Deep Learning Networks and Visual Perception

Grace W. Lindsay, Gatsby Computational Neuroscience Unit Sainsbury Wellcome Centre University College London London UK, W1T 4JG and Thomas Serre, Department of Cognitive Linguistic & Psychological Sciences Carney Institute for Brain Sciences Brown University Providence RI, 02912

https://doi.org/10.1093/acrefore/9780190236557.013.841

Published online: 31 August 2021

Summary

Deep learning is an approach to artificial intelligence (AI) centered on the training of deep artificial neural networks to perform complex tasks. Since the early 21st century, this approach has led to record-breaking advances in AI, allowing computers to solve complex board games, video games, natural language-processing tasks, and vision problems. Neuroscientists and psychologists have also utilized these networks as models of biological information processing to understand language, motor control, cognition, audition, and—most commonly—vision. Specifically, early feedforward network architectures were inspired by visual neuroscience and are used to model neural activity and human behavior. They also provide useful representations of the perceptual space of images. The extent to which these models match data, however, depends on the methods used to characterize and compare them. The limitations of these feedforward neural networks to account for, for example, simple visual reasoning tasks, suggests that feedback mechanisms may be necessary to solve visual recognition tasks beyond image categorization.

Keywords: perceptron, vision, convolutional neural network, perceptual spaces, attention, image categorization, feedforward network, recurrent networks

Subjects: Psychology and Other Disciplines

Introduction

Artificial neural networks, as their name suggests, are computational models inspired by the properties of biological neurons. The earliest artificial neural networks were introduced in 1943 by McCulloch and Pitts; however, 21st-century incarnations are more similar to the Perceptron algorithm created by Rosenblatt in 1958 (McCulloch & Pitts, 1943; Rosenblatt, 1958). In the following, artificial neurons will be referred to as *units*. Their activity will be described with a number (either a scalar or an integer).

In the most basic form of the Perceptron, a layer of input units connects to an output unit, the activity of which indicates whether the input belongs to a certain category or not (figure 1a). A different numerical value, known as a weight, connects each input unit to the output. To classify a given input, the activity of each input unit is multiplied by its respective weight and the sum is compared to a threshold. If this sum is above the threshold, the output unit is *active* (indicating the category is present); otherwise, it is *inactive* (indicating category is absent). In the Perceptron, the step function (also called staircase or threshold function) is used but in modern neural networks many different so-called “activation” functions or simply “nonlinearities” can be used (e.g., rectified linear unit or ReLu: $g(x) = x^+ = \max(x, 0)$ or the softplus function known as SmoothReLU: $g(x) = \ln(1 + e^x)$.

In this way, the neural network computes a function *F(x)* of its input stimulus, *x*. In the case of a simple binary classification, *F(x)* is a step function that can take only two values (say 0/1 or -1/1). However, this simple categorization problem can be easily extended to multi-class classification problems or to predicting a continuous output, as in regression problems (see Serre, 2016 for additional information). Most visual recognition tasks can be cast as categorization problems, from face detection to object recognition and even more general cases of visual reasoning.

In deep learning, the neural networks used are “deep,” meaning they are composed of not just a single layer of processing as in the Perceptron, but many layers—possibly hundreds of them. The layers that sit in between the input and output layers are called the *hidden* layers. Similar to the Perceptron, at each layer in a deep neural network, the weighted input from the layer below is combined and passed through each unit’s nonlinearity to create the activity at that layer. This allows the information that passes through the network—such as the pixel values of an image *x*—to be transformed several times until a useful output *F(x)*—for example, a label for the object in the image—is produced. In this sense, deep neural networks perform “function approximation”—that is, they combine several simple operations to approximate more complex real-world functions between inputs and outputs.

A common way of training neural networks to successfully learn such a transformation is through a particular kind of “optimization” called *supervised learning*. Supervised learning, as opposed to *unsupervised learning*, requires a set of training data that includes the desired output for each input (e.g., a set of natural images and the corresponding “ground-truth” label, i.e., what object category is in that image). Through knowledge of these pairings, the weights in the network are updated in order to make the network as a whole better at associating each input with its correct output.

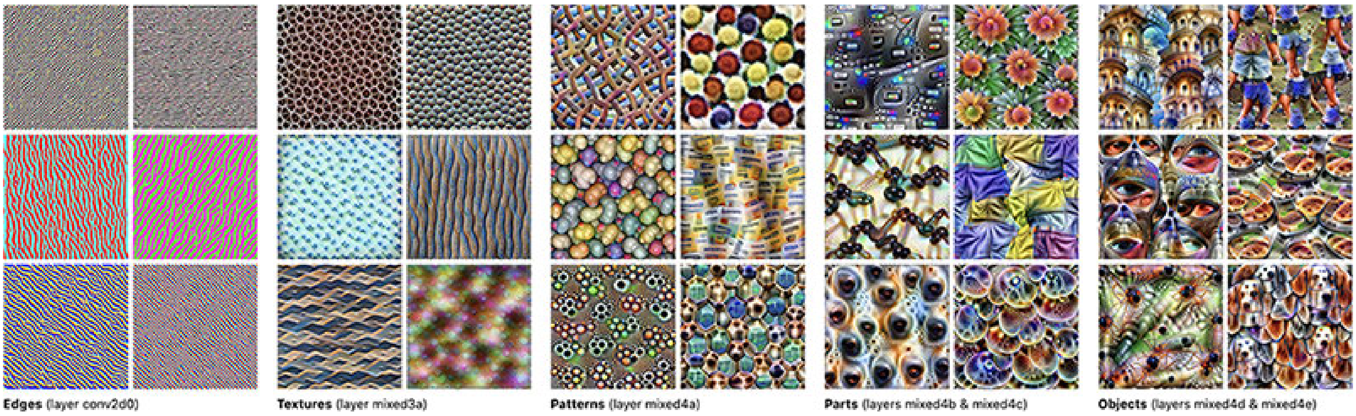

Figure 1. The relationship between artificial neural networks and visual circuitry.

Note: (a) The Perceptron architecture. The output unit of the Perceptron is said to be active if the sum of its inputs (black) times their respective weights (purple) is greater than 0 (example on right; total input = 0.8) and inactive otherwise (example on left; total input = -0.5). (b) Convolution. The convolutional layers of a CNN expand on the basic Perceptron architecture. Here, the weights are specified as a 2D grid. An image can also be specified as a 2D grid of pixel values (here dark pixels take a value of 0 and light pixels a value of 1). The grid of weights, or kernel, is multiplied by the individual pixel values at each location in the input image. The sum of these products (followed by a nonlinearity such as the ReLu, see Introduction for details) produces an output unit for each location. Here, the sum is calculated as (-1*0) + (-1*0) + (-1*0) + (1*1) + (1*1) + (1*1) + (-1*0) + (-1*0) + (-1*0) = 3. Calculating these sums of products is known mathematically as the convolution of the input pattern with the array of weights. Together these output units computed at all image locations comprise a feature map. (c) Hubel and Wiesel model. Receptive fields of neurons in the lateral geniculate nucleus (LGN) have a center-surround structure wherein they respond best to a spot of light surrounded by dark (bottom). (Note that the LGN response can itself be modeled as a convolution with a particular kernel known as a difference of Gaussians or DoG.) LGN cells with nearby receptive fields aligned along an axis provide input to a simple cell in the primary visual cortex (top). This gives the simple cell its orientation preference. A complex cell (not shown) would selectively pool over afferent simple cells with the same preferred orientation but different positions to build tolerance to translation of the preferred stimulus. Neuron drawing adapted from scidraw.io (under the creative commons license CC-BY). (d) A sample CNN architecture. After the image is convolved with several different filters (here, three) the same number of feature maps are created. The pooling layer downsamples these feature maps in the spatial dimensions. This activity is then passed into a final, non-convolutional layer with the same number of units as classes in the classification problem. When trained with supervised learning, this network can learn to classify images. In a deeper CNN, more layers of convolution and pooling would be added between the image and the classifier.

By far the most popular algorithm to train a deep neural network is the backpropagation algorithm (Rumelhart et al., 1986). The backpropagation algorithm is an application of the “chain rule” in calculus. It provides a means of calculating, for any weight in the network, how that weight should change in order to make the performance of the network better according to some metric. To apply backpropagation, one thus needs to define a “loss” or “objective” function which explicitly defines this metric. In supervised learning, the loss function is frequently a measure of how far the output of the network is from the correct output for any given input. For instance, for image classification, the network’s classification error calculated from a training set of images can be used as a loss. Historically, backpropagation has been viewed as biologically implausible. Nonetheless, the successes of deep neural networks have reinvigorated interest in backpropagation and several biologically plausible approximations of backpropagation have been proposed (see Lillicrap et al., 2020 for a review, including a discussion of a plausible neural substrate).

