Importantly, throughout these studies, it is reported that a perfect categorical oracle does not match neural correlation patterns better than convolutional architectures. This consistent result suggests that the computational goal of visual cortex dictates its functioning within a narrowly constrained class of architectures.

4.4. Biological Predictions

The previous paragraphs showed how invariant representations of visual input that support invariant recognition tasks can be measured in the brains of live and behaving primates. Moreover, it was established that artificial representations that better support invariant recognition better match those implied by neural recordings and reduce the sample complexity of learning tasks. Here, we summarize results showing that by constraining artificial models to compute representations that support invariant recognition, one obtains architectures that explain established properties of visual cortex. In particular, results showing that CNNs explicitly trained to support invariant recognition successfully predict the preferred stimuli of neurons in visual areas V1 and V2 as well as the mirror-symmetric preference to face orientation of neurons in face-selective area AL (lateral anterior patch). Moreover, beyond the tuning properties of individual neurons, computing invariant representations of visual input requires specialized modules rather than general-purpose visual processing machinery; these domain-specific regions, like the fusiform face area, are prominent features of the organization of visual cortex.

The preferred stimuli of individual neurons in brain area V1, the earliest cortical layer engaged in visual perception, are well characterized by a universal shape with quantitative parameters that

[1]ST-CNNs are a direct extension of CNNs that work with video stimuli.

hold across species. This property has been observed in cats (Jones & Palmer 1987), macaque monkeys (Ringach 2002), and mice (Niell & Stryker 2008). Recently, is has been shown that by constructing a simple convolutional architecture and programming it to learn a representation that is invariant to the translations it is exposed to during training, one obtains artificial neural units with preferred receptive fields that match the qualitative characteristics of the universal shape observed in cats, monkeys, and mice and, crucially, match the quantitative parameters that hold across species. Moreover, by applying the same ideas to more complex transformations, one obtains artificial receptive fields that resemble those preferred by monkey's V2 neurons (Mutch et al. 2017).

In a similar spirit, a simple CNN model of visual processing instructed to use Hebbian learning (Hebb 1949) to efficiently recognize faces of individuals across changes in 3D viewpoint was found to have artificial units that exhibit a mirror-symmetric preference to face orientation (Leibo et al. 2016). This property had been observed in macaque monkey face-selective area AL's neurons (Freiwald & Tsao 2010) and had yet to be accounted for by similar models of visual processing (see **Figure 5**).

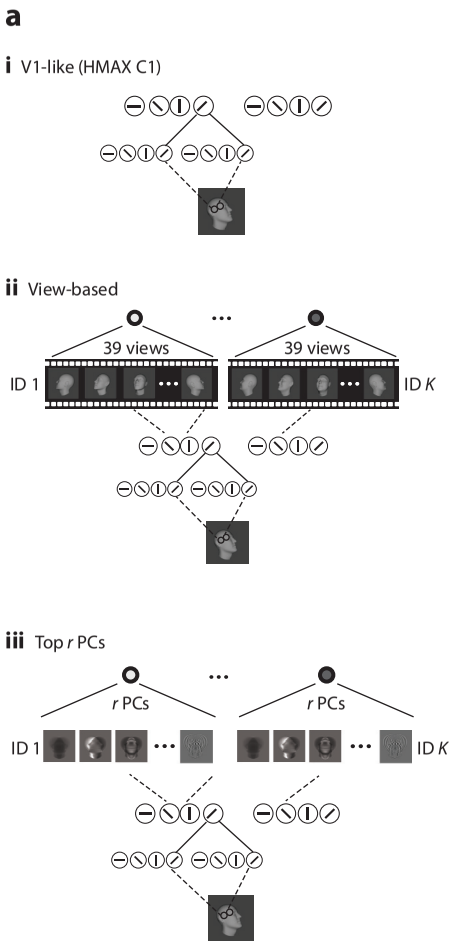
Beyond accurate modeling of individual units, assuming that visual cortex indeed computes invariant-to-transformations representations of visual input implies that objects that do not transform alike—for example, faces and bodies—cannot be processed efficiently by general purpose machinery but require specialized modules (Leibo et al. 2015). This principle explains the presence of domain-specific regions, a prominent feature of visual cortex's organization. For example, anatomically contained networks that selectively engage in the processing of faces (Kanwisher & Yovel 2006) or bodies (Downing et al. 2001) have been observed.

