Annual Review of Vision Science

Deep Learning: The Good, the Bad, and the Ugly

Thomas Serre

Department of Cognitive Linguistic and Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, Rhode Island 02818, USA; email: Thomas_Serre@brown.edu

Annu. Rev. Vis. Sci. 2019. 5:399–426

First published as a Review in Advance on August 8, 2019

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

https://doi.org/10.1146/annurev-vision-091718-014951

Keywords

deep learning, neural networks, artificial intelligence, computational neuroscience, object recognition, face recognition, action recognition, image segmentation

Abstract

Artificial vision has often been described as one of the key remaining challenges to be solved before machines can act intelligently. Recent developments in a branch of machine learning known as deep learning have catalyzed impressive gains in machine vision—giving a sense that the problem of vision is getting closer to being solved. The goal of this review is to provide a comprehensive overview of recent deep learning developments and to critically assess actual progress toward achieving human-level visual intelligence. I discuss the implications of the successes and limitations of modern machine vision algorithms for biological vision and the prospect for neuroscience to inform the design of future artificial vision systems.

1. INTRODUCTION

There is not a week that goes by without artificial intelligence (AI) making a news headline. AI has become increasingly ubiquitous in our everyday lives, challenging our superiority complex over machines: AI has now beaten the best human players at Atari games (Mnih et al. 2015), Go (Silver et al. 2016), chess, and Shogi (Silver et al. 2018), and it is also capable of achieving this feat without any human knowledge (Silver et al. 2017). The engine behind these tantalizing successes is a branch of machine learning known as deep learning. Because computers can effortlessly sift through data at scales far beyond human capabilities, deep learning is not only about to transform modern society, but also about to revolutionize science—crossing major disciplines from particle physics (Radovic et al. 2018) and organic chemistry (Segler et al. 2018) to biological research and biomedical applications (Baldi 2018, Wainberg et al. 2018). In the area of computer vision, we even hear claims that deep artificial neural networks have reached superhuman capabilities (Cireşan et al. 2012, He et al. 2016, Lee et al. 2017, Phillips et al. 2018) on a wide range of visual recognition problems (see **Figure 1**).

From a neuroscience perspective, the very success of modern artificial neural networks provides computational evidence for a toolkit of neural computations that was hypothesized decades ago. At the same time, critical limitations of modern architectures are becoming increasingly clear. Thus, the successes and failures of machine vision are already starting to shape our understanding of the computations underlying vision. A case in point is visual recognition, for which much of the early successes of deep learning stemmed from a class of feedforward preattentive neural networks known as convolutional neural networks (CNNs). On the one hand, CNN successes in natural image categorization (Cireşan et al. 2012, He et al. 2016, Lee et al. 2017, Phillips et al. 2018) provide computational evidence for the long-held hypothesis that rapid visual categorization is possible in the absence of cortical feedback, from a single feedforward sweep of activity through our visual cortex. On the other hand, known CNN limitations in solving basic visual reasoning tasks (Ellis et al. 2015, Stabinger et al. 2016, Kim et al. 2018) provide potentially novel hypotheses for the computational role of the many cognitive processes (including attention, memory, and executive control) that are lacking in these architectures and that are known to play a role in biological vision.

Section 2 provides a brief historical context and highlights key deep learning developments that have led to the recent breakthroughs in visual recognition. In Section 3, I highlight how these recent innovations have led to significant gains in the ability of CNN architectures to account for an array of brain data. In Section 4, I further highlight some of the main differences in the ways that machines and humans tackle various recognition problems and, in particular, CNNs' inability to account for a host of human behavioral data. In Section 5, I review work underscoring some of the most serious limitations of current architectures, from a symptomatic sensitivity to noise perturbations to a limited ability to learn abstract representations and to generalize beyond training data. Finally, I conclude this review by providing pointers to novel computational modules, motivated in some cases by neuroscience considerations, from attentional mechanisms to mnemonic and other arithmetic operations. While additional research will be needed to validate this extended toolkit of neural computations, I see this integration between brain and computer science as one of the most promising directions to move the field of machine vision forward.

2. DEEP LEARNING: RISE OF THE MACHINES

The deep learning revolution is often said to have started in 2012 when a CNN crushed the competition at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). That year, a

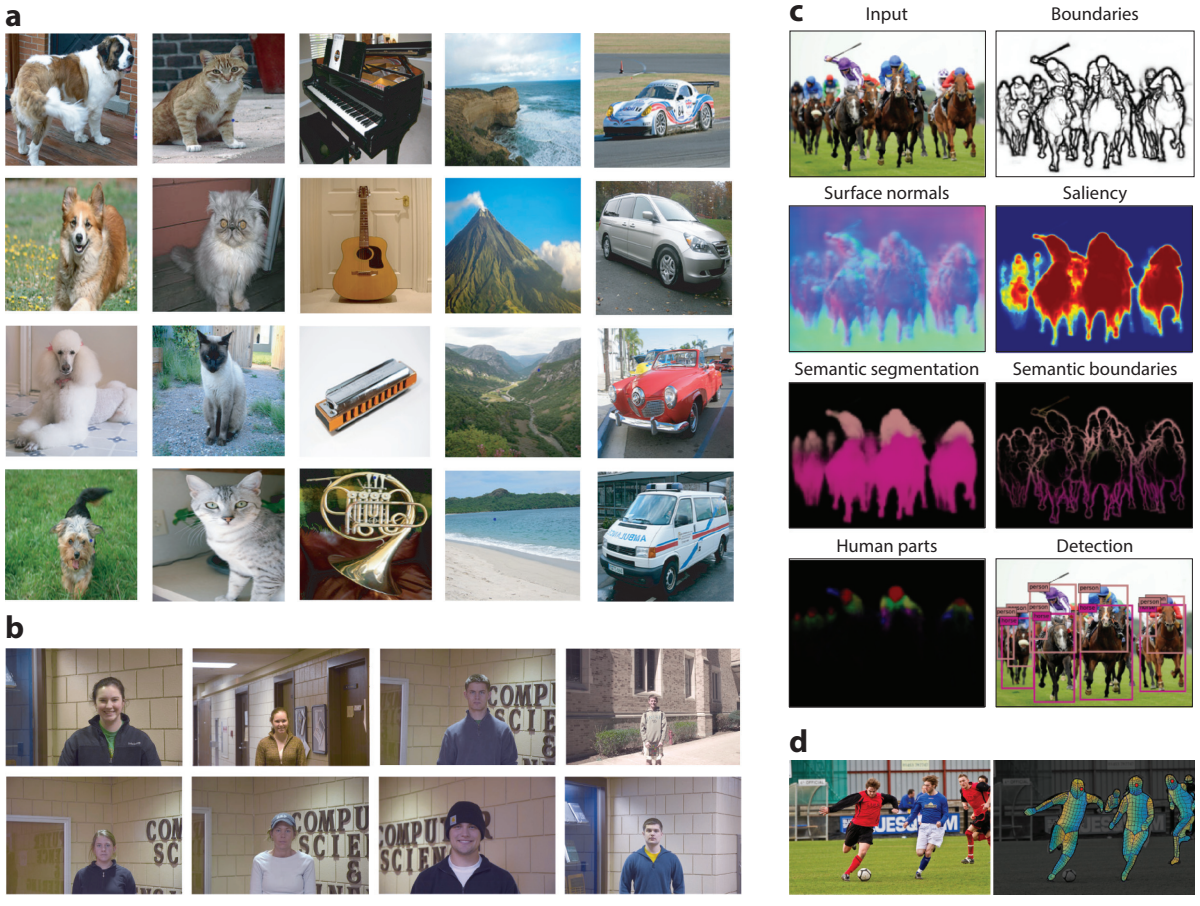

Figure 1

Modern visual recognition challenges used to evaluate machine vision algorithms. (*a*) ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Images are shown from 20 representative image subordinate categories sampled from the 1,000 represented in the data set organized into five basic categories in columns (from left to right and from top to bottom): domestic dogs (Saint Bernard, corgi, poodle, terrier), domestic cats (tiger, Persian, Siamese, Egyptian), musical instruments (piano, guitar, harmonica, French horn), geological formations (cliff, volcano, valley, seashore), and cars (race car, minivan, convertible, ambulance). Some of the best neural network architectures (Hu et al. 2019) achieve a top-5 accuracy above 96% (categorization is considered correct if the correct label is included in a system's top 5 predictions). Image credit: Derick Macklin Toth (adapted with permission). (*b*) Face recognition. Modern neural networks for face recognition already outperform facial forensic experts with an accuracy of approximately 96% (Phillips et al. 2018). Shown clockwise from the top left are representative face pairs used in the study, highlighting how difficult the task is: different-identity pair 5, same-identity pair 6, same-identity pair 4, and different-identity pair 1. Panel adapted from Phillips et al. (2018) with permission from the authors. (*c*) Sample dense prediction tasks solved with the same network, called UberNet (Kokkinos 2017). UberNet can perform tasks spanning low- to mid- and high-level vision within a single unified neural network architecture (see the sidebar titled Image Segmentation and Other (Per-Pixel) Dense Image Prediction Tasks, below). Image credit: Iasonas Kokkinos (adapted with permission). (*d*) Dense human pose estimation on the DensePose data set. Modern neural networks learn to predict 3D volumetric information about human bodies from single 2D images. Panel adapted from Güler et al. (2018) with permission from the authors.

network (Krizhevsky et al. 2012), since dubbed AlexNet (after its lead developer, Alex Krizhevsky), achieved a top-5 accuracy (response considered correct if the correct class label is included in the top 5 network outputs; chance level: 5/1000) of 83.6%—outperforming the second-best system by over 10% and cutting down the error rate compared to the previous year by over 60% (Russakovsky et al. 2015). The network shared many architectural details with earlier so-called feedforward hierarchical models of the visual cortex (see, e.g., Fukushima 1980, LeCun et al. 1998, Riesenhuber & Poggio 1999).

Some of these earlier models were indeed grounded in the anatomy and physiology of the ventral stream of the visual cortex, which is known to play a key role in recognition of objects (Riesenhuber & Poggio 1999, DiCarlo et al. 2012, Serre 2015). The central realization behind these models is that the gradual buildup in the tuning and invariance properties of neurons along the ventral stream could be approximated well by a cascade of convolutions (conv) and local pooling (pool) operations. Many of the key building blocks found in modern CNNs were already introduced in early computational neuroscience models, including the rectified linear unit or ReLu, to prevent negative firing rates and weight sharing (Fukushima 1980), as well as max pooling, subsampling, and contrast normalization (Riesenhuber & Poggio 1999). Even the distinction between convolutional layers for building visual feature representations and fully connected layers for classification was already present in these early models (Fukushima 1980, LeCun et al. 1998, Riesenhuber & Poggio 1999). One important distinction, however, between feedforward hierarchical models of the visual cortex (Fukushima 1980, Riesenhuber & Poggio 1999) and their computer vision relatives (LeCun et al. 1998, Krizhevsky et al. 2012) is the degree to which visual representations are constrained by task demand.