While the exact values of the weights between neurons are determined by the learning algorithm, the overall architecture of the network is usually pre-specified by the model builder. Architecture choices include the number of layers, the number of neurons per layer, the dimensions of the kernels used, and the properties of operations such as pooling, and so on. Neural networks for which the only type of connections is bottom up are called feedforward networks and those that allow for feedback signals are called recurrent networks. Recurrent networks allow re-entrant signals from connections within a layer (through "horizontal" or "lateral" connections) and/or top down from a later processing stage onto an earlier one.

One specific style of architecture that is widely used for visual processing is the convolutional neural network (CNN) (LeCun et al., 1989). The name was derived from a mathematical operation known as a "convolution" (figure 1b) but these networks also share a number of basic properties with the mammalian visual system. These properties were studied in cats by Hubel and Wiesel in the late 1950s (Hubel & Wiesel, 1959).

Hubel and Wiesel identified two main cell types in the primary visual cortex, which they labeled "simple" and "complex." Simple cells have a localized receptive field—the region of visual space where a stimulus must be presented in order to elicit a response from the neuron. And they have a preferred orientation: the angle at which a bar in their receptive field should be tilted to elicit a maximal response. This property is known as "selectivity." Complex cells also have preferred orientations but their receptive fields are larger (about twice as large as that of simple cells). This means that the same oriented bar can be placed in a wider range of locations and still elicit a response from a complex cell. This is known as the property of "invariance" or "tolerance" to position.

Hubel and Wiesel surmised that simple and complex cells both end up with these properties based on the kind of inputs they get from other cells. Mechanistically, simple-cell-like responses can be obtained by pooling the activity of a small set of cells tuned to spots of lights aligned along a preferred axis of orientation (figure 1c). Such neurons are found in the brain regions that provide input to simple cells, such as the lateral geniculate nuclear and layer IV of the primary visual cortex. Similarly, at the next stage, position tolerance at the complex cell level could be obtained by pooling over afferent simple cells with the same preferred (e.g., vertical) orientation but slightly different positions (see Ricci & Serre, 2020 for details).

These properties of selectivity and invariance are mimicked in a CNN through two main computations: convolution and pooling. Convolutions require kernels, which are a two- (or possibly three- or more) dimensional grids of weights. To convolve an image with a kernel means applying that set of weights (followed by a nonlinearity) separately at each location in the image (figure 1b). This is akin to "replicating" neurons with a particular selectivity (e.g., vertical orientation) at all locations in the visual field (Ullman & Soloviev, 1999). This creates a layer of artificial neurons with the same selectivity and the individual activity for each neuron is only influenced by a small region of image space—that is, they each have a localized receptive field. The exact values of the weights determine their selectivity (i.e., the features—e.g., what kind of visual patterns or lines—these artificial units respond to). Typically, many different kernels (on the order of tens to hundreds) will be applied to an image in parallel, resulting in several different "feature maps" within a processing layer (see Ricci & Serre (2020) for a more in-depth discussion).

In the visual cortex, neurons are arranged retinotopically—that is, nearby neurons have similar spatial receptive fields. At each retinotopic location there are different neurons with different selectivity or preferred features (e.g., spanning all possible orientations). Feature maps thus mimic this property of having selectivity to the same feature replicated across the visual field.

The pooling operation in a CNN typically comes after a convolution and does not involve weights (although technically it could). Rather, the activity of a unit in the pooling layer is calculated as the maximal activity of the units in a small region of the feature map below it. This results in a pooling layer with the same number of feature maps as the convolutional layer that came before it. In this way, units in the pooling layer have larger receptive fields but similar feature selectivity as those in the convolutional layer, much like complex cells relative to simple cells. Pooling can also be combined with a "stride" to downsample the spatial dimensions of the image representation. The "stride" determines the amount of overlap between complex cells' receptive fields—controlling how finely or coarsely activity at that layer is being sampled. The existence of such a max-pooling operation was originally hypothesized by Riesenhuber and Poggio (1999) in their HMAX computational model of object recognition and later confirmed experimentally (Lampl et al., 2004).

To create a deep neural network out of these operations, convolutions and pooling are simply stacked such that the output of the first pooling layer becomes the input for a second convolutional layer (see a relatively shallow CNN in figure 3d and an example of a deep CNN architecture in figure 6a). Such stacking of simple and complex cell-like computations was shown to be useful for producing an artificial visual system by Fukushima in a precursor to CNNs known as the Neocognitron (Fukushima, 1980). In modern-day CNNs, these stacks of convolutional and pooling layers are simply referred to as "convolutional" layers. These convolutional layers are typically followed by a classification stage—so-called "fully connected" layers whose purpose is to learn a mapping from the stimulus representations afforded by the convolutional layers with class labels. These fully connected layers are typically multilayer extensions of the Perceptron called a "multilayer perceptron" or MLP. In older models of vision in computational neuroscience, convolutional and fully connected layers used to be trained independently. In modern-day CNNs, the entire architecture is optimized jointly for a particular visual task and is said to be trained "end-to-end" (Serre, 2019). Hence, the task used to optimize the network is likely to shape the selectivity of units all the way down to early convolutional layers.

In 2012, a CNN composed of five layers of convolutions and pooling followed by three non-convolutional "fully connected" layers, achieved 62.5% accuracy on the ImageNet classification challenge (Deng et al., 2009; Krizhevsky et al., 2012). This challenge involves labeling 224 x 224 pixel color images as belonging to one of a thousand different object categories (chance level

<0.1%). The second-place winner of this challenge, which did not use a deep CNN, achieved only around 54% accuracy. These results kicked off a new age of extensive use of artificial neural networks, particularly CNNs, in computer vision and artificial intelligence (AI) more generally.

Not long after, computational neuroscientists began investigating how well these networks can inform understanding of the neural representations found in living brains. For an overview of the use of artificial neural networks to model the brain, see “Deep Neural Networks in Computational Neuroscience.”

Modeling Neural Activity with Deep Neural Networks

The similarities between the basic features of the mammalian visual cortex and the structure of CNNs make these networks a candidate model of visual processing in animals. Computational neuroscientists frequently use Marr’s levels to understand the functioning of a brain region or system and compare it to models (Marr, 1982). According to Marr, the “computational” level asks the purpose of the system; in this case, one could consider the purpose of part of the visual c to be to recognize and/or categorize objects in images. The “algorithmic” level then describes through which computational steps this goal is achieved and the “implementation” level identifies the physical instantiation of the algorithm (i.e., how the computation gets carried out by neural hardware). In this way, CNNs and the visual system may be considered comparable on the computational and algorithmic levels, while the specifics of their implementations differ (e.g., the feature replications in biology vs. convolutions in CNNs).

Aside from simply noting the broadly similar design of CNNs and the visual system, more direct methods of comparing their functions are possible. Specifically, relating the activity of units in a CNN trained/optimized for visual categorization on the ImageNet database with that of neurons in the visual system is a way of testing if the two systems really are functioning similarly (Kriegeskorte, 2015; Wardle & Baker, 2020; Yamins & DiCarlo, 2016). Comparing artificial to biological neural activity is generally done via two different kinds of analysis: representational similarity analysis (RSA) or a measure of variance explained. RSA (Kriegeskorte et al., 2008) is done at the population level; specifically the activity of each population of interest—for example, a population of neurons in a brain area or a pattern of functional magnetic resonance imaging (fMRI) activity and a population of units at a layer in a CNN—is recorded in response to the same set of images (Nili et al., 2014). A matrix is then created for each population. Each entry in this matrix corresponds to how similar the activity vector of that population is in response to two different images. These two matrices are compared across the populations as a quantitative measure of how similarly or dissimilarly they represent the images. This and related methods (such as canonical correlation analysis (Morcos et al., 2018) and centered kernel alignment (Kornblith et al., 2019)) have been used to compare brain areas to models (Kriegeskorte, Mur, Ruff, et al., 2008), brain areas to brain areas (Hunt et al., 2018), and models to models (Mehrer et al., 2020).

In the second method of analysis, the activity of a population (usually the activity of units in a CNN layer) is directly used to predict the activity of a single neuron (or possibly voxel) in the biological system (Yamins et al., 2014). For this, a separate linear regression model is trained on the CNN activities to predict the responses of individual neurons for a subset of the images. The ability of the regression model to predict neural activity on held-out data is then used to calculate how much variance in the neural response the model activity can explain.

Many studies have made such comparisons and found that deep CNNs that perform well on the challenging ImageNet classification task create patterns of artificial neural activity that are similar to the neural activity produced in the processing pathway of the visual system known as the ventral stream (see Lindsay (2020b) for a review). In particular, different layers in the CNN tend to correspond best to different areas in the ventral stream. For example, later layers best match the activity of the inferior temporal (IT) cortex while earlier layers have more in common with primary visual cortex (Antolík et al., 2016; 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). This finding holds in general for many different CNN architectures but scan break down for networks that are too deep (Storrs et al., 2020).