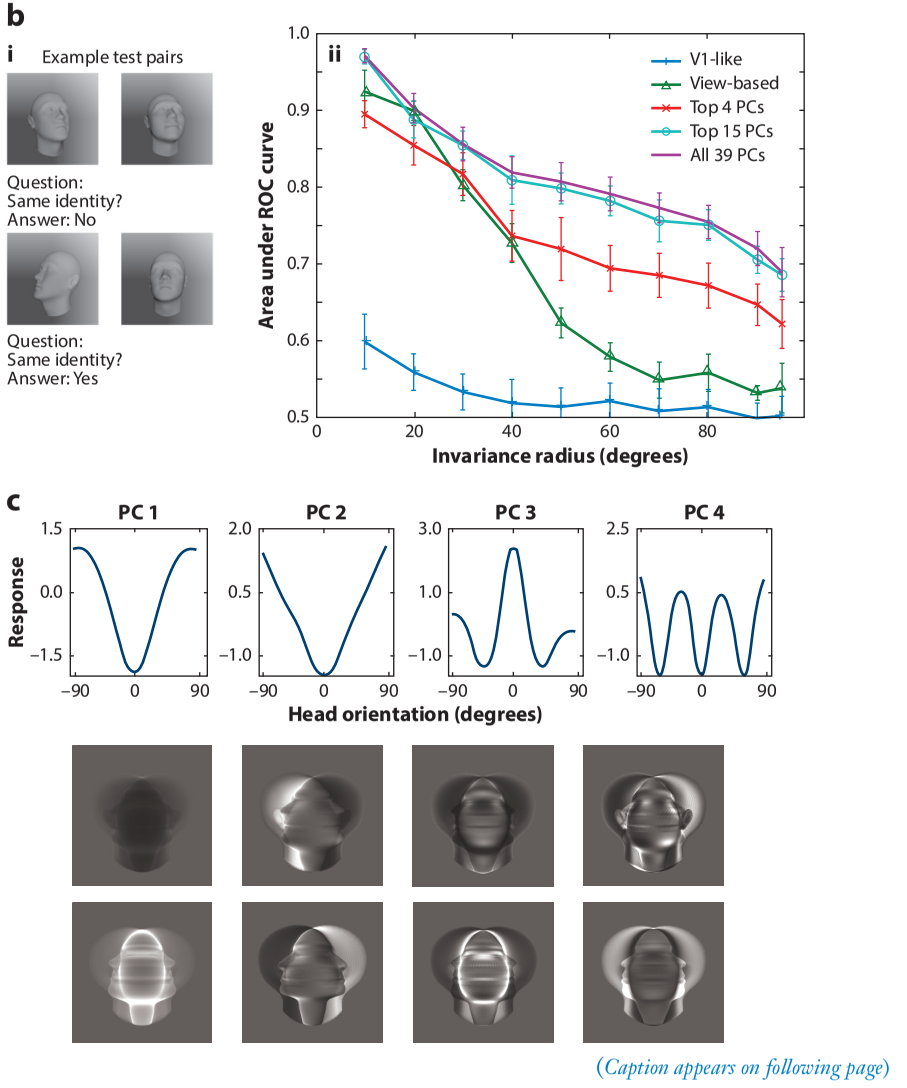
These findings show that models trained with the goal of robustness to complex transformations can recapitulate a wide range of biological characteristics. Importantly, these results are not limited to low-level aspects of neural functions but cover broader organizational principles as well.

5. CONCLUSIONS

We have presented an array of findings supporting the hypothesis that recognizing the semantic category of visual stimuli across photometric, geometric, and more complex changes is the computational principle dictating which representations of the outside world are computed by human visual cortex. In particular, some of the studies summarized here show that representations of the visual world that support invariant discrimination can be measured in the primate brain. Other works have shown how artificial representations that support this robust recognition match empirical dissimilarity structures constructed from neural data and lower the sample complexity of learning tasks. Taken as a whole, these results provide strong support for our claim that supporting invariant recognition is the computational goal dictating the neural computations taking place in human visual cortex and, importantly, span the entire gamut of vision science, from computational to theoretical and neurophysiological results.