Pooling: aggregating layer outputs locally (over a spatial range, typically via a mean or max operation)

ReLu: a unit that uses the rectified linear function defined as the positive part of its argument: $f(x) = x^+ = \max(0,x)$, where x is the input to the unit

Weight sharing: the same weight vector is shared across all locations to implement the convolution

Subsampling: also called downsampling, aims to reduce the spatial resolution of a visual representation via a pooling operation; its antonym is upsampling

Convolutional layers: implementation of convolutions (essentially the inner product between two vectors) for building visual feature representations

Fully connected layers: typically used as the final dense layers used to associate visual feature representations to classification units, and thus no weight sharing takes place

2.1. Learning Visual Representations

With the exception of the fully connected (classification) layers, which are trained with explicit supervision (i.e., using ground truth category labels), training of the convolutional layers in feedforward hierarchical models of the visual cortex used (Hebbian-like) unsupervised learning mechanisms. In a sense, these earlier models learn to represent visual features frequently encountered in natural scenes irrespective of whether these features are useful for visual recognition. This alleviates the need to propagate error signals backward from classification units in higher stages to lower convolutional stages during learning, as done with the backpropagation algorithm used in CNNs (see the sidebar titled Computing with Gradients), which has long been considered implausible for biological neural networks (Crick 1989; but for more biologically plausible approximations, see Bengio et al. 2015, Miconi et al. 2018, Moldwin & Segev 2018, Scellier & Bengio 2017, Whittington & Bogacz 2017). Unsupervised Hebbian-like learning also seems more consistent with neural recordings that had suggested that learning and plasticity in the ventral stream is driven by a subject's visual experience and is unaffected by reward signals (Li & Dicarlo 2012, Logothetis et al. 1995). Indeed, until relatively recently, the reliance on unsupervised learning mechanisms was not limited to computational neuroscience models; most of the work in computer vision before 2012 largely focused on the problem of engineering visual representations and/or learning them without any explicit task-driven supervision (for a review, see Tosic & Frossard 2011). Modern deep neural networks, however, use supervision from bottom to top layers (called end-to-end) to optimize their learned visual representations for specific tasks. Because of the large number of free parameters in these networks (i.e., the weights), training in this way requires very large image databases such as ILSVRC (with over 1 million samples for 1,000 object categories) (Russakovsky et al. 2015).

COMPUTING WITH GRADIENTS

The notion of a gradient, which is a multivariable generalization of the derivative, is key in nearly every aspect of deep learning. Unlike in earlier computational neuroscience network models, connection weights in convolutional neural networks are learned via a supervised training method known as backpropagation—a shorthand for the backward propagation of errors. Backpropagation follows a gradient descent approach to adjust the weights of a neural network so as to minimize a cost function (also called an error or energy function) for the network. There are many ways to define a network's cost function, and classification error is only one such method. Because backpropagation requires the computation of the gradient of this cost function, which typically (but not necessarily) means that a desired target value is known, it is considered to be a supervised learning method, although it is used in some unsupervised networks such as autoencoders. In effect, backpropagation is a generalization of the delta rule to multilayered feedforward networks, made possible by using the chain rule to iteratively compute the gradient for each layer. Because of the need to propagate gradient signals backward, backpropagation has long been considered implausible for biological neural networks (Crick 1989), but plausible alternatives have been described in recent years (e.g., Bengio et al. 2015, Miconi et al. 2018, Scellier & Bengio 2017, Whittington & Bogacz 2017).

Backpropagation is a special case of a more general technique called automatic differentiation (also called algorithmic differentiation or computational differentiation), which includes multiple numerical methods for evaluating the gradient of a cost function by exploiting the fact that the computations carried by a neural network can be decomposed as a sequence of elementary arithmetic operations and elementary functions. By applying the chain rule repeatedly to these operations, it is possible to compute gradients of arbitrary order automatically and accurately to working precision. Much of the success and widespread adoption of deep learning by the computer vision community stem from the public availability of algorithmic differentiation libraries such as Keras, Tensorflow (Abadi et al. 2016), or pyTorch (Paszke et al. 2017) that allow anyone (without even any knowledge of calculus) to train arbitrary networks. What used to take months, implementing and training a neural network, can now be done literally in minutes.

Beyond training networks, many of the methods mentioned in this review (including the feature visualization and attribution methods described in Section 3) require the computation of the gradient of a cost function that maps images to parameters (e.g., the gradient of arbitrary unit activations with respect to an image).

Perhaps one of the most surprising discoveries of the past few years is the degree of transfer exhibited by ILSVRC-optimized visual representations to novel tasks. Indeed, when developing a vision system for a novel recognition task, most researchers freeze the convolutional layers of their network (which are not retrained), reusing instead the convolutional layers from an ILSVRC-optimized network—only retraining the fully connected layers that implement task-specific classification circuits at the top by learning to associate visual representations derived from the convolutional layers with category labels. ILSVRC-optimized visual representations appear to be relatively generic in the sense that they appear to be good all-around visual representations that are useful for solving a wide variety of visual tasks beyond the specific ones used to train the network (Donahue et al. 2014, Oquab et al. 2014, Zhou et al. 2014). In a sense, the long-held dream of computational neuroscientists to identify the mechanisms by which the ventral stream would learn generic and invariant visual representations was realized by considering not unsupervised learning mechanisms (assumed to be biologically plausible), but rather supervised, task-specific learning mechanisms (assumed to be rather biologically implausible).

It remains unclear whether such invariant and generic visual representations can be learned without supervision. A large research effort focuses on unsupervised learning mechanisms (e.g., autoencoders, variational autoencoders, prediction networks) but so far results have been somewhat unconvincing—with visual representations trained without explicit supervision vastly

underperforming those trained with supervision. An alternative strategy has been to focus on the identification of so-called auxiliary tasks (i.e., tasks that can be used to pretrain a neural network but that are secondary to the main task for which the network is being trained). Examples of auxiliary tasks include asking neural networks to colorize grayscale images, to fill-in image holes, to solve jigsaw puzzles made from image patches, or to predict movement in videos (for an overview, see Doersch & Zisserman 2017). Researchers' hope is to discover auxiliary tasks that can be used to reduce the number of labeled examples needed to train modern neural networks.

2.2. Rolling in the Deep

In the past six years, the top-5 error on ILSVRC has continued to decrease, down from 16.4% in 2012 to 11.7% in 2013 and 6.7% in 2014, which marks the year in which the field realized that deeper networks perform better. While the core building blocks of modern neural networks were identified decades ago, recent gains achieved in computer vision would not have been possible without several innovations that have been discovered in the past few years and that have allowed the training of increasingly deep CNNs.

Two popular networks that are often considered to be the first truly deep networks include the 2014 ILSVRC winner, called GoogLeNet, with 22 layers (Szegedy et al. 2015), and the runner-up, called VGG (named after the Visual Geometry Group at Oxford), with 19 layers (Simonyan & Zisserman 2015). For comparison, earlier networks included eight layers or fewer (Fukushima 1980, Krizhevsky et al. 2012, LeCun et al. 1998, Riesenhuber & Poggio 1999). Another innovation of the GoogLeNet architecture includes the development of the inception module that allows for convolutions with different filter sizes and pooling operations within the same layer—allowing the network, in a sense, to identify which sizes to use through training. Another key building block introduced is the 1×1 convolution that allows an inception module to be implemented without blowing up the network dimensionality (by allowing larger $N \times M$ convolutional kernels to be approximated by 1×1 kernels, leading to a deeper architecture with much fewer parameters).

A milestone was achieved in 2015 when a deep residual network (ResNet) architecture (He et al. 2016) achieved an error rate of 3.5% with a mind-blowing 152 layers. The network is considered to be the first to have surpassed the human level of accuracy, estimated to have an error rate of approximately 5% (Russakovsky et al. 2015) (although the human accuracy measure was computed casually and should not be taken too literally). The main innovation in the ResNet included the design of the residual module: Building on the inception idea, in a residual module, an input goes through a series of conv-ReLu-conv stages, the result of which is then added to the original input. The main assumption is that residual modules help the gradient flow by allowing it to bypass processing layers that would cause it to dissipate and disrupt learning, in effect allowing the model to learn appropriate processing depth for a task. The latest trend in the design of deep neural networks involves the inclusion of such residual blocks in networks that are not only deeper but also wider, as in the ResNetXT (Xie et al. 2017). However, as of 2017, accuracy improvements on ILSVRC have started to hit a ceiling, and the challenge is considered by many to be saturated.

By now, deep networks have literally taken over the entire field of computer vision. In face recognition, CNNs have already been shown to perform at the level of professional facial forensic experts, as well as untrained super-recognizers from the general public (i.e., people who have significantly higher than average face recognition ability; see Phillips et al. 2018). CNNs have also achieved significant successes in the categorization of natural animal and plant species, mimicking tasks that require human expertise. Currently, the largest such database is called the iNaturalist plant and animal classification data set (Van Horn et al. 2018), which includes nearly 1 million photos from 5,089 taxa and 13 superclasses taken in the wild by amateur photographers. The

accuracy of these CNNs is claimed to be on par with experts and well beyond that of naive observers, with a current top-5 accuracy of 87.5% (or approximately 5,000 times better than chance).

Pose tracking and action recognition in videos are two additional areas where deep learning has pushed the state of the art. ResNets have already found an exciting application to body-pose tracking in animal biomedical research (Mathis et al. 2018). Human action recognition has been lagging somewhat behind image categorization because of a lack of a comparably large, labeled data set. There are now claims that action recognition in videos is about to catch up, given the recent availability of action data sets of magnitude comparable to ILSVRC (Hara et al. 2018). The recently released Kinetics data set (Kay et al. 2017) contains approximately 650,000 video clips covering 700 human action classes and at least 600 clips per class.

One of the main ideas behind extensions of CNNs from the processing of images to the processing of videos is to inflate spatial filters from 2D (spatial) convolutions to 3D (spatiotemporal) convolutions (Carreira & Zisserman 2017). Another key concept involves the combination of shape and motion information computed by these 2D and 3D convolutional architectures (Carreira & Zisserman 2017, Simonyan & Zisserman 2014). As in image categorization, these core ideas were identified by computational neuroscientists decades ago based on physiological and anatomical considerations (Adelson & Bergen 1985, Giese & Poggio 2003, Heeger et al. 1996, Jhuang et al. 2007). However, as in image categorization, progress has reached impressive levels, with accuracy rates now well over 80% (top-5) accuracy on the kinematics data set with 400 diverse human classes.