CNNs trained for face recognition have been shown to learn facial representations that are largely consistent with those of face-selective neurons found in the human ventral stream (Grossman et al., 2019). However, while higher layers of a face-optimized CNN seem largely consistent with the responses of neurons found in the anterior region of the face system, to date no single layer has been able to account for the tuning of neurons found in more posterior regions (Raman & Hosoya, 2020).

The Brain-Score platform was built to track how well different CNNs match brain data (Schrimpf et al., 2018). It compares different models against a collection of monkey electrophysiology datasets from four different brain areas in addition to human psychophysics data and keeps track of how well each model performs on image classification tasks. From this, several conclusions have emerged. For example, better performance on image classification is related to better neural predictivity but this relationship breaks down for some of the best performing networks. Also, networks with very different architectures can perform equally well on the Brain-Score measures (this is also true of the aforementioned Representational similarity analysis studies with human fMRI data, see Storrs et al., 2020). Finally, even the most predictive models cannot explain all of the neural activity. While such an explicit benchmark can help make comparisons across models clear, it should be noted that the Brain-Score results are still dependent on the specifics of the datasets and metrics used for comparison.

Many studies comparing CNNs to neural data start by training the CNN on a visual task. However, it is possible to train a network to replicate neural data directly. Such an approach has been helpful in explaining neural responses in the earliest stages of visual processing—in the retina. These retinal studies train a small CNN directly to model the activity of retinal ganglion cells (the output neurons of the retina) as a function of an input image. The structure and responses of the trained CNN can then be compared to the anatomy and activity of cells in the retina. Quantitative fits have been steadily improving by incorporating anatomical constraints into CNN models including the modular organization of the retina (Klindt et al., 2017; Maheswaranathan et al., 2018; Tanaka et al., 2019). In the cortex, fitting models to data has shown the importance of recurrence to neural representations (Kietzmann et al., 2019) and that data-trained models have similar success in fitting primary visual cortex (V1) neural data as task-trained networks (Cadena et al., 2019). Furthermore, training a network to both perform image classification and mimic the statistics of brain activity has been found to increase task performance (Federer et al., 2020).

Units in a CNN and Neurons: Visualizing What They Respond To

Another way of probing the function of CNNs and comparing them to the brain is to look at what kind of visual stimuli drive different units in the network. Such feature visualization can be done directly in CNNs because the full connectivity of the network is known (Olah et al., 2017; Qin et al., 2018). In the "optimization" approach to visualization, an initially meaningless image (e.g., a white noise image) is optimized via the backpropagation algorithm to reveal the features that drive a unit. Specifically, unlike network training (where backpropagation is used to change the network weights), during feature visualization the weights are held constant and the pixels in the input image are changed such that the activity of a selected unit increases. This is similar to how an experimenter may show different stimuli to an animal to find the features that drive a recorded neuron most strongly, but in the case of the artificial network, the mathematics allows this process to be done optimally and not through trial and error.

These methods reveal the increasing complexity of visual features that drive units in a CNN. Units in the early layers of a CNN trained to perform object classification respond strongly to oriented lines or patches of contrasting colors. This mimics the response properties of neurons in the early stages of the visual system of many mammals (Hubel & Wiesel, 1959; Scholl et al., 2013). Units in later layers get inputs from these units and gain certain invariances, for example detecting the same oriented patterns in a wider range of spatial locations or responding strongly to conjunctions of simple features or textures. Again, this mimics the responses seen in complex cells in the primary visual cortex as well as the responses of neurons in secondary visual areas (e.g., V2) (Anzai et al., 2007). Continuing through the network, feature visualization techniques have found curve-detecting units whose properties most easily map to V4 (Connor et al., 2007; Gallant et al., 1993). These eventually lead to partial or full object detectors such as shape detectors, eye detectors, face detectors, and so on, which would be analogous to cells of the later ventral stream in the temporal cortex (Kobatake & Tanaka, 1994). Example feature selectivities sampled from a deep neural network called Inception along different processing stages are shown in figure 2.

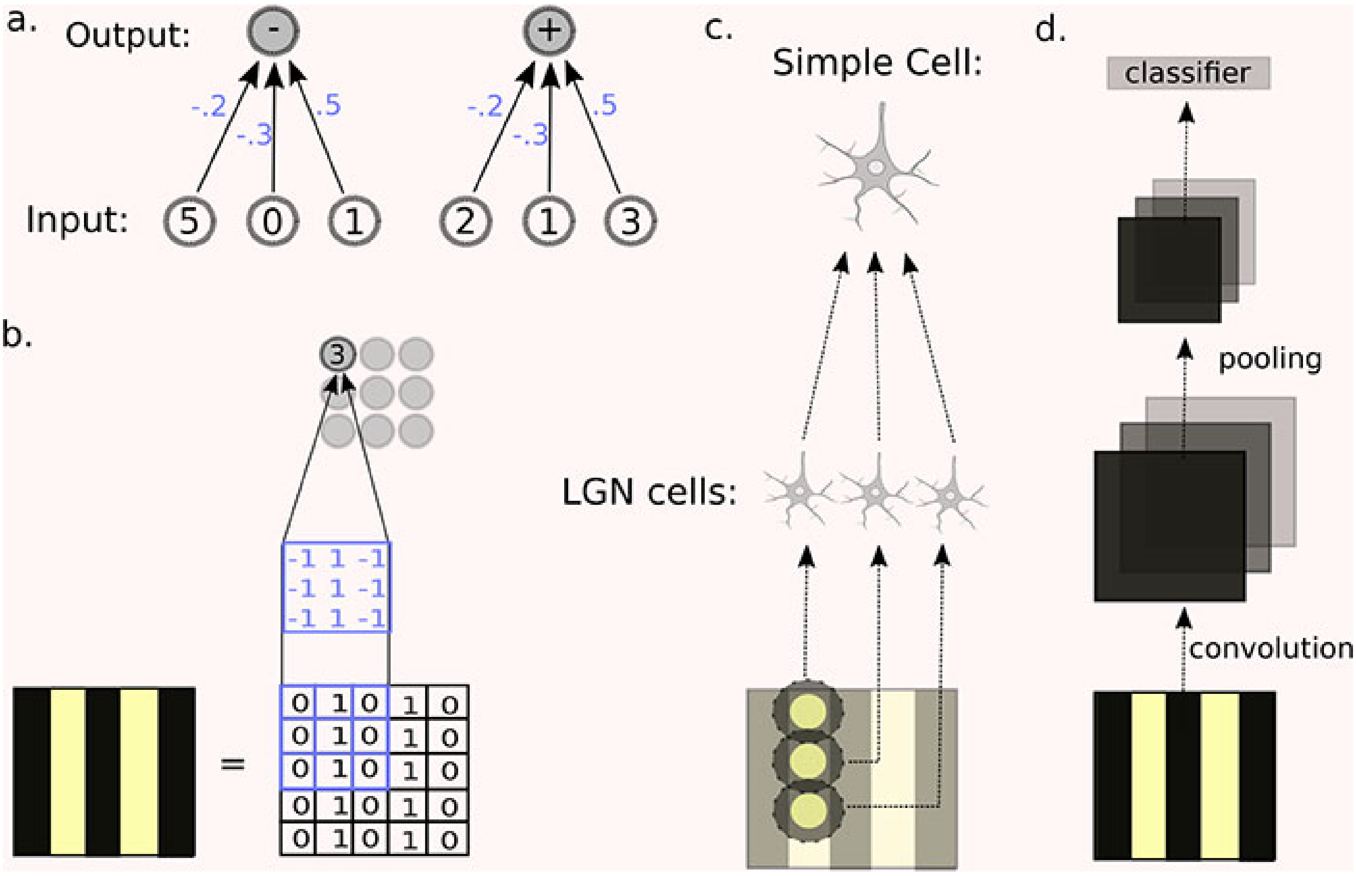

Figure 2. Applying a sample feature visualization method to probe the selectivity of neurons along different layers of a deep neural network.

Note: Along the network's visual hierarchy, units exhibit selectivity for increasingly complex preferred stimuli in a way that qualitatively mimics the organization of the mammalian visual system. Figure adapted from Olah et al. (2017) under Creative Commons Attribution CC-BY 4.0. Visualizations of all channels are available in Olah et al. (2017). Further analyses and interactive web tutorials on interpreting how deep networks categorize images can be found in Cammarata et al. (2020), Carter et al. (2019), Mordvintsev et al. (2018), Olah et al. (2018, 2020a, 2020b), and Sturmfels et al. (2020).