Humans rapidly make sense of the visual world around them. We effortlessly recognize the objects, faces, and actions that populate visual scenes, and we are capable of learning new visual concepts from very few examples. These remarkable abilities are supported by the representations of sensory input that our visual cortex computes in the first few hundred milliseconds of neural processing. Understanding the computational goals and requirements that shaped this brain region and how these relate to its neural circuitry is necessary to fully describe its most salient properties and will pave the way for replicating its functioning in artificial systems (see the sidebar titled Applied Invariant Representations).





(*Caption appears on following page*)

Figure 5 (*Figure appears on preceding page*)

(*a*) The structure of the models tested in panel *b*. (*a*, *i*) The V1-like model encodes an input image in the C1 layer of HMAX, which models complex cells in V1 (Riesenhuber & Poggio 1999). (*a*, *ii*) The view-based model encodes an input image as $\mu^k(x) = \sum_{i=1}^{|G|} \langle x, g_i w^k \rangle^2$, where x is the V1-like encoding. ID represents identity, and K denotes the number of individuals present in the templates. (*a*, *iii*) The top principal components (PCs) model encodes an input image as $\mu^k(x) = \sum_{i=1}^{r} \langle x, w_i^k \rangle^2$, where x is the V1-like encoding. The term r represents the number of PCs used. (*b*) The test of depth-rotation invariance required discriminating unfamiliar faces. That is, the template faces did not appear in the test set, so this is a test of depth-rotation invariance from a single example view. (*b*, *i*) In each trial, two face images appear and the task is to indicate whether they depict the same or different faces. They may appear at different orientations from each other. For classification of an image pair (a, b) as depicting the same or a different individual, the cosine similarity of the two representations was compared to a threshold. The threshold was varied systematically to compute the area under the receiver operating characteristic (ROC) curve (AUC). (*b*, *ii*) In each test, 600 pairs of face images were sampled from the set of faces with orientations in the current testing interval. Three hundred pairs depicted the same individual, and 300 pairs depicted different individuals. Testing intervals were $[-x, x]$ for $x = 5°, \ldots, 95°$. The radius of the testing interval x, dubbed the invariance radius, is the abscissa. AUC declines as the range of testing orientations is widened. As long as enough PCs are used, the proposed model performs on par with the view-based model. It even exceeds its performance if the complete set of PCs is used. Both models outperform the baseline HMAX C1 representation. The error bars were computed over repetitions of the experiment with different template and test sets. (*c*) Mirror-symmetric orientation tuning of the raw pixels–based model. $\langle x_\theta, w_i \rangle$ is shown as a function of the orientation of x_θ. Here, each curve represents a different PC. Shown below are the PCs w_i^k, visualized as images. Figure adapted from Leibo et al. (2016), with permission from Elsevier.

APPLIED INVARIANT REPRESENTATIONS

Invariant recognition as a framework to study perception has led to models of visual cortex that reproduce low-level properties as well as organizational principles of the biological system they replicate. Moreover, systems that are built following the notion that a representation of sensory input that is useful to make sense of the world should be invariant to irrelevant transformations and support efficient recognition have been successfully applied to artificial perception problems like object recognition (Krizhevsky et al. 2012, Soatto & Chiuso 2016), face identification (Liao et al. 2014), texture classification (Bruna & Mallat 2011, Freeman & Simoncelli 2011), and speech recognition (Evangelopoulos et al. 2014; Voinea et al. 2014; Zhang et al. 2014a,b). These applied results, beyond being remarkable engineering achievements in their own merit, strengthen our claim concerning visual cortex's computational goal by finding a link between the proposed computational principle and the ability to solve applied problems.

In particular, it has been shown that by letting a CNN architecture learn its templates through memorizing video frames depicting human faces, a biologically plausible learning mechanism, and by letting its pooling units' receptive fields span continuous video segments, one obtains a model representation that is invariant to clutter and successfully supports face identification in a challenging benchmark data set (Liao et al. 2014).

Similarly, in speech recognition, architectures with pooling units' receptive fields extending over changes in pitch support discrimination between spoken digits, even across gender or age (Voinea et al. 2014). The same idea, pooling across pitch shifts, has been applied to more sophisticated CNNs, resulting in a representation that was invariant to changes in pitch and could significantly reduce the sample complexity in a phone- and music-classification task, compared to a noninvariant baseline (Evangelopoulos et al. 2014; Zhang et al. 2014a,b).