Overall, the recent gains in accuracy achieved with CNNs would have been hard to imagine just a decade ago, and today, it is nearly impossible to find an area of computer vision that is not dominated by deep learning. What is perhaps even more surprising is how many of these gains fueled by computer science and machine learning advances have turned into parallel gains in visual neuroscience, which I discuss in the next section.

3. DEEP LEARNING: THE GOOD

The history of artificial neural networks is rooted in biology, but two important developments have led to the recent gains observed in computer vision: the end-to-end learning of visual representations optimized for large-scale visual recognition tasks and the successful training of increasingly deep visual architectures. These advances have, in turn, prompted neuroscientists to explore the ability of these improved vision models to account for the properties of neurons in our visual cortex (for reviews, see Kriegeskorte 2015, Yamins & DiCarlo 2016).

3.1. Deep Learning in the Visual System

Early computational neuroscience work started with AlexNet (Krizhevsky et al. 2012) and ZFnet (Zeiler & Fergus 2014), which were shown to improve the fit to neural data in intermediate and higher areas of the ventral stream of the visual cortex compared to earlier computational models trained without explicit supervision (Cadieu et al. 2014, Khaligh-Razavi & Kriegeskorte 2014). More recent work includes deeper networks such as VGG (Simonyan & Zisserman 2015), but these models did not significantly improve the goodness-of-fit beyond the shallower AlexNet and ZFnet (Abbasi-Asl et al. 2018, Kalfas et al. 2017). A recent study (Cadena et al. 2019) has shown that intermediate VGG layers provide a better fit to V1 monkey electrophysiology data compared to simpler linear-nonlinear models. Work by Hong et al. (2016) has also shown that multiple image properties beyond object categories (e.g., object position, 3D size and pose) remain relatively well encoded in higher processing stages in both neural and CNN representations. This provides

further evidence that both visual hierarchies are able to learn visual representations that are invariant to task-irrelevant transformations while maintaining information for categorical-orthogonal properties.

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

One of the main challenges associated with the fitting of computational models such as the CNNs discussed above arises because of the large dimensionality of the space of model parameters and the comparatively small amount of neural data available to estimate those parameters. As a result, all the above-mentioned studies had to use feature activations from ILSVRC-optimized networks to fit linear models to neural data, thus reducing the model parameter space to a more manageable size given the number of samples available from these neuroscience experiments. Recently, however, researchers have started to take into account the anatomy of the visual system to reduce the dimensionality of the space of model parameters, which has, in turn, allowed them to train models end-to-end without the need for pretraining.

For instance, it is possible to exploit the fact that neurons with overlapping receptive fields receive a common pool of inputs to more effectively fit model parameters to multiple V1 neurons simultaneously, making it possible to learn a common feature space from which one can then linearly predict the activity of each individual neuron (Antolík et al. 2016). Similarly, researchers have shown that it is possible to leverage the modular organization of the retina to fit the spiking activity of a population of retinal ganglion cells—markedly outperforming previous models (Batty et al. 2016, Klindt et al. 2017, Maheswaranathan et al. 2018, McIntosh et al. 2016).

3.2. In Silico Electrophysiology

Above, I highlight that our progress in deep learning has led to computational models with better fit to neuroscience data, but the interaction between computer scientists and brain scientists has not been unidirectional. Interestingly, several of the common methods used to try to characterize the visual representations learned by deep networks are indeed reminiscent of classic methods used in neurophysiology studies, from searching for a preferred stimulus (Gross et al. 1972) to methods for feature simplification (Kobatake & Tanaka 1994) and reverse correlation (Jones & Palmer 1987).

Feature visualization methods aim to identify preferred stimuli for individual network units (**Figure 2*a–d***) (for an overview, see Olah et al. 2017). Some of the simplest methods include the

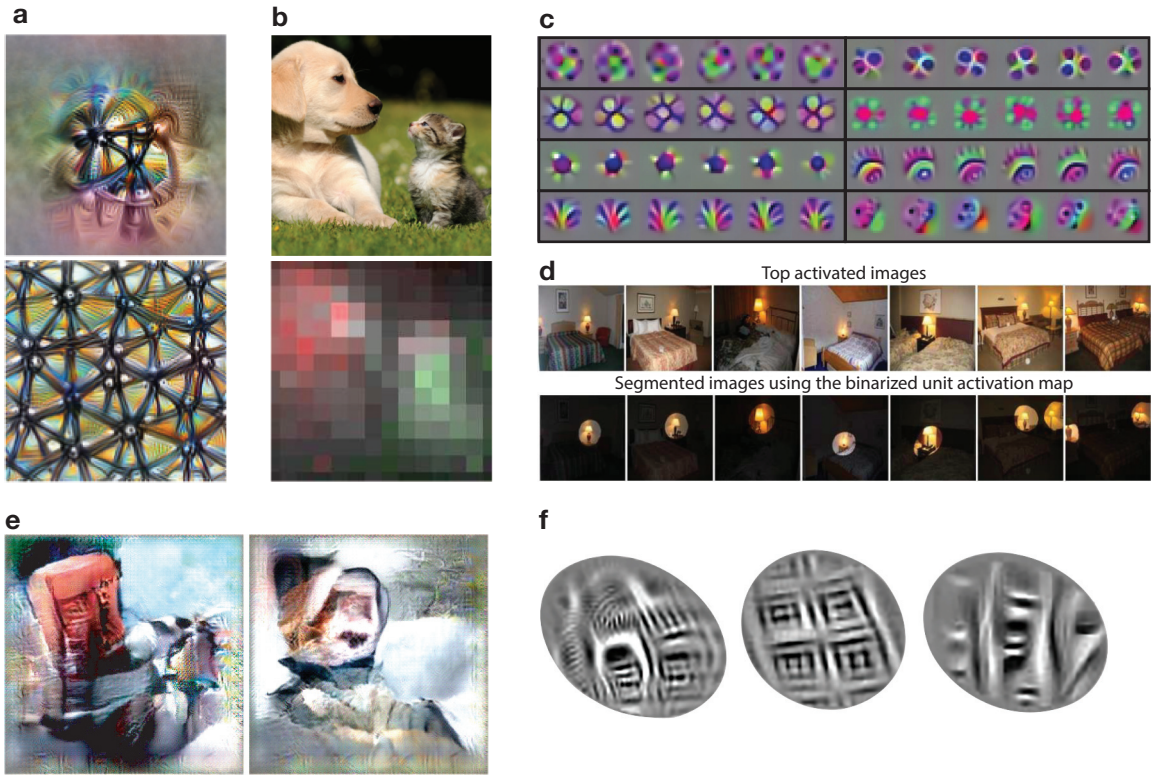

Figure 2

Representative methods for understanding deep neural network representations and interpreting their decisions. (*a*) Feature visualization methods aim to identify the preferred stimulus of units in a neural network. (*b*) Attribution methods aim to identify the part of an image responsible for a network activating in a particular way. Shown are pixels that contribute to the network decision for "Labrador retriever" (*red*) and "tiger cat" (*green*). Panels *a* and *b* adapted from Olah et al. (2017) under Creative Commons Attribution CC-BY 4.0. (*c*) Invariant subspaces. Representative samples showing texture-like and shape-like detectors in VGG [layer conv3_1 from Cadena et al. (2018)]. Panel adapted with permission from the authors. (*d*) Network dissection (Zhou et al. 2019). The method identifies image regions that selectively activate a particular network unit and assess the overlap with the associated semantic interpretation using manual image annotations. Image credit: Bolei Zhou (adapted with permission). (*e*) Stimulus generated using a generative adversarial network (GAN) combined with a genetic algorithm to probe the selectivity of inferotemporal cortex neurons (Ponce et al. 2019). Image credit: Carlos Ponce (adapted with permission). (*f*) Sample stimuli generated using the feature visualization method of panel *a* on a network fitted to V4 neural responses. These responses are in turn shown to activate V4 units maximally. Shown are one of the features generated to maximally activate neural sites 13, 16 and 17. Image credit: Pouya Bashivan (adapted with permission).

activation maximization approach, which aims to identify the top visual stimulus (Erhan et al. 2009) or the average (over-the-top N) visual stimuli (Zhou et al. 2014) that maximize an individual unit response as their preferred stimuli. So-called attribution methods (for a review, see Olah et al. 2018) have also been proposed to help determine salient image features responsible for the network activating in a particular way, also using variants of backpropagation (**Figure 2*b***). Another related line of work (Cadena et al. 2018) has focused on visualizing what image transformations units in a CNN are invariant to (**Figure 2*c***).

More complex optimization-based feature visualization methods, referred to as deep dreaming, have also been proposed that use variants of backpropagation (see the sidebar titled Image Synthesis) to synthesize novel images that are optimal to drive specific units of a network

IMAGE SYNTHESIS

Starting from a noise image, it is possible to use gradient descent methods, including backpropagation, to iteratively synthesize a new image that shares the same representation (e.g., the same unit activations in certain layers of the network) with a reference image up to some tolerance (for one of the first successful approaches prior to deep learning, see Portilla & Simoncelli 2000). Unlike when training the network, where each iteration of the backpropagation algorithm yields an update of the network weights, synthesis and related transfer methods do not update the weights and update the input image instead because the gradient is computed with respect to the input image and not the network weights. The art in image synthesis is to discover interesting differentiable image parametrizations to produce appealing images (for an overview, see Mordvintsev et al. 2018). Another recent method for image synthesis that is gaining in popularity is the generative adversarial network (GAN) (Goodfellow et al. 2014). GANs use two neural networks: A generator tries to synthesize content (say, an image), and a discriminator tries to discriminate between real images and synthetic ones. This can lead to impressive artificially generated images, including the synthetic faces and artistic paintings shown in **Figure 3**. The field has now boomed with methods developed for the more general class of problems known as image-to-image translation to learn to convert images from one domain to another (as in converting day scenes into night scenes, summer pictures into winter ones, etc.) (Zhu et al. 2017).

(Olah et al. 2017). The use of neural networks for synthesizing novel images has bloomed in recent years, with broad applications in computational photography and even art production (**Figure 3**). These image synthesis methods have even been used to control cortical populations of neurons. Bashivan et al. (2019) fitted CNNs to V4 neural responses and used these models to construct images designed to either broadly activate large populations of neurons or selectively activate one population while keeping another unchanged. Recent work has also focused on the application of a new breed of architectures called generative adversarial networks (GANs; see the sidebar titled Image Synthesis) to find optimal stimuli for neurons (Ponce et al. 2019) or to help reconstruct visual stimuli from fMRI activity (Güçlütürk et al. 2017; but for reconstruction approaches that use more standard CNNs, see also Horikawa & Kamitani 2017a,b).