Traditionally, identifying which stimuli maximally drive neurons in the visual system requires an extensive process of trial and error: different stimuli are shown and the responses of neurons recorded. There are many downsides to this approach; for example, it is labor-intensive and can provide only a partial picture of a cell's response that is biased by which stimuli were tested. CNNs that can predict single neuron responses have been used to perform the same kind of image optimization used to visualize the preferences of artificial units. In Bashivan et al. (2019), this method generated stimuli that could reliably drive the activity of V4 neurons beyond that elicited by the images of curved lines frequently used to study V4. Another study (Ponce et al., 2019) used a different style of CNN, a generative adversarial network (GAN) (Goodfellow et al., 2014), to produce effective artificial stimuli for the human IT cortex. Similar methods have also been used to reconstruct visual inputs from fMRI data during perception (Mozafari et al., 2020; Seeliger et al., 2018).

Visualization techniques using CNNs have also provided a means to measure mental imagery. In Shen et al. (2019), a mapping between activations in a CNN and patterns of fMRI activity was made by showing the network and participants the same natural images. This mapping was then used to visualize images that participants were imagining. Specifically, the fMRI activity during mental imagery in the absence of visual inputs was mapped to CNN activity, and this activity was used to construct an image that could elicit such activity in the network. Though still very crude, these constructed images did bear some basic resemblance to the images subjects were instructed to imagine.

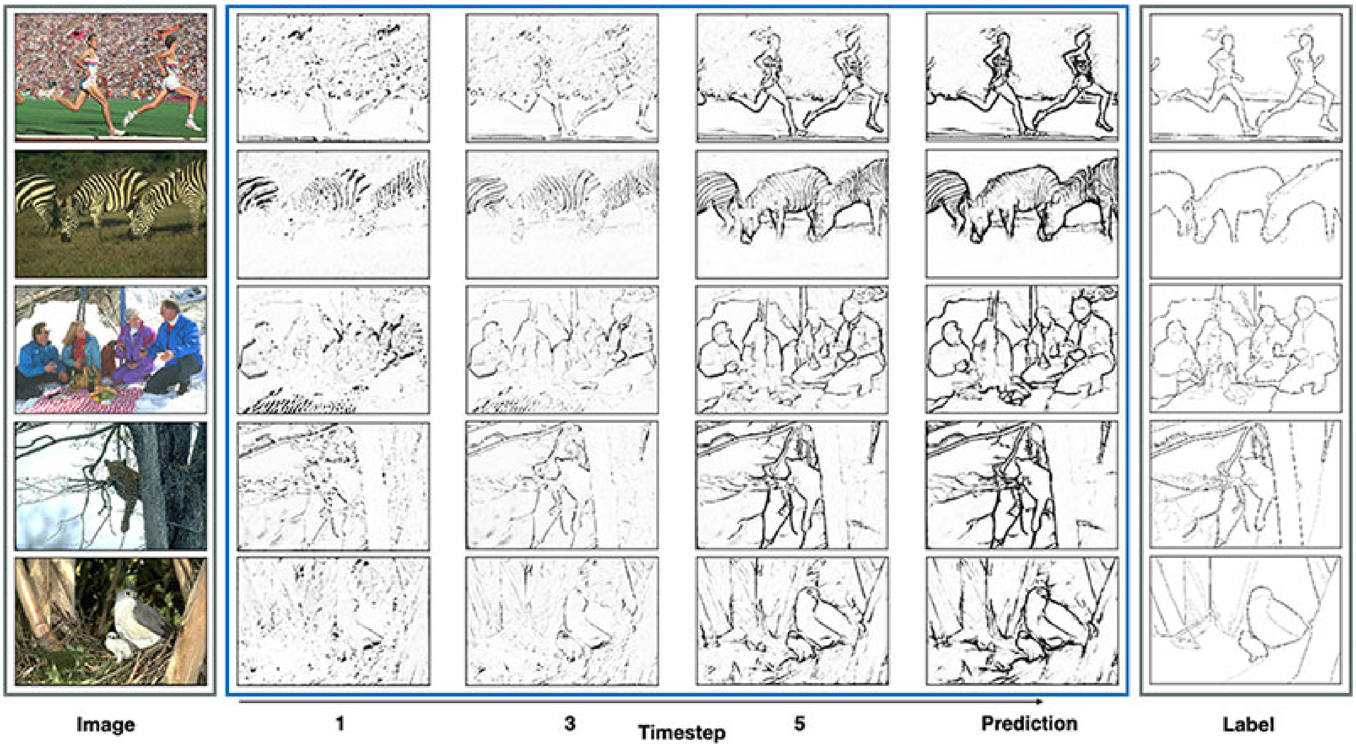

Figure 3. Interpreting how human observers vs. deep neural networks categorize natural images.

Note: From left to right columns: sample images from ImageNet and associated category labels, feature importance maps derived from human observers using a gamified version of classic psychophysics methods (Linsley et al., 2017) and using a similar attribution method on a CNN known as sensitivity analysis (Zeiler & Fergus, 2014), and saliency maps derived from human observers for comparison.

Source: Image credit: Drew Linsley.

In psychophysics, researchers have used “classification images” to understand classification behavior in humans and animals. This process involves adding random perturbations to an image to determine which features of the image help or hinder classification (Murray, 2011). Similar methods have been applied to CNNs (Zeiler & Fergus, 2014). However, the image features relevant for classification in these networks can also be calculated directly through “attribution” methods (Olah et al., 2018). These methods involve creating “importance” maps; this is done by using backpropagation to determine how changes in pixel values impact the classifier (Simonyan et al., 2013). Some of these methods are in some sense related to classic animal electrophysiology and human psychophysics methods ranging from reverse correlation to classification image methods (Schyns et al., 2002) and have been used to reveal, in particular, how deep neural networks discriminate between faces (Xu et al., 2019). Similar methods have also been used to compare the visual representations learned by deep neural networks and human observers (Linsley et al., 2019) and to demonstrate that deep neural networks are able to leverage shortcuts such as watermarks to categorize images (Lapuschkin et al., 2019). Figure 3 shows importance maps derived from human observers using psychophysics methods (Linsley et al., 2017) and a deep neural network using standard attribution methods. Modern CNNs trained to categorize images using the ImageNet dataset appear to leverage visual features that fall somewhere between human-derived “top-down” features that are used by human observers for image categorization and “bottom-up” saliency features (see Geirhos, Jacobsen, et al., 2020; Linsley et al., 2017), although they can be cued to attend to features that are important for human observers, leading to learned visual representations that match those derived from human observers much more closely (Linsley et al., 2019).

CNN Representations to Understand the Space of Visual Stimuli

Parameterizing simple visual stimuli is relatively straightforward. On the one hand, the images of oriented lines used to probe the function of the primary visual cortex can be defined mainly by the angle of the orientation, and certain other factors such as spatial frequency or contrast. On the other hand, more naturalistic stimuli such as real-world images are harder to describe with just a small set of values.

Passing an image through a CNN, however, can produce a new representation of that image, the number of dimensions of which is equal to the number of units in the layer being considered. Particularly, using the last or penultimate layer of a CNN can offer a representation that is smaller than the number of pixels in the image and contains semantically relevant information. While this representation doesn’t necessarily have a neat or direct mapping to known image features such as object identity, color, size, rotation, and so on, it can still help experimenters to understand and represent their stimulus space in a compact and meaningful way.

In Bao et al. (2020), the authors collected the activity of the second-to-last layer of a CNN in response to a stimulus set of objects and performed dimensionality reduction on it. Dimensionality reduction is a technique that allows the activity patterns of a large neural population response to be represented with fewer dimensions if there are correlations among a network’s units. In this study, the authors reduced the activity to a 50-dimensional stimulus space. They showed that many individual IT neurons were tuned to different dimensions in this space—that is, the neuron’s firing rate in response to an image was as a function of where that image fell on a particular axis. On average, a neuron was tuned to seven of these dimensions and its activity did not vary significantly in response to changes in the others. The CNN representation thus provided a compact description of the image features relevant to the neural response in this area. What's more, using fMRI during the viewing of natural and artificially generated images, the authors were also able to identify a topographic map in IT. In this map, cells are clustered into four quadrants with nearby cells having similar preferred stimuli. These quadrants are defined by the first two axes of the CNN-identified stimulus space (which correspond roughly to "spiky" versus "stubby" and inanimate versus animate objects). The understanding of stimulus space provided by CNNs thus helps explain the physical layout of neurons in the brain (figure 4). In a similar vein, Lee et al. (2020) predicted the spatial arrangement of cells in IT by placing units in their model with similar responses near each other. This created clusters of units with different selectivity properties that resembled the face patch system.

$$g(x) = \ln\left(1 + e^{x}\right)$$

Figure 4. Representation space in IT.

Note: (a) The first two principal components of the object representation learned by a CNN correspond roughly to a stubby vs. spiky and an animate vs. inanimate axis. (b) Cells that respond preferentially to objects in the four quadrants of this space can be found in different locations in IT.

Source: Figure adapted from Bao et al. (2020) with permission.