Finally, wavelet-based architectures whose output is robust, although not invariant, to translations and local deformations have been proven relevant to texture classification (Bruna & Mallat 2011).

The systems described here were designed to construct invariant representations of visual and auditory stimuli and, remarkably, achieve state-of-the-art performance on challenging perceptual tasks. When considered in the context of this review, the success of these methods provides strong evidence that invariant representations of sensory input are key to solving perceptual tasks and explain abilities that are unique to human visual intelligence, like learning from few examples and generalizing across complex transformations.

Much work still needs to be done for us to fully close the gap between biological perception systems and their artificial replicas. In the following paragraphs, we highlight some of these open questions and frame them in the context of this review.

5.1. Future Directions

5.1.1. Doing away with full supervision. The impressive recent achievements of CNN-like artificial systems have inspired a great deal of research aimed at connecting CNNs to their biological counterpart (Agrawal et al. 2014, Khaligh-Razavi & Kriegeskorte 2014, Yamins et al. 2014, Mutch et al. 2017). Overall, the consensus is that these systems are accurate models of human visual cortex and, when properly trained, quantitatively match many aspects of our visual system. However, the vast majority of these results were obtained using CNN architectures trained with large amounts of supervised data. In contrast, biological systems, like developing children, do not require constant supervised input to learn useful representations of the visual world. This gap has inspired techniques to learn with other sources of information. For instance, the results presented by Tacchetti et al. (2017b), Hénaff et al. (2017), and Wiskott & Sejnowski (2002) suggest that temporal continuity might be sufficient to learn invariant representations. Other ideas have been put forward—for example, colorization, or learning to predict the color version of a black and white image, provides an implicit supervision signal that is sufficient to learn useful representations (Larsson et al. 2017). Likewise, inpainting, in which the values of a few pixels cut out of an image are predicted, has been used to learn useful semantic representations (Pathak et al. 2016). Despite these recent efforts, a definite mechanism delineating how our brain might learn to compute representations of visual input that are invariant to complex transformations and yet selective to the semantics of a visual scene has not yet been described.

5.1.2. Gradient-based learning. Requiring large amounts of supervision is not the only discrepancy between CNNs and human development. In fact, whether and how biological systems could implement the gradient-based learning methods used for parameter tuning in modern CNNs is the subject of active research (Mazzoni et al. 1991, Bengio et al. 2015, Liao et al. 2015). Irrespective of the precise biological mechanisms that could carry out performance optimization on invariant discriminative tasks, computational studies point to its relevance to understanding neural representations of visual scenes (Khaligh-Razavi & Kriegeskorte 2014, Yamins et al. 2014, Yamins & DiCarlo 2016, Tacchetti et al. 2017a).

These considerations provide inspiration for new research, with potentially interesting applications. Identifying substitutes for a fully supervised supervision signal, for example, would make large-scale learning of image representations economical. Similarly, fully understanding whether and how biological systems might implement performance-optimization algorithms might have biological implications well beyond visual perception.

5.1.3. Invariant recognition and sparse coding. Alongside invariant recognition, other computational principles have been put forward to explain neural representations in the ventral stream. In particular, efficient coding (Barlow 1972), especially in the form of sparsity on some function basis (Olshausen & Field 1996), has inspired a number of successful results in modeling the receptive fields of simple cells in V1. More recently, modern CNNs like the ones discussed here have reproduced these same results in V1 with higher accuracy and extended them throughout the ventral stream (see Section 4.4). Further, these CNNs do not require any explicit sparsity constraint, although the convolutional structure of the template-matching layers does impose an

implicit sparsity pressure on the learned templates. More critically, however, the assumption underlying the efficient coding hypothesis, which is apparent in the dictionary-learning framework where it is normally used (Bell & Sejnowski 1997), is that the goal of the ventral stream is to represent the visual world with a degree of detail that affords a complete reconstruction of the scene. The error signal that would enable learning such a representation during development, however, seems unnatural, as we cannot readily compare visual stimuli and their reconstructions directly in the input space. Looking ahead, it is important to note that sparsity, or efficiency, and invariant recognition are not mutually exclusive, and a representation that is useful for recognition (and is learned as such) can be sparse. Combining these two requirements is a direction of research that is ripe for further exploration.