In addition to visualization and attribution methods, several selectivity measures have been proposed to better characterize the visual representations learned by CNNs. The network dissection method (Zhou et al. 2019) tries to quantify the interpretability of individual units by estimating the receptive field location of individual units and then considering the purity of their selectivity for a variety of visual attributes (e.g., color, object category, material properties) using a large database of manually annotated images (see **Figure 2*d***). The class-conditional mean activity selectivity (Morcos et al. 2018) inspired by category indexes used in monkey neurophysiology studies (Freedman et al. 2001) measures the contrast between the highest class-conditional mean activity and the mean activity across all other classes. Overall, one of the most striking findings is that, following ILSVRC training, a significant number of units become highly selective to object categories, including people and animals (Agrawal et al. 2014, Zhou et al. 2019). In addition, it has also been shown that the representations learned by CNNs, including the top-most layers, are able to maintain a surprising amount of image information, including the viewpoint of a face (Parde et al. 2017) and the position, size, and pose of an object (Hong et al. 2016).

Overall, a growing body of literature suggests that depth and supervision in CNNs, which have led to major gains in accuracy in machine vision applications, have supported the concomitant learning, allowing for the development of visual representations that are more consistent with

Figure 3

Art and image synthesis with deep neural networks. (*a*) Style transfer with convolutional neural networks (CNNs). The content of an image (*top*, *first panel*) is combined with the style of three distinct paintings (*top*, *second*, *third*, and *fourth panels*) to synthesize a new artistic image (*bottom row*) using a neural network (Gatys et al. 2017). Image credit: Matthias Bethge (adapted with permission). (*b*) The *Portrait of Edmond Belamy* produced by a generative adversarial network (GAN) and sold by Christie's for $432,500 in October 2018. Reproduced with permission from the copyright holder. Sotheby's Contemporary Art Day Auction held in March 2019 also featured a machine installation that used neural networks to generate an infinite stream of portraits. Mainstream artists including Refik Anadol, Trevor Paglen, and Jason Salavon have started to incorporate neural networks as a part of their artistic process. (*c*) Latest improvements in style transfer by leveraging attentional mechanism to produce transfers that respect the semantic content of the original image (Park & Lee 2019). Image credit: Dae Y. Park and Kwang H. Lee (adapted with permission). (*d*) Synthetic images generated by a large generative adversarial network named BigGAN (Brock et al. 2019). Adapted with permission from the authors.

those found in our visual cortex. Concurrently, novel methods have been developed, inspired in part by electrophysiology studies, to try to better characterize these learned visual representations. At the same time, a growing body of literature is suggesting that modern artificial neural networks leverage visual recognition strategies that differ from those used by their biological cousins, which I discuss in the next section.

4. DEEP LEARNING: THE BAD

Surprisingly, the success of CNNs in explaining neural data has not turned into consistent improvements in explaining human judgments and behavioral decisions. An important human benchmark for CNNs includes psychophysics data derived from rapid (speeded) categorization tasks because these tasks do not require attention and engage predominantly feedforward neural processes akin to those implemented in CNNs (see Section 2.1). Indeed, at a coarse level, CNNs' sensitivity to image transformations appears to be largely consistent with human rapid categorization data (Kheradpisheh et al. 2016a,b). At the same time, decisions derived from CNNs at the individual image level appear to differ significantly from those made by human observers (Eberhardt et al. 2016, Rajalingham et al. 2018). A systematic comparison between predictions derived from individual network layers and those derived from humans revealed that the correlation peaks for convolutional layers (Eberhardt et al. 2016), suggesting that the decision processes involved in human behavioral judgments are not well modeled by the final (fully connected) layers of a CNN.

Consistent with this interpretation, a recent object naming study (Devereux et al. 2018) has shown that a network model of semantics, explicitly trained to learn a mapping from visual features onto object semantic attributes, was better able to explain fMRI activation patterns in higher visual areas compared to both convolutional and fully connected layers of a CNN. In addition, it has been shown that neural signals in intermediate areas of the ventral stream of the visual cortex are predictive of nonhuman primate accuracy and reaction times during rapid visual categorization (Cauchoix et al. 2016). It is thus possible that human participants rely on visual representations of lesser complexity than those learned by CNNs. How then is the visual cortex able to improve its accuracy when more time is allowed for visual recognition? One hypothesis is that longer response times allow recurrent processes to take place, and greater processing depth is achieved through time rather than static depth, as in CNNs.

Beyond rapid visual categorization, several studies have highlighted qualitative differences between the visual strategy used by CNNs and that used by human observers. One important feature of the primate visual system is its sensitivity to shape information starting early in life. Initial studies have shown that, indeed, ILSVRC-optimized CNNs exhibit a sensitivity to shape features that is much improved compared to earlier computational models of object recognition and qualitatively consistent with neural and behavioral studies (Kubilius et al. 2016, Ritter et al. 2017), including the preference to categorize objects according to shape rather than color that is found in young word learners (Landau et al. 1988).

At the same time, studies have also found systematic differences between CNN and human judgments for the perception of object silhouettes (Kubilius et al. 2016, Pramod & Arun 2016). One possible explanation for this discrepancy is that, as hypothesized in cognitive psychology (Biederman 1987), more explicit structural representations may be needed to close the gap. This assumption was directly validated by Erdogan & Jacobs (2017), who showed that a Bayesian model of shape inference with an object-centered coordinate system captures human observers' similarity judgments significantly better than do CNN models. In addition, a very recent study has shown that CNNs may have access to some shape information in the form of local edge relations, but they do not seem to have access to global object shapes (Baker et al. 2018). Another study has shown that CNNs do not seem to perceive illusory contours, suggesting that they deal with partial occlusions using a strategy that likely differs from the amodal completion mechanisms used by human observers (Kellman et al. 2017). Two studies have also found that modern CNNs underperformed simpler (older) models of early visual processing (inspired by the physiology of the early visual system) for explaining human perceptual sensitivity to image perturbations and to temporal changes in image sequences (Berardino et al. 2017, Hénaff et al. 2019).

In a similar vein, Ullman et al. (2016) found that, when presented with small object crops, human participants depend critically on the inclusion of a key diagnostic image feature to recognize an object. In contrast, CNNs fail to exhibit the same all-or-nothing dependence on key visual features during object recognition. A subsequent study by Linsley et al. (2017) tried to compare more directly the visual representations learned by CNNs and those used by human observers. Using Clicktionary, a collaborative web-based game that they developed to identify diagnostic visual features for human object recognition, they were able to directly compare attribution maps derived from representative CNNs (for an overview of these methods, see Section 3.2) with importance maps derived from human observers. The analysis revealed that CNNs and humans favor dissimilar visual features during object categorization. It is likely that these differences arise because of a lack of explicit mechanisms for perceptual grouping and figure-ground segmentation in CNNs, which are known to play a key role in the development of our visual system (Johnson 2001, Ostrovsky et al. 2009). In the absence of figure-ground mechanisms, CNNs are compelled to associate foreground objects and their context as single perceptual units. Consistent with this idea, it has been shown that CNNs do not generalize well to atypical scenes, such as when objects are presented outside of their usual context and in the presence of clutter and occluders (Rosenfeld et al. 2018b, Saleh et al. 2016, Tang et al. 2018, J. Wang et al. 2018). Follow-up work by Linsley et al. (2019) demonstrated that it is, however, possible to cue CNNs to attend to image regions that human observers deem to be important for object recognition. Such co-training leads to neural architectures that generalize better and that learn visual representations that are more consistent with those used by humans.

Beyond visual categorization, several early studies have successfully used CNNs to predict human typicality ratings (Lake et al. 2015) and memorability (Dubey et al. 2015) for natural object images. More recent work has shown that important features of human judgments are missing from CNN representations (Jozwik et al. 2017). However, the difference between the two systems may be more quantitative rather than qualitative, as it has been shown that a simple linear transformation of these representations [analog to the concept of dimensional attention in cognitive psychology (Nosofsky 1987)] leads to substantial improvements in the goodness-of-fit of these models (Peterson et al. 2018). At the same time, a recent study aiming to target higher-level concepts in similarity ratings demonstrated that, despite the authors' best efforts, none of the tested CNNs were able to reproduce the image matching produced by human observers (Rosenfeld et al. 2018a). The authors attributed this discrepancy to a variety of factors that are known to affect human similarity judgments, including abstraction (as in abstracting a doorway for a mountain passageway) and context dependence (as in flexibly ignoring color cues or pose to match shape).

Overall, while initial studies highlighted the substantial similarities between visual representations learned by artificial and biological neural networks, a growing body of literature is starting to demonstrate systematic differences between machine and human visual recognition judgments.

5. DEEP LEARNING: THE UGLY

While cognitive psychologists have found systematic differences between human and machine judgments, computer scientists have discovered several critical limitations of modern deep neural network architectures—adding further support to the idea that CNNs might constitute at best an incomplete model of human vision.

5.1. Adversarial Images

An adversarial image is an image that has been slightly altered to fool a deep neural network by getting the network to misclassify an image that would have been otherwise correctly classified

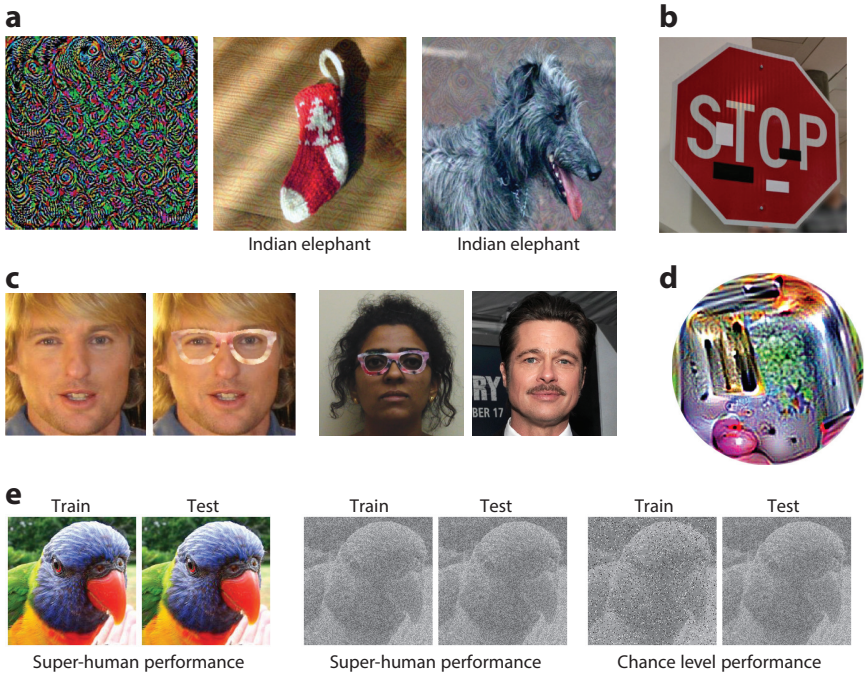

Figure 4

Deep neural networks are sensitive to noise perturbations. (*a*) Universal adversarial attacks. State-of-the-art deep neural networks are vulnerable to universal (image-agnostic) adversarial attacks such that a very small noise mask (shown magnified for GoogLeNet) applied to images would lead to misclassification with high probability (actual adversarial samples with noise overlaid shown). Panel adapted from Moosavi-Dezfooli et al. (2017) with permission. (*b–d*) Sample adversarial physical attacks (*b*) conducted on a traffic sign system using stickers masked as graffiti, (*c*) using printed glasses to fool a face verification system, and (*d*) using a sticker to cause a deep neural network to see bananas everywhere. Panel *b* adapted from Eykholt et al. (2018) with permission. Panel *c* adapted from Sharif et al. (2019) with permission. Panel *d* adapted from Brown et al. (2017) with permission. (*e*) Deep neural networks do not generalize to unseen noise patterns, as shown for a network trained with additive Gaussian noise and tested on salt-and-pepper noise, which appear nearly identical to human observers. Panel adapted from Geirhos et al. (2018) with permission.