The study of facial processing has relied on such representational spaces for decades, starting with the introduction of the notion of a "face space" in the early 1990s. Mathematically decomposing face images into their component parts (known as eigenvectors) provides a set of dimensions on which any face image can be placed, and location in this space can predict identity features and recognizability (O'Toole et al., 1993). Later work has explored how the representation learned by CNNs trained on facial recognition compares to this traditional face space (O'Toole et al., 2018). CNN-based representations are arranged opposite to traditional face space with uninterpretable images (such as low-quality and off-center face images) grouped in the center (figure 5). Dimensions in this space do not correspond neatly to physical face features or pose, illumination, and so on; however, this information is present even in the last layer of a face-trained CNN.

While the representations provided by CNNs can help parameterize natural images, they are not themselves immediately interpretable. Some studies have aimed to increase the interpretability of CNN representations while also ensuring they still capture biological data (Jha et al., 2020).

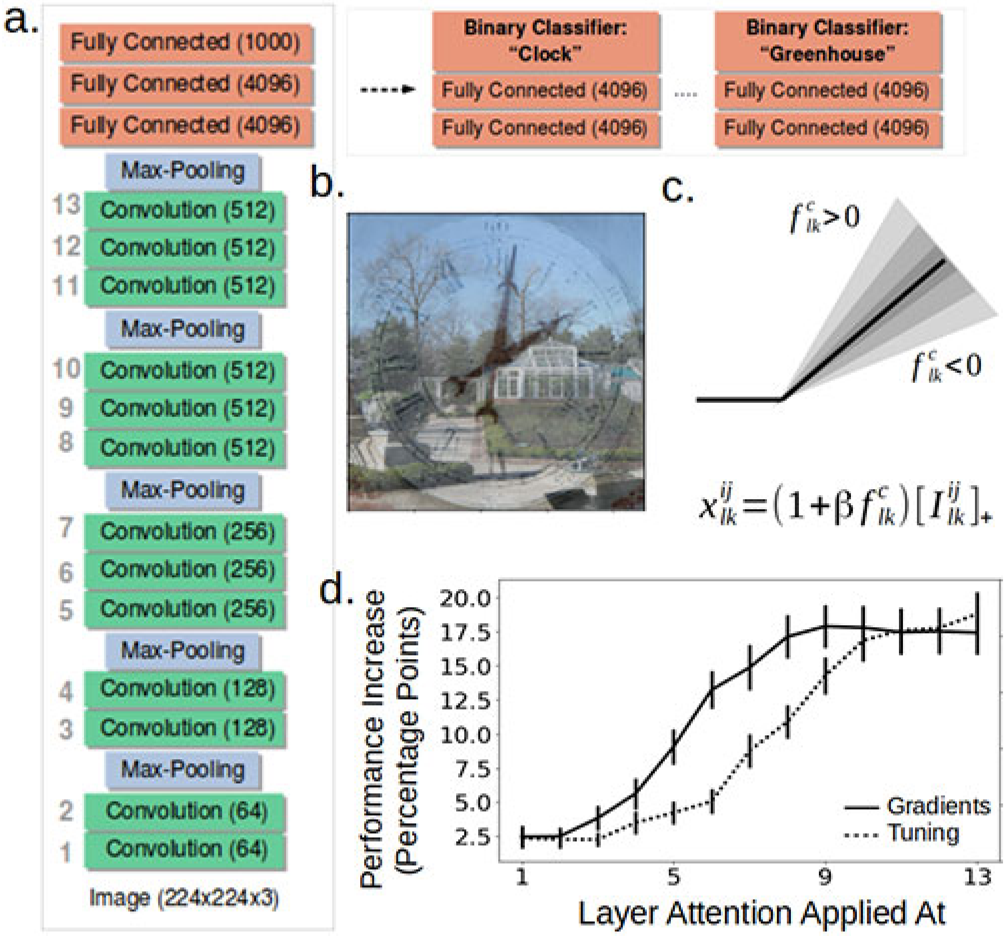

Figure 5. ace space as determined by a CNN trained on face recognition.

Note: Images cluster based on quality and image features. For example, poor-quality images are at the center. Outside the center, off-center facial views can be found in inner rings and frontal views are found more peripherally. The space is reduced to two dimensions and plotted via the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm.

Source: Image modified from O'Toole et al. (2018) with permission.

CNN Representations to Understand Human Behavior

Human annotations have played a significant role in the history of computer vision by providing ground truth for visual tasks from contour annotations (Januszewski et al., 2018; Martin et al., 2001) to image categorization (Oliva et al., 2001). Behavioral studies test the input–output mapping of the visual system and establish constraints on computation. CNN representations have allowed for testing between competing hypotheses regarding the nature of behavior and, in particular, which visual tasks require attention and which do not, and the role of bottom-up and top-down processing.

The visual system contains at least three dozen visual areas that are almost always reciprocally connected. In principle, this could lead to very complex visual dynamics (Kreiman & Serre, 2020). However, to a first approximation, the underlying visual dynamics can be roughly simplified into two main modes of visual processing: an initial—primarily feedforward or bottom-up—sweep of activity followed by re-entrant or recurrent/feedback signals. It has long been hypothesized that when observers are forced to operate at their temporal limit such as during speeded ultra-rapid categorization tasks when stimuli are briefly flashed for a few tenths of a second and responses are forced to be as fast as can be made, visual recognition may proceed based on only a single feedforward sweep of activity (Thorpe et al., 1996)—also known as core recognition—with limited contributions from feedback processes.

Matching the level of accuracy of human participants during ultra-rapid categorization tasks was indeed the main goal of computational models of the ventral stream of the visual cortex during the pre-deep learning era (Serre, Oliva, et al., 2007). Rapid categorization performance measures are a primary benchmark for CNNs (Eberhardt et al., 2016; Geirhos et al., 2018; Kheradpisheh et al., 2016; Rajalingham et al., 2015). However, for specific visual recognition tasks, there are already unspeeded tasks for which deep neural networks appear to outperform human observers including object (He et al., 2016) and face (Phillips et al., 2018) categorization, contours detection, and image segmentation (He et al., 2019; Lee et al., 2017). These claims have to be taken with a grain of salt as there is ample evidence that CNNs fall short of the robustness and generalization ability of the visual system (Geirhos et al., 2018; Jo & Bengio, 2017; Papernot et al., 2017; Rosenfeld, Zemel, et al., 2018; see Serre (2019) for a more in-depth review).

A more nuanced measure of accuracy for multi-class categorization problems includes the confusion matrix. A confusion matrix is defined such that each row of the matrix provides the probability of assigning an image of a particular class to each of the possible classes. A perfectly performing system would produce a confusion matrix where all diagonal entries are one (i.e., the probability of assigning the correct class is one for each class). An unbiased classification system at chance would produce any class label with equal probability for each category and its confusion matrix would thus be a unit matrix (subject to some normalization constant because rows have to sum up to one). Confusion matrices were used to compare the level of accuracy of CNNs vs. human observers across image transformations and object categories (Ghodrati et al., 2014; Kheradpisheh et al., 2016; Rajalingham et al., 2015). One study leveraged the Amazon Mechanical Turk crowdsourcing platform (Buhrmester et al., 2016) to collect over 500,000 behavioral decisions from a few thousand participants for 10,000 natural images from 10 object categories (Battleday et al., 2019). These choices were compared to model-generated distributions, allowing a more precise evaluation of the behavior of these models and a better understanding of the computations humans use to categorize images.

With a large enough number of participants, human accuracy can also be computed per individual stimulus to be correlated with CNN confidence scores (measured by their signed output responses to individual stimuli) (Eberhardt et al., 2016; Geirhos, Meding, et al., 2020; Kubilius et al., 2016; Rajalingham et al., 2018). The Brain-Score discussed in the section "Modeling Neural Activity with Deep Neural Networks" also includes a per-image human accuracy score to be correlated with candidate models (Schrimpf et al., 2018). It should be noted, however, a later study has shown that much of the observed correlation between CNNs and human participants is due to chance, and when proper statistical measures are used much of the consistency vanishes (Geirhos, Meding, et al., 2020).

Other possible measures used to compare CNNs with human behavior include similarity judgments. These can be derived from direct or indirect similarity scoring between all pairs of stimuli. Initial studies using simple object silhouette and natural images had found that CNNs account well for human shape judgments (Kubilius et al., 2016; Peterson et al., 2018). More generally, CNNs appear to be good models of 2D vision as assessed by a battery of psychophysics tests; for example, they show a primate-like confusion for mirror-symmetric views of an object (Logothetis & Sheinberg, 1996), an advantage for object categorization in congruent vs. incongruent scenes, and an adherence to Weber's law (Jacob et al., 2020). However, a growing body of literature suggests key differences in the visual strategies used by ImageNet-trained CNNs vs. human participants in their ability to process shape information: unlike human observers, modern deep neural networks appear to rely largely on texture cues and other shortcuts for object recognition (Geirhos et al., 2019; Geirhos, Jacobsen, et al., 2020). In contrast, human observers appear to judge shape similarity in a viewpoint-insensitive manner based on 3D shape features (Erdogan & Jacobs, 2016; German & Jacobs, 2020; Pramod & Arun, 2016). CNNs may have access to some shape information in the form of local edge relations, but they do not seem to encode global object shapes (Baker et al., 2018). Another study has shown that CNNs do not seem to register 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).