5.1.4. The role of the architecture and the need for a theory. Arguably, the most interesting future research direction connected to the work presented here concerns the role and origin of the convolutional and hierarchical organization of primate visual cortex itself. Throughout this review, we have summarized recent results in support of a precise computational theory of visual perception. All the computational evidence we have presented has made use of CNNs to model and understand the human visual system. However, whether the well-known convolutional and hierarchical structure typical of this brain region is necessary to compute invariant and selective representations of visual input, and therefore a consequence of this very same computational goal, or the architecture was determined by external factors and simply constrains the kinds of representations visual cortex could compute remains an open question (Poggio et al. 2016). This research direction is tightly linked to the lack of a theoretical understanding of CNN-like architectures. While these systems vastly outperform other architectures in a broad range of artificial perception tasks, our current theoretical tools do not cover CNN-like architectures (Zhang et al. 2017a). This knowledge gap might leave the impression that by modeling the human brain with CNNs, we are abandoning the study of a hopelessly complex system to embrace that of another system that is just as complex (Kriegeskorte 2015). However, the access to faithful and behaviorally accurate models of human perception has allowed us to explicitly investigate various computational hypotheses, and, moreover, much theoretical work is in progress to push the boundaries of our theoretical frameworks to include these systems (Poggio & Liao 2017, Zhang et al. 2017b).

5.2. Final Remarks and Broad Outlook

In this review, we have sought to validate the hypothesis that supporting invariant recognition is the chief computational goal of human visual cortex, a goal that shapes which representations of the visual world this region computes and passes along to the rest of the visual system. In the last several decades, this brain region has been the subject of intense investigation, an effort that produced a detailed understanding of its function. Grounded in this knowledge, and inspired by recent advances in artificial perception systems, the results we have presented here show that a simple computational principle—robust recognition of semantic entities—can relate findings across the entire gamut of vision science. These connections will hopefully accelerate our understanding of human visual intelligence and catalyze its replication in artificial systems.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This review was supported by the Center for Brains, Minds and Machines and funded by the National Science Foundation Science and Technology Center award CCF-1231216. We thank the McGovern Institute for Brain Research at MIT for its support.

LITERATURE CITED

Agrawal P, Stansbury D, Malik J, Gallant J. 2014. Pixels to voxels: modeling visual representation in the human brain. arXiv:1407.5104v1 [q-bio.NC]

Anselmi F, Leibo JZ, Rosasco L, Mutch J, Tacchetti A, Poggio T. 2016a. Unsupervised learning of invariant representations. *Theor. Comput. Sci.* 633:112–21

Anselmi F, Rosasco L, Poggio T. 2016b. On invariance and selectivity in representation learning. *Inf. Inference* 5(2):134–58

Barlow HB. 1972. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1(4):371–94

Bell AJ, Sejnowski TJ. 1997. The "independent components" of natural scenes are edge filters. *Vis. Res.* 37(23):3327–38

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

Bruna J, Mallat S. 2011. Classification with scattering operators. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1561–66. Los Alamitos, CA: IEEE

Carlson T, Tovar D, Alink A, Kriegeskorte N. 2013. Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13(10):1

Connor C, Brincat S, Pasupathy A. 2007. Transformation of shape information in the ventral pathway. *Curr. Opin. Neurobiol.* 17(2):140–47

Deng J, Dong W, Socher R, Li L-J, Li K, Li F-F. 2009. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–55. Los Alamitos, CA: IEEE

Desimone R, Albright T, Gross C, Bruce C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4(8):2051–62

DiCarlo J, Cox D. 2007. Untangling invariant object recognition. *Trends Cogn. Sci.* 11(8):333–41

DiCarlo J, Zoccolan D, Rust N. 2012. How does the brain solve visual object recognition? *Neuron* 73(3):415–34

Downing P, Jiang Y, Shuman M, Kanwisher N. 2001. A cortical area selective for visual processing of the human body. *Science* 293(5539):2470–73

Evangelopoulos G, Voinea S, Zhang C, Rosasco L, Poggio T. 2014. Learning an invariant speech representation. arXiv:1406.3884 [cs.SD]

Felleman D, Van Essen D. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1(1):1

Freeman J, Simoncelli E. 2011. Metamers of the ventral stream. *Nat. Neurosci.* 14(9):1195–201

Freiwald WA, Tsao DY. 2010. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* 330(6005):845–51