with high confidence. That such adversarial attacks are possible is hardly surprising, since an image can always be manipulated to change the appearance of an individual or an object. What is surprising is that, in practice, artificial neural networks can be fooled by minute manipulations that are barely visible to a human eye (see **Figure 4**). While such failure cases seem quite hard to interpret, recent work has shown that, surprisingly, human observers are capable of correctly anticipating a CNN's classification output over such adversarial images—even for images described as "totally unrecognizable to human eyes" (Zhou & Firestone 2019). Originally described as an odd CNN behavior (Szegedy et al. 2013), adversarial images have become increasingly well studied, with researchers simultaneously working toward the design of methods for attacking networks and methods of defending networks against these attacks.

For many common images, finding such perturbation appears to be strikingly easy—requiring a single step in the direction of the gradient (see the sidebar titled Computing with Gradients) to produce adversarial examples that are transferable from one network to another trained for the same problem but with a different architecture. Adversarial attacks originally began as pixel perturbations optimized for individual images, but in recent years, attacks have become increasingly

pernicious: Adversarial attacks can now be carried out using universal (image-agnostic) adversarial perturbations, which when applied to almost any image will lead to misclassification (Moosavi-Dezfooli et al. 2017), and they can be carried out in the real world—posing a significant risk for biometric applications and for autonomous vehicles (for a review, see Gilmer et al. 2018). Representative examples are shown in **Figure 4**, with adversarial examples shown in the domain of object and traffic sign recognition using physical stickers (Brown et al. 2017, Eykholt et al. 2018), as well as face recognition with physical 3D glasses worn by impersonating individuals (Sharif et al. 2019).

While it is possible to carry out adversarial attacks using noise perturbations so small that they are imperceptible to human observers (Brendel et al. 2017), it is also possible to fool networks using completely unrecognizable images (Nguyen et al. 2015). Both of these extremes highlight significant differences between the visual strategies leveraged by biological and artificial neural networks.

5.2. The Need for Generalization

In addition to their sensitivity to adversarial noise, as discussed in the previous section, a growing body of literature is showing that current deep neural networks are severely limited in their ability to generalize to novel stimuli beyond the image data set used to train them.

Beyond adversarial noise, a small amount of filtering in the Fourier domain, which would lead to very small behavioral differences with human observers, ends up significantly affecting the recognition capabilities of deep neural networks (Jo & Bengio 2017). Similarly, while CNNs can learn to handle specific kinds of noise when noisy images are incorporated in their training data sets, they are unable to generalize to unseen noise conditions—even when the noise patterns are similar (see **Figure 4**) (Geirhos et al. 2018). The lack of generalization of CNNs also extends to objects embedded in novel contexts or occluded by out-of-context objects (Rosenfeld et al. 2018b, Saleh et al. 2016, J. Wang et al. 2018). Even translation invariance, the property that motivated the original architectural design behind CNNs (Fukushima 1980, LeCun et al. 1998, Riesenhuber & Poggio 1999), has been shown to be surprisingly limited, in contrast to common belief (Azulay & Weiss 2018). In fact, the deeper is the network, the less invariance to translation it exhibits, with a shift by only a few pixels greatly affecting the network output.

Even more problematic, state-of-the-art neural architectures are capable of achieving a near-perfect classification accuracy when trained on ILSVRC with randomly shuffled image labels (C. Zhang et al. 2016). In other words, they are able to memorize associations between images and random class labels. After all, this is not all that surprising: Given the very large number of free parameters in these networks (on the order of tens of millions), they technically have the capacity to memorize random associations between images and class labels, even for a large data set such as ILSVRC (1.5 million images). Still, such an experiment opens up the possibility that the high level of accuracy achieved by state-of-the-art neural networks could be due to their ability to memorize images combined with the limited image variability and known biases present in benchmark computer vision data sets (Torralba & Efros 2011).

Indeed, a subsequent study demonstrated that state-of-the-art CNNs trained on one image data set do not generalize well to novel data from a novel test set generated by the authors, despite the fact that the training and test sets exhibit near-identical statistics (Recht et al. 2018, 2019). This highlights the possibility that much of the recent progress achieved in image categorization could be due in part to the fact that researchers have been building on each other's work (since 2010 in the case of ILSVRC) and have thus broken one of the most important rules in machine learning: not to use the test set to tune any free parameters to maximize accuracy (i.e., by tuning hyperparameters such as the number of layers).

OBJECT DETECTION AND LOCALIZATION

Unlike in the image categorization tasks described in Section 2, where entire images are associated with a single class label, object detection and localization involve the detection of one or multiple objects and the ability to draw a bounding box around them. Region-based approaches are extensions of the CNNs that achieve state-of-the-art results for object detection and localization. Rather than exhaustively scanning an image across all positions and scales and classifying every such window, as in early computer vision algorithms, the basic idea behind region-based approaches is to first run a generic object detector over the image, as in the R-CNN (Girshick et al. 2014), to bring down the number of windows to be classified (called the region proposals) to a reasonable number (from millions to a few thousands). These windows are then classified by a CNN to yield a class label for each bounding box (including an option to reject the bounding box as containing any of the objects of interest). The approach was improved in a series of papers from the fast R-CNN (Girshick 2015) to the faster R-CNN (Ren et al. 2015) and the region-based fully convolutional networks (R-FCN) (Dai et al. 2016) by sharing convolutional layers between the region proposal stage and the detection and localization stages—thus allowing the training of a single efficient CNN for the entire system. A popular architecture includes YOLO (Redmon & Farhadi 2017), which can run with near state-of-the-art accuracy but in real time at 30 fps. Rather than considering region proposals, YOLO looks at the whole image at test time, so its predictions can potentially be informed by context. More recent developments include the RetinaNet (Lin et al. 2019), which is both fast and accurate, and the Mask-RCNN (He et al. 2019), which learns to predict an object mask that is more detailed than the coarse bounding box returned by other architectures. It is worth noting that region-based approaches for detection and localization detect object-like shapes in an image exhaustively before attempting to recognize these objects. This does not seem consistent with the strategy used by human observers, who have been shown during visual searches, unlike deep networks, to often miss targets that have an atypical size relative to the surrounding objects in the scene—presumably highlighting a strategy used to rapidly ignore distractors during visual search tasks (Eckstein et al. 2017).

A major issue for the field is that modern neural networks have become truly humongous. In the past few years, the depth of state-of-the-art architectures has been rapidly increasing, with some of the deepest CNNs now containing approximately 60 million parameters. This trend is continuing with recent CNN extensions for visual recognition tasks beyond image categorization (**Figure 1**), including object localization (see the sidebar titled Object Detection and Localization), semantic segmentation, depth estimation, and dense pose estimation (see the sidebar titled Image Segmentation and Other (Per-Pixel) Dense Image Prediction Tasks). As the number of free parameters continues to exceed the number of samples available for training, these neural architectures maintain an ability to solve complex recognition tasks via brute-force memorization of feature templates.

Visual reasoning tasks, for instance, offer a vivid example of this issue. The Synthetic Visual Reasoning Test (SVRT) is a collection of 23 binary classification problems in which opposing classes differ based on whether images obey an abstract rule (Fleuret et al. 2011). For the most part, the visual relations depicted in the 23 SVRT problems are rather intuitive and obvious to a human observer (Fleuret et al. 2011). Such an ability is by no means limited to human perception. In a striking example from the work of Martinho & Kacelnik (2016), newborn ducklings were shown to imprint on an abstract concept of sameness from a single training example at birth. Yet modern CNNs appear to be unable to learn such same–different tasks even after being presented with millions of training examples (Ellis et al. 2015, Kim et al. 2018, Stabinger et al. 2016). To make matters worse, the issue has been overshadowed by the recent successes obtained with CNN extensions called relational networks (RNs) (Santoro et al. 2017) on seemingly challenging visual question-answering benchmarks (Johnson et al. 2017). To directly test the RN ability to learn

IMAGE SEGMENTATION AND OTHER (PER-PIXEL) DENSE IMAGE PREDICTION TASKS

Dense image prediction tasks have become increasingly popular in recent years (**Figure 1**). Such tasks are referred to as dense predictions because a neural network has to produce entire prediction maps for each individual image (as opposed to a single class label in image categorization), as in boundary detection or local surface orientation prediction, saliency, semantic segmentation, human part detection, and object detection to name just a few (additional classic tasks not shown in **Figure 1** include dense optical flow and depth prediction). Also shown in **Figure 1** is a task referred to as dense (3D body) pose prediction (Güler et al. 2018).

How can one go from architectures that return a single class label for the entire image, as in a standard CNN, to an architecture capable of labeling every pixel in an image? Most of these methods follow the same general architecture, which combines a CNN, referred to as the encoder because it turns the input image into a downsampled visual representation, with a decoder, followed by a readout processing stage that upsamples the visual representation and turns it back into a map the size of the original image. The fully convolutional network architecture was one of the first successful approaches (Long et al. 2015, Maninis et al. 2018, Papandreou et al. 2015, Xie & Tu 2017). Decoders in these models use 1×1 convolutions to combine upsampled activity maps from several layers of the encoder. There also exist featured candidate operations for incorporating contextual information into local convolutional activities, including dilated convolutions, which involve applying a stride to the kernel before convolving the input (Long et al. 2015, Yu & Koltun 2016) and have helped improve performance in multiple computer vision problems, from denoising (T. Wang et al. 2017) and semantic segmentation (Chen et al. 2018, Hamaguchi et al. 2017) to colorization tasks (R. Zhang et al. 2016). Another approach to dense prediction uses skip connections to connect specific layers of a model's encoder to its decoder. This approach was first described by Long et al. (2015) as a method for more effectively merging coarse-layer information into a model's decoder and later extended in the U-Net (Ronneberger et al. 2015). Unpooling models eliminate the need for feature map upsampling by routing decoded activities to the locations of the winning max-pooling units derived from the encoder. Unpooling is also a leading approach for a variety of dense per-pixel prediction tasks, including segmentation, which is exemplified by SegNet (Badrinarayanan et al. 2017). This model has a decoder that mirrors its encoder, with unpooling operations replacing its pooling.

to solve same–different visual tasks, Kim et al. (2018) trained an RN using millions of examples sampled from one set of objects; the network was incapable of recognizing the relation when tested on a novel set of objects.