Beyond matching accuracy and similarity judgments, there is also a trend toward assessing the sensitivity of CNNs to visual illusions. Face recognition is particularly relevant because CNNs have been shown to match human-level accuracy for face matching (Phillips et al., 2018). Interestingly, CNNs have been shown to be sensitive to the Thatcher illusion (such that it is more difficult to detect local feature changes in an upside-down face, despite identical changes being obvious in an upright face; see Thompson, 1980) only after being trained to recognize (upright) faces but not after being trained for generic object recognition. The implication is that the Thatcher illusion arises as a consequence of the training of neural mechanisms specifically for face recognition as opposed to object recognition (Jacob et al., 2020). In the domain of contour processing, Linsley et al. (2020) have shown that feedback mechanisms are necessary for CNNs trained for contour detection to be sensitive to the tilt illusion (O'Toole & Wenderoth, 1977).

In a similar vein, Ullman et al. (2016) found that, when presented with small cropped regions from object images, 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 (but see also Funke et al., 2021). A subsequent study by Linsley et al. (2017) tried to compare more directly the visual representations learned by CNNs with those used by human observers. Using Clicktionary, a collaborative web-based game they developed to identify diagnostic visual features for human object recognition, they were able to compare importance maps derived from representative CNNs directly using attribution methods (see "Units in a CNN and Neurons: Visualizing What They Respond To") with importance maps derived from human observers. The analysis revealed that CNNs and human observers 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 the 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, Zemel, et al., 2018; Saleh et al., 2016; Tang et al., 2018; Wang, Zhang, et al., 2017).

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. However, more work has shown that important features of human judgments are missing from CNN representations (Jozwik et al., 2017). Yet the difference between the two systems may be more quantitative rather than qualitative: it has been shown that a simple linear transformation of these representations (analogous 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 study targeting higher-level concepts using similarity ratings demonstrated that despite the authors' best efforts, none of the tested CNNs were able to reproduce the image judgments produced by human observers (Rosenfeld, Solbach, et al., 2018). 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 had 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.

Taking diagnostic tests from cognitive science can help identify the precise strategies of CNNs and how they differ from humans. In particular, careful stimulus selection and variation can reveal behaviors not immediately obvious from simple classification (see Ma & Peters (2020) for a more in-depth treatment).

Computational Modeling beyond the Feedforward Sweep

Core object recognition, the ability to rapidly recognize objects under different visual settings, has been extensively studied using both electrophysiological (DiCarlo et al., 2012) and psychophysical methods (Fabre-Thorpe, 2011). The very success of CNNs in capturing several neural and behavioral aspects of core object recognition suggests that much of the computations required to successfully categorize images can be largely approximated as a bottom-up cascade of filtering, rectification, and normalization operations. This provides computational evidence for the feedforward hypothesis (Serre, Kreiman, et al., 2007).

Despite these successes, it is also becoming increasingly clear that modern deep neural network models remain outmatched by the power and versatility of the primate brain (see Kreiman & Serre (2020) and Serre (2019) for reviews). CNNs can be easily fooled by small (perceptually invisible) levels of noise applied to images (Goodfellow et al., 2015) and they struggle with occlusions (Wang, Xie, et al., 2017) and clutter (Volokitin et al., 2017), possibly due to their inability to process and represent individual objects separately from context. Beyond basic object recognition tasks, these networks struggle to learn to solve rather simple but abstract visual reasoning problems such as judging whether two never-seen-before items are the same or different—even after observing millions of training examples (Ellis et al., 2015; Fleuret et al., 2011; Gülçehre & Bengio, 2013; Kim et al., 2018; see Ricci et al., 2020, for a review). Although such algorithms are reasonably good at recognizing the presence of certain objects in the scene, they often fail miserably at flexibly interpreting the fundamental gist of complex visual scenes, human actions, social interactions, and events depicted in images. To date, no known artificial system is capable of passing a visual Turing test as defined in Geman et al. (2015).

The limitations of modern deep neural network models underlie critical aspects of visual cognition that are not accounted for by purely feedforward networks. Primate vision is often described as roughly composed of two distinct processing phases. First, an early “bottom-up phase” primarily carried by feedforward processes during the first 150 milliseconds after visual onset and which seems to be well approximated by CNNs and other related deep neural networks. However, this bottom-up phase is typically followed by a late “re-entrant” phase carried by feedback processes (Lamme & Roelfsema (2000); but see also Bullier (2001) for evidence of early contributions of feedback on neural responses). The very success of CNNs for image categorization suggests that feedforward processing is sufficient for rapid categorization tasks (Serre, Oliva, et al., 2007; Thorpe et al., 1996; VanRullen, 2007) but their relative difficulty in addressing more general visual reasoning problems suggest that they may necessitate bringing in feedback signals (Kreiman & Serre, 2020).

Consistent with this idea, several studies have noted that purely feedforward CNNs struggle to predict neural activity later in the response to object images, especially for more challenging images. Kar et al. (2019) showed that adding lateral recurrence—that is, connections between units at the same layer in the CNN—can make these models a better fit. In Kietzmann et al. (2019), both lateral connections and feedback from later layers to earlier ones were added to a feedforward CNN and trained to reproduce neural dynamics. These connections also enhanced classification performance.

More specific computational uses of recurrent connections have also been studied. For example, the ability of dynamic computations to implement a speed-accuracy tradeoff was explored using CNNs in Spoerer et al. (2020). Behavioral, neurophysiological, and computational evidence suggests that purely bottom-up computations are generally insufficient to perform pattern completion of heavily occluded objects (Wyatte et al., 2014). During natural visual conditions, many objects are partially visible either because they are occluded by other objects in front of them or because of poor illumination or unusual viewing angles. Despite such challenging visual conditions, primate visual recognition is quite robust even when most of the object is occluded, in the absence of contextual cues, and when subjects have minimal prior experience with the object in question (Tang et al., 2014). These behavioral and neurophysiological observations are further corroborated by computational models: state-of-the-art feedforward network models struggle during recognition of heavily occluded objects unless they are extensively trained with occluded examples of those specific objects (Rosenfeld, Zemel, et al., 2018; Wang et al., 2018). Recurrent connections are also believed to aid the processing of noisy or occluded images (Iyer et al., 2020; Spoerer et al., 2017; Tang et al., 2018; Xiao et al., 2020).

Several forms of attention are known to require top-down mechanisms (Itti et al., 2005). In Lindsay and Miller (2018), the activity of units in a CNN was modulated to replicate the ways in which neurons in the ventral stream are modulated by feature-based attention (figure 6). According to one of the leading models of attention, the Feature Similarity Gain model, during a visual search task, neurons in the visual cortex should be modulated according to their "tuning," that is, the degree to which they respond preferentially to images of a given category (Martinez-Trujillo & Treue, 2005). In short, searching for a red fire hydrant should preferentially facilitate neurons tuned to red while searching for a banana should preferentially facilitate neurons tuned to yellow. Replicating these activity changes in a CNN led to performance changes in the network that mimicked the performance enhancements that subjects in attention tasks show. Furthermore, the study suggests an intriguing alternative proposal for attentional modulation within the visual cortex. In a deep neural network, an optimal modulation can be explicitly computed for individual object categories using backpropagation to calculate "gradient" values. Gradient values indicate the ways in which feature map activities should change in order to make the network more likely to classify an image as being of a certain object category. As shown in figure 6, such a modulation scheme yields higher accuracy gains than the changes inspired by the Feature Similarity Gain model (particularly when applied at middle layers). These gradient values were found to only partially correlate with predictions derived from the Feature Similarity Gain model, suggesting a theoretically grounded alternative to that model. Such an approach shows how CNNs can be used to study the connection between neural activity modulation and behavior under the influence of attention (Lindsay, 2020a; Thorat et al., 2019).