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

Gallant J, Braun J, Van Essen D. 1993. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science* 259(5091):100–3

Giese MA, Poggio T. 2003. Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* 4(3):179–92

Granlund GH. 1978. In search of a general picture processing operator. *Comput. Graph. Image Proc.* 8(2):155–73

Grill-Spector K, Weiner KS. 2014. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* 15(8):536–48

Gross CG, Schonen SD. 1992. Representation of visual stimuli in inferior temporal cortex [and Discussion]. *Philos. Trans. Biol. Sci.* 335(1273):3–10
Hebb D. 1949. *The Organization of Behavior*. New York: Psychol. Press
Hénaff O, Goris R, Simoncelli E. 2017. Perceptual straightening of natural video trajectories. *J. Vis.* 17(10):402
Hinton G, Salakhutdinov R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–7
Hubel D, Wiesel T. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160(1):106–54
Hung CP, Kreiman G, Poggio T, DiCarlo JJ. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310(5749):863–66
Isik L, Meyers E, Leibo J, Poggio T. 2014. The dynamics of invariant object recognition in the human visual system. *J. Neurophysiol.* 111(1):91–102
Isik L, Tacchetti A, Poggio T. 2017. A fast, invariant representation for human action in the visual system. *J. Neurophysiol.* 119(2):631–40
Jhuang H, Serre T, Wolf L, Poggio T. 2007. *A biologically inspired system for action recognition*. Paper presented at the 11th International Conference on Computer Vision, Rio de Janeiro, Braz., Oct. 14–21
Jones J, Palmer L. 1987. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* 58(6):1233–58
Kanwisher N, Yovel G. 2006. The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. B* 361(1476):2109
Khaligh-Razavi S-M, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput. Biol.* 10(11):e1003915
Kriegeskorte N. 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1:417–46
Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4
Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–105. Red Hook, NY: Curran Assoc.
Lake B, Ullman T, Tenenbaum J, Gershman S. 2016. Building machines that learn and think like people. arXiv:1604.00289 [cs.AI]
Larsson G, Maire M, Shakhnarovich G. 2017. Colorization as a proxy task for visual understanding. arXiv:1703.04044 [cs.CV]
LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44
LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, et al. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1(4):541–51
Leibo J, Liao Q, Anselmi F, Freiwald W, Poggio T. 2016. View-tolerant face recognition and Hebbian learning imply mirror-symmetric neural tuning to head orientation. *Curr. Biol.* 27(1):1–6
Leibo J, Liao Q, Anselmi F, Poggio TA. 2015. The invariance hypothesis implies domain-specific regions in visual cortex. *PLOS Comput. Biol.* 11(10):e1004390
Leibo J, Mutch J, Poggio T. 2011. *Learning to discount transformations as the computational goal of visual cortex*. Paper presented at the IEEE Conference on Vision and Pattern Recognition, Colorado Springs, CO, June 20–25
Liao Q, Leibo J, Poggio T. 2014. Unsupervised learning of clutter-resistant visual representations from natural videos. arXiv:1409.3879v1 [cs.CV]
Liao Q, Leibo J, Poggio T. 2015. How important is weight symmetry in backpropagation? arXiv:1510.05067 [cs.LG]
Liu H, Agam Y, Madsen JR, Kreiman G. 2009. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62(2):281–90
Mallat SG. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 11(7):674–93
Marr D. 1982. *Vision*. Cambridge, MA: MIT Press
Marr D, Nishihara HK. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. B* 200(1140):269–94

Mazzoni P, Andersen R, Jordan M. 1991. A more biologically plausible learning rule for neural networks. *PNAS* 88(10):4433–37

Moeller S, Freiwald WA, Tsao DY. 2008. Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science* 320(5881):1355

Mutch J, Anselmi F, Tacchetti A, Rosasco L, Leibo J, Poggio T. 2017. Invariant recognition predicts tuning of neurons in sensory cortex. In *Computational and Cognitive Neuroscience of Vision*, ed. Q Zhao, pp. 85–104. Singapore: Springer Singapore

Mutch J, Lowe D. 2008. Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* 80(1):45–57

Niell C, Stryker M. 2008. Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* 28(30):7520–36

Niyogi P, Girosi F, Poggio T. 1998. Incorporating prior information in machine learning by creating virtual examples. *Proc. IEEE* 86(11):2196–208