It is becoming increasingly clear that the ability of modern neural networks to generalize beyond training data is indeed quite limited, and that artificial networks do not yet truly support neuronal representations and processes that naturally allow for flexible, rich reasoning about, e.g., objects and their relations in visual scenes.

6. CONCLUSIONS AND PATH FORWARD

Progress in the area of machine vision has been significant—arguably beyond what any researcher in the field would have predicted only a decade ago. The variety and difficulty of modern visual challenges that are driving the field are impressive—from recognizing thousands of objects, to identifying individual faces among millions of distractors, to parsing complex natural scenes and estimating the poses of articulated bodies. The accuracy of current machine vision systems on these tasks is such that it is tempting to conclude that vision has been solved. Furthermore, these machine vision innovations have been met with concurrent improvements in the ability of modern

deep neural networks to account for neural data from the visual cortex. As a result, deep convolutional neural networks have become de facto models of primate vision.

While progress has been significant, many obstacles remain: These network architectures can be easily fooled by small noise perturbations that are barely perceptible to human observers, and it is becoming increasingly clear that they exhibit relatively limited generalization capabilities beyond the data used to train them. Current artificial vision systems have only really been successful in visual categorization tasks in scenarios that are limited to preattentive recognition in primates. They are not yet capable of abstraction and exhibit a limited ability to learn abstract relational rules beyond rote memorization.

Overall, it is unlikely that all the necessary building blocks of vision have been identified. Current trends in automatic machine learning, where large-scale optimization algorithms automatically identify optimal configurations of existing building blocks (Jaderberg et al. 2017), are likely to yield further quantitative gains in accuracy, but they are unlikely to produce the leap needed to yield truly intelligent seeing machines. Such qualitative improvements are likely to require additional computations beyond those employed in current architectures.

Fortunately, motivated in some cases by neuroscience considerations, computer scientists have already started to increase the size of the toolbox of computations available to deep neural networks, from attentional mechanisms, to more complex elementary units such as capsules (Hinton et al. 2018, Sabour et al. 2017), to memory units with persistent states, all the way to content- and location-addressable memories.

Attention, for instance, has become the subject of intensive research within the deep learning community, and some of the most recent gains achieved in visual recognition can be attributed to the inclusion of attention mechanisms in CNNs. While biology is sometimes mentioned as a source of inspiration (Biparva & Tsotsos 2017, Chen et al. 2017, F. Wang et al. 2017), the attentional mechanisms that have been considered remain rather limited in comparison to the rich and diverse array of processes used by the primate visual system. Yet initial work has already shown that using human supervision to cue these architectures to attend to image regions that are deemed diagnostic for human visual categorization is showing great promise (Linsley et al. 2019).

More generally, it is likely that the depth of state-of-the-art CNN architectures already exceeds the depth of our visual system. The main assumption is thus that longer response times allow for greater processing depth through time via recurrent circuits. The fastest behavioral responses would thus reflect preattentive feedforward processing compatible with CNN architectures. However, longer response times would likely reflect the involvement of re-entrant and other top-down signals when more time is available for visual processing.

Indeed, computer vision work is starting to demonstrate the benefits of systems that integrate recurrent mechanisms: Feedback signals can provide prior information in predictive coding models (Han et al. 2018, Lotter et al. 2016, O'Reilly et al. 2017); help to solve contour tracing, incremental grouping, and other visual recognition tasks that require learning long-range statistical dependencies (Brosch et al. 2015, Linsley et al. 2018); and help improve recognition performance (George et al. 2017, Sabour et al. 2017, Spoerer et al. 2017, Tang et al. 2018). There is also currently a substantial increase of excitement for deep generative models for both language and vision, which includes the generative adversarial networks (GANs) discussed in Section 3.2, and related models such as variational auto-encoders (Kingma & Welling 2014) and deep Boltzmann machines (Salakhutdinov & Hinton 2009), which combine the complementary strengths of latent variable models and deep neural networks. Such deep generative architectures constitute rich models of visual scenes (Eslami et al. 2018) and provide an intriguing new perspective on the role of feedback in the visual cortex (Lee et al. 1998).

Similarly, our brains rely on multiple memory systems (Hassabis et al. 2017), and recent work has shown the benefit of incorporating a memory into artificial neural networks to allow better one-shot generalization (Vinyals et al. 2016). Popular recurrent neural networks such as Long Short Term Memory (LSTM) units (Hochreiter & Schmidhuber 1997) and Gated Recurrent Units (GRUs) (Cho et al. 2014) include gate mechanisms that control how the information is maintained over time and are reminiscent of working memory. However, in both LSTMs and GRUs, the sequence controller and memory storage are intermingled, whereas in our brains, the two are separate (Tonegawa et al. 2015). Researchers have started to address these issues with the development of a differential neural controller that learns to attend to and to read and write from an external memory matrix to solve complex reasoning problems (Graves et al. 2016).

These extended neural network models, and those to come, are likely to help scientists to better understand how living brains work and help engineers create better thinking machines.

SUMMARY POINTS

1. The computational building blocks of modern deep neural networks were first identified by neuroscientists. However, recent progress has been driven entirely by computer scientists via the development of additional computational blocks.
2. The visual representations derived from modern neural networks optimized for large-scale visual categorization challenges initially yielded concurrent improvements in our ability to explain visual representations found in the visual cortex.
3. However, the ability of current neural network architectures to explain experimental data seems to have hit a ceiling. Recent work has highlighted key differences between the visual strategies used by modern deep neural networks and those used by human observers.
4. The ability of neural network architectures to generalize beyond training data appears significantly more limited than was initially thought. Their ability to solve more general visual reasoning tasks beyond visual categorization appears limited.
5. These limitations suggest the need to identify additional neural computations beyond those already implemented in CNNs.
6. Recent innovations in deep learning that are at least in part motivated by neuroscience considerations and extend the feedforward neural architectures described in this review via attentional, mnemonic, and other feedback mechanisms offer exciting future directions to try to address current shortcomings in the development of machine vision algorithms capable of truly intelligent visual behavior.

DISCLOSURE STATEMENT

T.S. serves as a scientific advisor for Vium, Inc., which may potentially benefit from the research results.

ACKNOWLEDGMENTS

I would like to thank Rufin VanRullen, Gabriel Kreiman, and Drew Linsley for their helpful feedback on the manuscript. I was funded in part by a Defense Advanced Research Projects Agency

(DARPA) young faculty award and director's award (grant N66001-14-1-4037) and a National Science Foundation (NSF) early career award (grant IIS-1252951). Additional funding was provided by the Office of Naval Research (ONR) (grant N00014-19-1-2029) and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI).

LITERATURE CITED

Abadi M, Barham P, Chen J, Chen Z, Davis A, et al. 2016. TensorFlow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, pp. 265–83. Berkeley, CA: USENIX Assoc.

Abbasi-Asl R, Chen Y, Bloniarz A, Oliver M. 2018. The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. bioRxiv 465534

Adelson EH, Bergen JR. 1985. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A.* 2(2):284–99

Agrawal P, Girshick R, Malik J. 2014. Analyzing the performance of multilayer neural networks for object recognition. In *Computer Vision: ECCV 2014*, pp. 329–44. Berlin: Springer

Antolík J, Hofer SB, Bednar JA, Mrsic-Flogel TD. 2016. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLOS Comput. Biol.* 12(6):e1004927

Azulay A, Weiss Y. 2018. Why do deep convolutional networks generalize so poorly to small image transformations? arXiv:1805.12177 [cs.CV]

Badrinarayanan V, Kendall A, Cipolla R. 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(12):2481–95

Baker N, Lu H, Erlikhman G, Kellman PJ. 2018. Deep convolutional networks do not classify based on global object shape. *PLOS Comput. Biol.* 14(12):e1006613

Baldi P. 2018. Deep learning in biomedical data science. *Annu. Rev. Biomed. Data Sci.* 1:181–205

Bashivan P, Kar K, DiCarlo JJ. 2019. Neural population control via deep image synthesis. *Science* 364(6439):eaav9436

Batty E, Merel J, Brackbill N, Heitman A, Sher A, et al. 2016. *Multilayer recurrent network models of primate retinal ganglion cell responses*. Paper presented at the 5th International Conference on Learning Representations (ICLR), Toulon, France

Berardino A, Laparra V, Ballé J, Simoncelli E. 2017. Eigen-distortions of hierarchical representations. In *Advances in Neural Information Processing Systems 30*, ed. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, pp. 3530–39. Red Hook, NY: Curran Assoc.

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

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94(2):115–47

Biparva M, Tsotsos J. 2017. *STNet: selective tuning of convolutional networks for object localization*. Paper presented at the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Venice, Italy

Brendel W, Rauber J, Bethge M. 2017. *Decision-based adversarial attacks: reliable attacks against black-box machine learning models*. Paper presented at the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada

Brock A, Donahue J, Simonvan K. 2019. *Large scale GAN training for high fidelity natural image synthesis*. Paper presented at the 7th International Conference on Learning Representations (ICLR), New Orleans, LA

Brosch T, Neumann H, Roelfsema PR. 2015. Reinforcement learning of linking and tracing contours in recurrent neural networks. *PLOS Comput. Biol.* 11(10):e1004489

Brown TB, Mané D, Roy A, Abadi M, Gilmer J. 2017. Adversarial patch. arXiv:1712.09665 [cs.CV]

Cadena SA, Weis MA, Gatys LA, Bethge M, Ecker AS. 2018. Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Computer Vision: ECCV 2018*, ed. V Ferrari, M Hebert, C Sminchisescu, Y Weiss, pp. 225–40. Berlin: Springer

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

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, et al. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Comput. Biol.* 10(12):e1003963

Carreira J, Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–33. Piscataway, NJ: IEEE

Cauchoix M, Crouzet SM, Fize D, Serre T. 2016. Fast ventral stream neural activity enables rapid visual categorization. *NeuroImage* 125:280–90

Chen L, Zhang H, Xiao J, Nie L, Shao J, et al. 2017. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6298–306. Piscataway, NJ: IEEE

Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4):834–48

Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, et al. 2014. *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. Paper presented at the 2014 Conference on Empirical Methods in Natural Language, Doha, Qatar

Cichy RM, Khosla A, Pantazis D, Oliva A. 2017. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage* 153:346–58

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

Cireşan D, Meier U, Masci J, Schmidhuber J. 2012. Multi-column deep neural network for traffic sign classification. *Neural Netw.* 32:333–38

Crick F. 1989. The recent excitement about neural networks. *Nature* 337(6203):129–32

Dai J, Li Y, He K, Sun J. 2016. R-FCN: object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 379–87. Red Hook, NY: Curran Assoc.