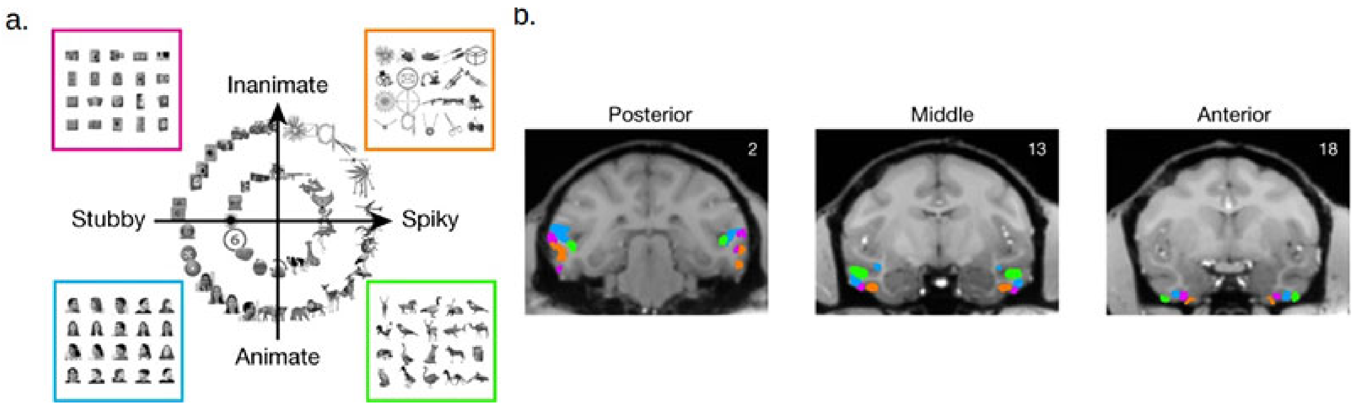

Figure 6. Effects of applying feature-based attention on individual layers of a deep neural network.

Note: (a) This study uses a pretrained deep neural network (VGG-16) that contains 13 convolutional layers (labeled in gray, number of feature maps given in parenthesis) and is trained on the ImageNet dataset to do 1,000-way object classification. To perform feature-based attention tasks, the final layer that was implementing 1,000-way softmax classification is replaced by binary classifiers (logistic regression), one for each category tested (two shown here, 20 total). These binary classifiers are trained on standard ImageNet images. (b) Merged images contain two transparently overlaid ImageNet images of different categories and are used in the attention tasks for the network. (c) Schematic of how attention modulates the activity function. All units in a feature map are modulated the same way. The slope of the activation function is altered based on the tuning (or gradient) value, f^{c}_{lk}, of a given feature map (here, the kth feature map in the lth layer) for the attended category, c, along with an overall strength parameter β. I^{ij}_{lk} is the input to this unit from the previous layer. (d) Average increase in binary classification of merged images as a function of the layer at which attention is applied (solid line represents tuning values, dashed line gradient values, error bars ± S.E.M.). See Lindsay and Miller (2018) for details.

Source: Modified from Lindsay and Miller (2018) with permission.

In a study by Linsley et al. (2019), it was shown that while a state-of-the-art deep neural network learned visual representations that are quite different from those used by human observers for visual categorization, cueing these same networks to attend to image locations that are important for human observers not only helps categorization accuracy but also leads to learned representations that are much more similar to those used by human observers.

Beyond image categorization, it has long been assumed that feedback mechanisms play a key role in perceptual grouping (Gilbert & Li, 2013; Grossberg et al., 1997; Li, 2002). Yet the successes of deep convolutional networks for contour detection and image segmentation in seemingly challenging visual tasks (e.g., He et al., 2019; Lee et al., 2017) have made the role of feedback unclear. Linsley et al. (2018) described a simple visual recognition challenge inspired by cognitive psychology tasks (see Roelfsema et al., 2000 for review) called the “Pathfinder,” which involves judging whether there exists a path linking two markers in an image. Using this task, they showed that while deep feedforward neural networks could solve the task using a brute-force strategy, a single layer of a highly recurrent neural network that imbue neurons with the ability to incorporate context through horizontal connections was able to perform on par or better than all tested feedforward hierarchical baselines, despite the fact that these feedforward networks contained orders of magnitude more parameters.

In follow-up work, Linsley et al. (2020) showed that such recurrent neural networks, which allow for contextual interactions to take place among neighboring neurons, yield state-of-the-art results for contour detection—on par with human observers (figure 7). Interestingly, the recurrent neural network was shown to learn to solve contour detection tasks with better sample efficiency than state-of-the-art feedforward networks, while also exhibiting a classic perceptual illusion, known as the orientation-tilt illusion (O’Toole & Wenderoth, 1977). Correcting this illusion significantly reduced the network’s contour detection accuracy by driving it to prefer low-level edges over high-level object boundary contours. This suggests that the orientation-tilt illusion is a byproduct of neural circuits that help biological visual systems achieve robust and efficient contour detection, and that incorporating such circuits in artificial neural networks can improve computer vision.

Anatomically, one can distinguish between two kinds of recurrent mechanisms: horizontal/ lateral connections (within a processing stage) vs. top-down connections (from a higher to a lower processing stage). The role of these horizontal vs. top-down connections was studied by Kim et al. (2020), who extended the Pathfinder challenge, which stresses low-level gestalt cues, to a task which they called "cluttered ABC" (cABC), which emphasizes high-level object cues for perceptual grouping. The authors found a double dissociation between horizontal and top-down connections: horizontal connections are needed to solve the Pathfinder featuring gestalt cues by relying on incremental spatial propagation of activities, while top-down connections help rescue learning on tasks such as cABC featuring object cues by propagating coarse predictions about the expected position of the target object. These findings thus disassociate the computational roles of bottom-up, horizontal, and top-down connectivity, and demonstrate how a recurrent network model featuring all these interactions can more flexibly form perceptual groups.

Beyond perceptual grouping, several other computer vision tasks have been shown to benefit from a similar inclusion of recurrent processing including image generation (Van Den Oord et al., 2016), object recognition (Liang & Hu, 2015; Liao & Poggio, 2016; O'Reilly et al., 2013; Zamir et al., 2017), and super-resolution tasks (Kim et al., 2016).

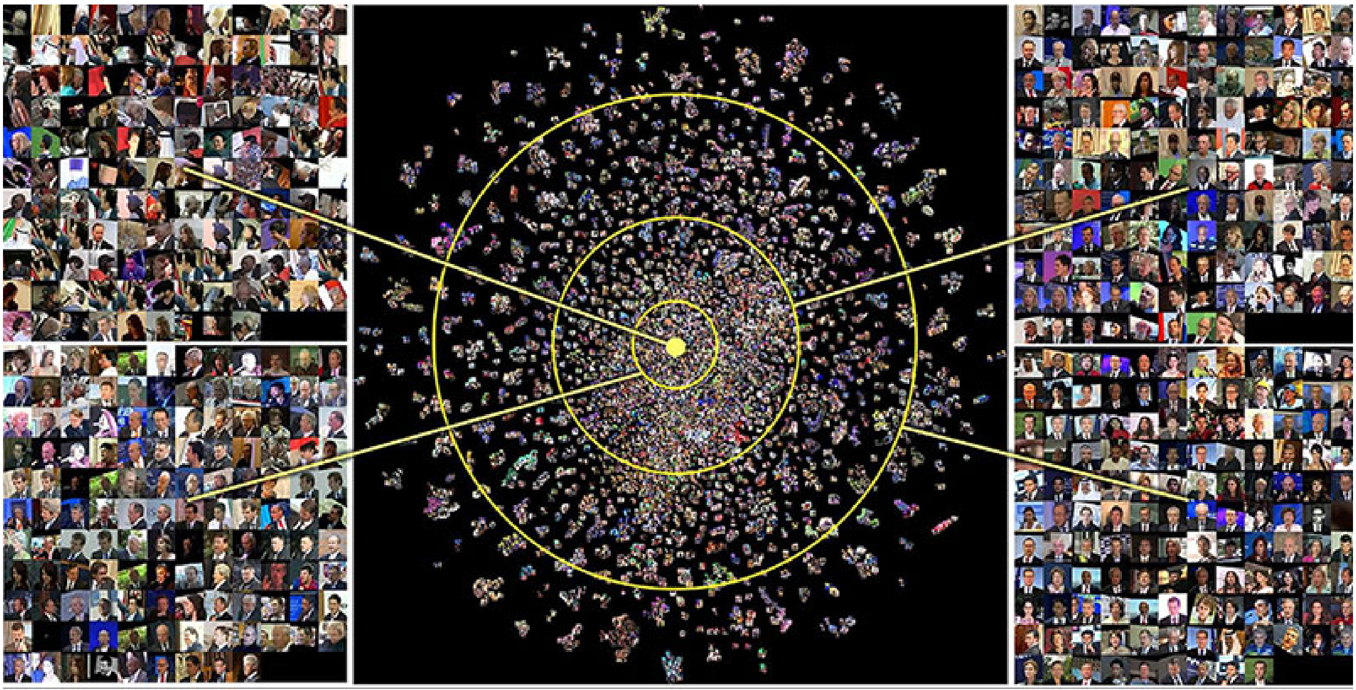

Figure 7. Recurrent neural network solving a contour detection task.