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

Pathak D, Krähenbühl P, Donahue J, Darrell T, Efros A. 2016. Context encoders: feature learning by inpainting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–44. Los Alamitos, CA: IEEE

Poggio TA, Anselmi F. 2016. *Visual Cortex and Deep Networks: Learning Invariant Representations*. Cambridge, MA: MIT Press

Poggio T, Edelman S. 1990. A network that learns to recognize three-dimensional objects. *Nature* 343(6255):263–66

Poggio T, Liao Q. 2017. Theory II: landscape of the empirical risk in deep learning. arXiv:1703.09833 [cs.LG]

Poggio T, Mhaskar H, Rosasco L, Miranda B, Liao Q. 2016. Why and when can deep—but not shallow—networks avoid the curse of dimensionality: a review. arXiv:1611.00740 [cs.LG]

Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2(11):1019–25

Ringach D. 2002. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* 88(1):455–63

Rust N, DiCarlo JJ. 2010. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* 30(39):12978–95

Rust N, Mante V, Simoncelli E, Movshon JA. 2006. How MT cells analyze the motion of visual patterns. *Nat. Neurosci.* 9:1421–31

Schroff F, Kalenichenko D, Philbin J. 2015. FaceNet: a unified embedding for face recognition and clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–23. Los Alamitos, CA: IEEE

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. 2013. OverFeat: integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229 [cs.CV]

Serre T, Kreiman G, Kouh M, Cadieu C. 2007a. A quantitative theory of immediate visual recognition. *Prog. Brain Res.* 165:33–56

Serre T, Oliva A, Poggio T. 2007b. A feedforward architecture accounts for rapid categorization. *PNAS* 104(15):6424–29

Simoncelli EP, Freeman WT. 1995. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *Proceedings of the International Conference on Image Processing*, Vol. 3, pp. 444–47: Los Alamitos, CA: IEEE

Singer J, Sheinberg D. 2010. Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. *J. Neurosci.* 30:3133–45

Soatto S, Chiuso A. 2016. Visual representations: defining properties and deep approximations. arXiv:1411.7676 [cs.CV]

Tacchetti A, Isik L, Poggio T. 2017a. Invariant recognition drives neural representations of action sequences. *PLOS Comput. Biol.* 13(12):e1005859

Tacchetti A, Voinea S, Evangelopoulos G. 2017b. Discriminate-and-rectify encoders: learning from image transformation sets. arXiv:1703.04775 [cs.CV]

Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. *Nature* 381:520–22

Tsao D, Freiwald W, Tootell R. 2006. A cortical region consisting entirely of face-selective cells. *Science* 311(5761):670

Tsao D, Livingstone M. 2008. Mechanisms of face perception. *Annu. Rev. Neurosci.* 31:411–37

Voinea S, Zhang C, Evangelopoulos G, Rosasco L, Poggio T. 2014. Word-level invariant representations from acoustic waveforms. In *Proceedings of the Annual Conference of the International Speech Communication Association*, *INTERSPEECH*, pp. 2385–89. Red Hook, NY: Curran Assoc.

Wiskott L, Sejnowski T. 2002. Slow feature analysis: unsupervised learning of invariances. *Neural Comput*. 14(4):715–70

Yamins D, DiCarlo J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19(3):356–65

Yamins D, Hong H, Cadieu C, Solomon E, Seibert D, DiCarlo J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111(23):8619–24

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. 2017a. Understanding deep learning requires rethinking generalization. arXiv:1611.03530 [cs.LG]

Zhang C, Evangelopoulos G, Voinea S, Rosasco L, Poggio T. 2014a. A deep representation for invariance and music classification. arXiv:1404.0400 [cs.SD]

Zhang C, Liao Q, Rakhlin A, Sridharan K, Miranda B, et al. 2017b. *Theory of deep learning III: generalization properties of SGD*. Center Brains Minds Mach. Memo 067. **https://cbmm.mit.edu/sites/default/files/publications/CBMM-Memo-067.pdf**

Zhang C, Voinea S, Evangelopoulos G, Rosasco L, Poggio T. 2014b. Phone classification by a hierarchy of invariant representation layers. In *Proceedings of the Annual Conference of the International Speech Communication Association*, *INTERSPEECH*, pp. 2346–50. Red Hook, NY: Curran Assoc.