Devereux BJ, Clarke A, Tyler LK. 2018. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Sci. Rep.* 8(1):10636

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

Doersch C, Zisserman A. 2017. Multi-task self-supervised visual learning. In *Proceedings of the 2017 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–79. Piscataway, NJ: IEEE

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, et al. 2014. DeCAF: a deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 647–55. La Jolla, CA: Int. Conf. Machine Learn.

Dubey R, Peterson J, Khosla A, Yang M-H, Ghanem B. 2015. What makes an object memorable? In *Proceedings of the 2015 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1089–97. Piscataway, NJ: IEEE

Eberhardt S, Cader JG, Serre T. 2016. How deep is the feature analysis underlying rapid visual categorization? In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 1100–8. Red Hook, NY: Curran Assoc.

Eckstein MP, Koehler K, Welbourne LE, Akbas E. 2017. Humans, but not deep neural networks, often miss giant targets in scenes. *Curr. Biol.* 27(18):2827–32.e3

Eickenberg M, Gramfort A, Varoquaux G, Thirion B. 2017. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* 152:184–94

Ellis K, Solar-Lezama A, Tenenbaum J. 2015. Unsupervised learning by program synthesis. In *Advances in Neural Information Processing Systems 28*, ed. C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett, pp. 973–81. Red Hook, NY: Curran Assoc.

Erdogan G, Jacobs RA. 2017. Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychol. Rev.* 124(6):740–61

Erhan D, Bengio Y, Courville A, Vincent P. 2009. *Visualizing higher-layer features of a deep network*. Tech. Rep., Univ. Montreal

Eslami SMA, Jimenez Rezende D, Besse F, Viola F, Morcos AS, et al. 2018. Neural scene representation and rendering. *Science* 360(6394):1204–10

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, et al. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1625–34. Piscataway, NJ: IEEE

Fleuret F, Li T, Dubout C, Wampler EK, Yantis S, Geman D. 2011. Comparing machines and humans on a visual categorization test. *PNAS* 108(43):17621–25

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. 2001. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291:312–16

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

Gatys LA, Ecker AS, Bethge M. 2017. Texture and art with deep neural networks. *Curr. Opin. Neurobiol.* 46:178–86

Geirhos R, Temme J, Rauber J, Schutt M, Bethge M, Wichmann FA. 2018. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems 31*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 7549–61. Red Hook, NY: Curran Assoc.

George D, Lehrach W, Kansky K, Lázaro-Gredilla M, Laan C, et al. 2017. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science* 358(6368):eaag2612

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

Gilmer J, Adams RP, Goodfellow I, Andersen D, Dahl GE. 2018. Motivating the rules of the game for adversarial example research. arXiv:1807.06732 [cs.LG]

Girshick R. 2015. Fast R-CNN. In *Proceedings of the 2015 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1440–48. Piscataway, NJ: IEEE

Girshick R, Donahue J, Darrell T, Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–87. Piscataway, NJ: IEEE

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 2672–80. Red Hook, NY: Curran Assoc.

Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature* 538(7626):471–76

Greene MR, Hansen BC. 2018. Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Comput. Biol.* 14(7):e1006327

Gross CG, Rocha-Miranda CE, Bender DB. 1972. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35(1):96–111

Güçlü U, van Gerven MAJ. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35(27):10005–14

Güçlü U, van Gerven MAJ. 2017. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage* 145(Pt. B):329–36

Güçlütürk Y, Güçlü U, Seeliger K, Bosch S, van Lier R, van Gerven MAJ. 2017. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems 30*, ed. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, et al., pp. 4246–57. Red Hook, NY: Curran Assoc.

Güler RA, Neverova N, Kokkinos I. 2018. DensePose: dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–306. Piscataway, NJ: IEEE

Hamaguchi R, Fujita A, Nemoto K, Imaizumi T, Hikosaka S. 2017. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision*, art. 00162. Piscataway, NJ: IEEE

Han K, Wen H, Zhang Y, Fu D, Culurciello E, Liu Z. 2018. Deep predictive coding network with local recurrent processing for object recognition. In *Advances in Neural Information Processing Systems 31*, pp. 9221–33. Red Hook, NY: Curran Assoc.

Hara K, Kataoka H, Satoh Y. 2018. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18–22. Piscataway, NJ: IEEE

Hassabis D, Kumaran D, Summerfield C, Botvinick M. 2017. Neuroscience-inspired artificial intelligence. *Neuron* 95(2):245–58

He K, Gkioxari G, Dollar P, Girshick R. 2019. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* In press

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. arXiv:1512.03385 [cs.CV]

Heeger DJ, Simoncelli EP, Movshon JA. 1996. Computational models of cortical visual processing. *PNAS* 93(2):623–27

Hénaff OJ, Goris RLT, Simoncelli EP. 2019. Perceptual straightening of natural videos. *Nat. Neurosci.* 22(6):984–91

Hinton GE, Sabour S, Frosst N. 2018. *Matrix capsules with EM routing*. Paper presented at the 6th International Conference on Learning Representations, Vancouver, Canada

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–80

Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ. 2016. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19(4):613–22

Horikawa T, Kamitani Y. 2017a. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* 8:15037

Horikawa T, Kamitani Y. 2017b. Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Front. Comput. Neurosci.* 11:4

Hu J, Shen L, Albanie S, Sun G, Wu E. 2019. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* In press. **https://doi.org/10.1109/TPAMI.2019.2913372**

Jaderberg M, Dalibard V, Osindero S, Czarnecki WM, Donahue J, et al. 2017. Population based training of neural networks. arXiv:1711.09846 [cs.LG]

Jhuang H, Serre T, Wolf L, Poggio T. 2007. A biologically inspired system for action recognition. In *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8. Piscataway, NJ: IEEE

Jo J, Bengio Y. 2017. Measuring the tendency of CNNs to learn surface statistical regularities. arXiv:1711.11561 [cs.LG]

Johnson J, Hariharan B, van der Maaten L, Li F-F, Zitnick CL, Girshick R. 2017. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–97. Piscataway, NJ: IEEE

Johnson SP. 2001. Visual development in human infants: binding features, surfaces, and objects. *Vis. Cogn.* 8(3–5):565–78

Jones JP, Palmer LA. 1987. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* 58(6):1187–211

Jozwik KM, Kriegeskorte N, Storrs KR, Mur M. 2017. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* 8:1726

Kalfas I, Kumar S, Vogels R. 2017. Shape selectivity of middle superior temporal sulcus body patch neurons. *eNeuro* 4(3):ENEURO.0113-17.2017

Karras T, Aila T, Laine S, Lehtinen J. 2018. *Progressive growing of GANs for improved quality, stability, and variation*. Paper presented at the International Conference on Learning Representations, Vancouver, BC, Canada

Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, et al. 2017. The Kinetics Human Action Video dataset. arXiv:1705.06950 [cs.CV]

Kellman P, Baker N, Erlikhman G, Lu H. 2017. Classification images reveal that deep learning networks fail to perceive illusory contours. *J. Vis.* 17(10):569

Khaligh-Razavi S-M, Henriksson L, Kay K, Kriegeskorte N. 2017. Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* 76(Pt. B):184–97

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

Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T. 2016a. Deep networks can resemble human feed-forward vision in invariant object recognition. *Sci. Rep.* 6:32672

Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T. 2016b. Humans and deep networks largely agree on which kinds of variation make object recognition harder. *Front. Comput. Neurosci.* 10:92

Kim JK, Ricci M, Serre T. 2018. Not-So-CLEVR: learning same-different relations strains feedforward neural networks. *Interface Focus* 8(4):20180011

Kingma DP, Welling M. 2014. *Auto-encoding variational Bayes*. Paper presented at the International Conference on Learning Representations, Banff, Canada

Klindt D, Ecker AS, Euler T, Bethge M. 2017. Neural system identification for large populations separating "what" and "where." In *Advances in Neural Information Processing Systems 30*, ed. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, et al., pp. 3506–16. Red Hook, NY: Curran Assoc.

Kobatake E, Tanaka K. 1994. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* 71(3):856–67

Kokkinos I. 2017. UberNet: training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5454–63. Piscataway, NJ: IEEE

Kriegeskorte N. 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1:417–46

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, ed. F Pereira, CJC Burges, L Bottou, KQ Weinberger, pp. 1097–105. Red Hook, NY: Curran Assoc

Kubilius J, Bracci S, Op de Beeck HP. 2016. Deep neural networks as a computational model for human shape sensitivity. *PLOS Comput. Biol.* 12(4):e1004896

Lake BM, Zaremba W, Fergus R, Gureckis TM. 2015. Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pp. 1243–48. Seattle, WA: Cogn. Sci. Soc.

Landau B, Smith LB, Jones SS. 1988. The importance of shape in early lexical learning. *Cogn. Dev.* 3(3):299–321

LeCun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE.* 86(11):2278–324

Lee K, Zung J, Li P, Jain V, Seung HS. 2017. Superhuman accuracy on the SNEMI3D connectomics challenge. arXiv:1706.00120 [cs.CV]

Lee TS, Mumford D, Romero R, Lamme VA. 1998. The role of the primary visual cortex in higher level vision. *Vis. Res.* 38(15–16):2429–54

Li N, Dicarlo JJ. 2012. Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *J. Neurosci.* 32(19):6611–20

Lin T-Y, Goyal P, Girshick R, He K, Dollar P. 2019. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* In press. **https://doi.org/10.1109/TPAMI.2018.2858826**

Linsley D, Eberhardt S, Sharma T, Gupta P, Serre T. 2017. What are the visual features underlying human versus machine vision? In *Proceedings of the IEEE ICCV Workshop on the Mutual Benefit of Cognitive and Computer Vision*, pp. 2706–14. Piscataway, NJ: IEEE

Linsley D, Kim JK, Veerabadran V, Windolf C, Serre T. 2018. Learning long-range spatial dependencies with horizontal gated recurrent units. In *Advances in Neural Information Processing Systems 31*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 152–64. Red Hook, NY: Curran Assoc.

Linsley D, Schiebler D, Eberhardt S, Serre T. 2019. *Learning what and where to attend*. Paper presented at the Seventh International Conference on Learning Representations, New Orleans, LA

Logothetis NK, Pauls J, Poggio T. 1995. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5:552–63

Long J, Shelhamer E, Darrell T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–40. Piscataway, NJ: IEEE

Lotter W, Kreiman G, Cox D. 2016. *Deep predictive coding networks for video prediction and unsupervised learning*. Paper presented at the 5th International Conference on Learning Representations, Toulon, France

Maheswaranathan N, Kastner DB, Baccus SA, Ganguli S. 2018. Inferring hidden structure in multilayered neural circuits. *PLOS Comput. Biol.* 14(8):e1006291

Maninis K-K, Pont-Tuset J, Arbelaez P, Van Gool L. 2018. Convolutional oriented boundaries: from image segmentation to high-level tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4):819–33

Martinho A, Kacelnik A. 2016. Ducklings imprint on the relational concept of "same or different." *Science* 353(6296):286–88

Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, et al. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* 21(9):1281–89

McIntosh L, Maheswaranathan N, Nayebi A, Ganguli S, Baccus S. 2016. Deep learning models of the retinal response to natural scenes. In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 1369–77. Red Hook, NY: Curran Assoc.

Miconi T, Clune J, Stanley KO. 2018. Differentiable plasticity: training plastic neural networks with backpropagation. In *Proceedings of the 35th International Conference on Machine Learning (ICML2018)*, pp. 3556–65. La Jolla, CA: Int. Conf. Machine Learn.

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–33

Moldwin T, Segev I. 2018. Perceptron learning and classification in a modeled cortical pyramidal cell. bioRxiv 464826

Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, Frossard P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1765–73. Piscataway, NJ: IEEE

Morcos AS, Barrett DGT, Rabinowitz NC, Botvinick M. 2018. *On the importance of single directions for generalization*. Paper presented at the 6th International Conference on Learning Representations, Vancouver, Canada

Mordvintsev A, Pezzotti N, Schubert L, Olah C. 2018. Differentiable image parameterizations. *Distill*, July 25. **https://distill.pub/2018/differentiable-parameterizations/**

Nguyen A, Yosinski J, Clune J. 2015. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–36. Piscataway, NJ: IEEE

Nosofsky RM. 1987. Attention and learning processes in the identification and categorization of integral stimuli. *J. Exp. Psychol. Learn. Mem. Cogn.* 13(1):87–108

Olah C, Mordvintsev A, Schubert L. 2017. Feature visualization. *Distill*, Nov. 7. **https://distill.pub/2017/feature-visualization/**

Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, et al. 2018. The building blocks of interpretability. *Distill*, March 6. **https://distill.pub/2018/building-blocks/**

Oquab M, Bottou L, Laptev I, Sivic J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717–24. Piscataway, NJ: IEEE

O'Reilly RC, Wyatte DR, Rohrlich J. 2017. Deep predictive learning: a comprehensive model of three visual streams. arXiv:1709.04654 [q-bio.NC]

Ostrovsky Y, Meyers E, Ganesh S, Mathur U, Sinha P. 2009. Visual parsing after recovery from blindness. *Psychol. Sci.* 20(12):1484–91

Papandreou G, Kokkinos I, Savalle P-A. 2015. Modeling local and global deformations in Deep Learning: epitomic convolution, Multiple Instance Learning, and sliding window detection. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 390–99. Piscataway, NJ: IEEE

Parde CJ, Castillo C, Hill MQ, Colon YI, Sankaranarayanan S, et al. 2017. Face and image representation in deep CNN features. In *Proceedings of the 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 673–80. Piscataway, NJ: IEEE

Park DY, Lee KH. 2019. Arbitrary style transfer with style-attentional. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5873–81. Piscataway, NJ: IEEE

Paszke A, Gross S, Chintala S, Chanan G, Yang E, et al. 2017. *Automatic differentiation in PyTorch*. Paper presented at the Neural Information Processing Systems Autodiff Workshop, Long Beach, CA

Peterson JC, Abbott JT, Griffiths TL. 2018. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cogn. Sci.* 42(8):2648–69

Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, et al. 2018. Face recognition accuracy of forensic examiners, super-recognizers, and face recognition algorithms. *PNAS* 115(24):6171–76

Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* 177(4):999–1009.e10

Portilla J, Simoncelli EP. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40(1):49–71

Pramod RT, Arun SP. 2016. Do computational models differ systematically from human object perception? In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1601–9. Piscataway, NJ: IEEE

Radovic A, Williams M, Rousseau D, Kagan M, Bonacorsi D, et al. 2018. Machine learning at the energy and intensity frontiers of particle physics. *Nature* 560(7716):41–48

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. 2018. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* 38(33):7255–69

Recht B, Roelofs R, Schmidt L, Shankar V. 2018. Do CIFAR-10 classifiers generalize to CIFAR-10? arXiv:1806.00451 [cs.LG]

Recht B, Roelofs R, Schmidt L, Shankar V. 2019. Do ImageNet classifiers generalize to ImageNet? arXiv:1902.10811 [cs.CV]

Redmon J, Farhadi A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–25. Piscataway, NJ: IEEE

Ren S, He K, Girshick R, Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, ed. C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett, pp. 91–99. Red Hook, NY: Curran Assoc.

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

Ritter S, Barrett DGT, Santoro A, Botvinick MM. 2017. Cognitive psychology for deep neural networks: a shape bias case study. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2940–49. La Jolla, CA: Int. Conf. Machine Learn.

Ronneberger O, Fischer P, Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2015*, pp. 234–41. Berlin: Springer

Rosenfeld A, Solbach MD, Tsotsos JK. 2018a. Totally looks like—how humans compare, compared to machines. arXiv:1803.01485 [cs.CV]

Rosenfeld A, Zemel R, Tsotsos JK. 2018b. The elephant in the room. arXiv:1808.03305 [cs.CV]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115:211–52

Sabour S, Frosst N, Hinton GE. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems 30*, ed. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, pp. 3859–69. Red Hook, NY: Curran Assoc.

Salakhutdinov R, Hinton G. 2009. Deep Boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, ed. D van Dyk, M Welling, pp. 448–55. New York: ACM

Saleh B, Elgammal AM, Feldman J, Farhadi A. 2016. Toward a taxonomy and computational models of abnormalities in images. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 3588–96. Palo Alto, CA: Assoc. Adv. Artif. Intell.

Santoro A, Raposo D, Barrett DGT, Malinowski M, Pascanu R, et al. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30*, ed. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, pp. 4974–83. Red Hook, NY: Curran Assoc.

Scellier B, Bengio Y. 2017. Equilibrium propagation: bridging the gap between energy-based models and backpropagation. *Front. Comput. Neurosci.* 11:24

Segler MHS, Preuss M, Waller MP. 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555(7698):604–10

Serre T. 2015. Hierarchical models of the visual system. In *Encyclopedia of Computational Neuroscience*, ed. D Jaeger, R Jung, pp. 1309–18. Berlin: Springer

Sharif M, Bhagavatula S, Bauer L, Reiter MK. 2019. A general framework for adversarial examples with objectives. *ACM Trans. Priv. Secur.* 22(3):6

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–89

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362(6419):1140–44

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–59

Simonyan K, Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27*, ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 568–76. Red Hook, NY: Curran Assoc.

Simonyan K, Zisserman A. 2015. *Very deep convolutional networks for large-scale image recognition*. Paper presented at the 3rd International Conference on Learning Representations, San Diego, CA

Spoerer CJ, McClure P, Kriegeskorte N. 2017. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* 8:1551

Stabinger S, Rodríguez-Sánchez A, Piater J. 2016. 25 years of CNNs: Can we compare to human abstraction capabilities? In *Artificial Neural Networks and Machine Learning: ICANN 2016*, pp. 380–87. Berlin: Springer

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, et al. 2015. Going deeper with convolutions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9. Piscataway, NJ: IEEE

Szegedy C, Zaremba W, Sutskever I. 2013. Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV]

Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, et al. 2018. Recurrent computations for visual pattern completion. *PNAS* 115(35):8835–40

Tonegawa S, Pignatelli M, Roy DS, Ryan TJ. 2015. Memory engram storage and retrieval. *Curr. Opin. Neurobiol.* 35:101–9

Torralba A, Efros AA. 2011. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–28. Piscataway, NJ: IEEE

Tosic I, Frossard P. 2011. Dictionary learning. *IEEE Signal Process. Mag.* 28(2):27–38

Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4489–97. Piscataway, NJ: IEEE

Ullman S, Assif L, Fetaya E, Harari D. 2016. Atoms of recognition in human and computer vision. *PNAS* 113(10):2744–49

Van Horn G, Mac Aodha O, Song Y, Cui Y, Sun C, et al. 2018. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–78. Piscataway, NJ: IEEE

Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 3630–38. Red Hook, NY: Curran Assoc.

Wainberg M, Merico D, Delong A, Frey BJ. 2018. Deep learning in biomedicine. *Nat. Biotechnol.* 36(9):829–38

Wang F, Jiang M, Qian C, Yang S, Li C, et al. 2017. Residual attention network for image classification. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–58. Piscataway, NJ: IEEE

Wang J, Zhang Z, Xie C, Zhou Y, Premachandran V, et al. 2018. Visual concepts and compositional voting. *Ann. Math. Sci. Appl.* 3(1):151–88

Wang T, Sun M, Hu K. 2017. Dilated deep residual network for image denoising. In *Proceedings of the 29th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1272–79. Piscataway, NJ: IEEE

Whittington JCR, Bogacz R. 2017. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Comput.* 29(5):1229–62

Xie S, Girshick R, Dollár P, Tu Z, He K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–95. Piscataway, NJ: IEEE

Xie S, Tu Z. 2017. Holistically-nested edge detection. *Int. J. Comput. Vis.* 125(1):3–18

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

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

Yu F, Koltun V. 2016. *Multi-scale context aggregation by dilated convolutions*. Paper presented at the 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico

Zeiler MD, Fergus R. 2014. Visualizing and understanding convolutional networks. In *Computer Vision: ECCV 2014*, pp. 818–33. Berlin: Springer

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. 2016. *Understanding deep learning requires rethinking generalization*. Paper presented at the 4th International Conference on Learning Representations, Toulon, France, April 24–26

Zhang R, Isola P, Efros AA. 2016. Colorful image colorization. In *Computer Vision: ECCV 2016*, pp. 649–66. Berlin: Springer

Zhou B, Bau D, Oliva A, Torralba A. 2019. Interpreting deep visual representations via network dissection. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(9):2131–45

Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. 2014. Learning deep features for scene recognition using Places Database. In *Advances in Neural Information Processing Systems 27*, ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 487–95. Red Hook, NY: Curran Assoc.

Zhou Z, Firestone C. 2019. Humans can decipher adversarial images. *Nat. Commun.* 10(1):1334

Zhu J-Y, Park T, Isola P, Efros AA. 2017. *Unpaired image-to-image translation using cycle-consistent adversarial networks*. Paper presented at the IEEE International Conference on Computer Vision (ICCV), Venice, Italy