Note: The evolution of recurrent neural network predictions across time steps of processing (Linsley et al., 2020). Predictions are initially coarse and are then refined over processing time steps to select figural object contours. Unlike in a standard feedforward neural network such as a CNN, which processes information in a single pass, a recurrent neural network such as the one used here increases processing depth through time and is able to dynamically update perceptual decisions.

Source: Replicated with permission from the authors.

Conclusion

Progress in deep learning has spawned great successes in many engineering applications. As a prime example, CNNs, a type of feedforward neural networks, are approaching—and sometimes even surpassing—human accuracy on a variety of visual recognition tasks. 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 and behavioral data from human observers. As a result, deep CNNs have become de facto models of primate vision. From a neuroscience and psychology 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 commonly studied architectures are becoming increasingly clear. Rigorous methods of comparing artificial and biological visual processing are important for determining exactly where the two differ. Through the identification of these differences, better models can be made and understanding of the biological underpinnings of vision can advance. A better understanding of biological vision can, in turn, lead to better artificial visual systems. However, this is not always a straightforward outcome nor is it the goal of research on biological vision. In sum, the interplay between machine vision and biology is complex. Yet advances in the former are already starting to shape understanding of the computations underlying vision and will likely continue to do so.

Acknowledgments

Grace W. Lindsay was funded by a Marie Skłodowska-Curie Individual Fellowship (No. 844003). Thomas Serre was supported by Office of Naval Research (N00014-19-1-2029), National Science Foundation (IIS-1912280), and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004).

Further Reading

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. In *Oxford research encyclopedia of neuroscience*.

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. *Current Biology*, *29*(7), R231–R236.

Lindsay, G. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 1–15.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3*(3), e10.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., . . . Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*, 55–67.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Visual Neuroscience*, *5*, 399–426.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

References

Antolík, J., Hofer, S. B., Bednar, J. A., & Mrsic-Flogel, T. D. (2016). Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLoS Computational Biology*, *12*(6), e1004927.

Anzai, A., Peng, X., & Van Essen, D. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, *10*(10), 1313–1321.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis *<https://doi.org/10.1126/science.aav9436>*. *Science*, *364*(6439), 1–11.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2019, April 26). Capturing human categorization of natural images at scale by combining deep networks and cognitive models *<http://arxiv.org/abs/1904.12690>*. *arXiv [cs.CV]*.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 133–139). American Psychological Association.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research. Brain Research Reviews*, *36*(2–3), 96–107.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, *15*(4), e1006897.

Cammarata, N., Goh, G., Carter, S., Schubert, L., Petrov, M., & Olah, C. (2020). Curve Detectorsdetectors *<https://doi.org/10.23915/distill.00024.003>*. *Distill*, *5*(6).

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019). Exploring neural networks with activation atlases <*https://distill.pub/2019/activation-atlas/*>. *Distill*, March 6.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 346–358.

Cichy, R. M., 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. *Scientific Reports*, *6*, 27755.

Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology, Figure 17*(2), 140–147.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, Li. (2009). *ImageNet: A large-scale hierarchical image database* [Paper presentation]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255.

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1), 10636.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What makes an object memorable? *IEEE International Conference on Computer Vision (ICCV)*, 1089–1097.

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, Vol. 29 (pp. 1100–1108). Curran Associates.

Ellis, K., Solar-Lezama, A., & Tenenbaum, J. (2015). Unsupervised learning by program synthesis. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28, pp. 973–981). Curran Associates.

Erdogan, G., & Jacobs, R. A. (2016). A 3D shape inference model matches human visual object similarity judgments better than deep convolutional neural networks *<https://pdfs.semanticscholar.org/f4bd/26b49aaeabc1da023534bae3e0dd731454bb.pdf>*. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 2543–2548). Cognitive Science Society

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology*, *2*(October), 243.

Federer, C., Xu, H., Fyshe, A., & Zylberberg, J. (2020). Improved object recognition using neural networks trained to mimic the brain’s statistical properties. *Neural Networks: The Official Journal of the International Neural Network Society*, *131*, 103–114.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(43), 17621–17625.

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

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S. A., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*(3), 16.

Gallant, J. L., Braun, J., Van Essen, D. C., J braun, & VanEssen, D. C. (1993). Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science*, *259*(5091), 100–103.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020, December 18). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency *<http://arxiv.org/abs/2006.16736>*. *arXiv [cs.CV]*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems*, Vol. 31 (pp. 7549–7561). Curran Associates.

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual Turing test for computer vision systems. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3618–3623.

German, J. S., & Jacobs, R. A. (2020). Can machine learning account for human visual object shape similarity judgments? *Vision Research*, *167*, 87–99.

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in Computational Neuroscience*, *8*, 74.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews. Neuroscience*, *14*(5), 350–363.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, Vol. 27 (pp. 2672–2680). Curran Associates.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015, March 20). Explaining and harnessing adversarial examples *<http://arxiv.org/abs/1412.6572>*. *arXiv [cs.CV]*.

Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Computational Biology*, *14*(7), e1006327.

Grossberg, S., Mingolla, E., & Ross, W. D. (1997). Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neurosciences*, *20*(3), 106–111.

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., Khuvis, S., Herrero, J. L., Irani, M., Mehta, A. D., & Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, *10*(1), 4934.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Güçlü, U., & van Gerven, M. A. J. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*(Pt B), 329–336.

Gülçehre, Ç., & Bengio, Y. (2013, July 13). Knowledge Mattersmatters: Importance of prior information for optimization <http://arxiv.org/abs/1301.4083>. *arXiv [cs.CV]*.

He, J., Zhang, S., Yang, M., Shan, Y., & Huang, T. (2019). Bi-directional cascade network for perceptual edge detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3828–3837.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, December 10). Deep residual learning for image recognition <http://arxiv.org/abs/1512.03385>. *arXiv [cs.CV]*.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat’s striate cortex. *The Journal of Physiology*, *148*, 574–591.

Hunt, L. T., Malalasekera, W. M. N., de Berker, A. O., Miranda, B., Farmer, S. F., Behrens, T. E. J., & Kennerley, S. W. (2018). Triple dissociation of attention and decision computations across prefrontal cortex. *Nature Neuroscience*, *21*(10), 1471–1481.

Itti, L., Rees, G., & Tsotsos, J. K. (2005). *Neurobiology of attention*. Academic Press.

Iyer, R., Hu, B., & Mihalas, S. (2020). Contextual integration in cortical and convolutional neural networks. *Frontiers in Computational Neuroscience*, *14*, 31.

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2020, March 5). Do deep neural networks see the way we do? <https://doi.org/10.1101/860759>*bioRxiv*.

Januszewski, M., Kornfeld, J., Li, P. H., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J. B., Tyka, M., Denk, W., & Jain, V. (2018). High-precision automated reconstruction of neurons with flood-filling networks <https://www.nature.com/articles/s41592-018-0049-4>. *Nature Methods*, *15*, 605–610.

Jha, A., Peterson, J., & Griffiths, T. L. (2020, May 29). Extracting low-dimensional psychological representations from convolutional neural networks <http://arxiv.org/abs/2005.14363>. *arXiv [q-bio.NC]*.

Jo, J., & Bengio, Y. (2017, November 30). Measuring the tendency of CNNs to learn surface statistical regularities <http://arxiv.org/abs/1711.11561>. *arXiv [cs.LG]*.

Johnson, S. P. (2001). Visual development in human infants: Binding features, surfaces, and objects. *Visual Cognition*, *8*(3–5), 565–578.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726.

Kalfas, I., Kumar, S., & Vogels, R. (2017). Shape selectivity of middle superior temporal sulcus body patch neurons *<https://doi.org/10.1523/ENEURO.0113-17.2017>*. *eNeuro*, *4*(3).

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior *<https://www.nature.com/articles/s41593-019-0392-5>*. *Nature Neuroscience*, *22*, 974–983.

Kellman, P., Baker, N., Erlikhman, G., & Lu, H. (2017). Classification images reveal that deep learning networks fail to perceive illusory contours. *Journal of Vision*, *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. *Journal of Mathematical Psychology*, *76*(Pt B), 184–197.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, *6*, 32672.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(43), 21854–21863.

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Deeply-recursive convolutional network for image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1637–1645.

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2020). *Disentangling neural mechanisms for perceptual grouping* [Paper presentation]. International Conference on Learning Representations (ICLR), Addis Ababa.

Kim, J., Ricci, M., & Serre, T. (2018). Not-so-CLEVR: Learning same–different relations strains feedforward neural networks. *Interface Focus*, *8*(4), 20180011.

Klindt, D., Ecker, A. S., Euler, T., & Bethge, M. (2017). Neural system identification for large populations separating "what" and "where." In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*, Vol. 30 (pp. 3506–3516). Curran Associates, Inc.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*(3), 856–867.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, July 19). Similarity of neural network representations revisited *<http://arxiv.org/abs/1905.00414>*. *arXiv [cs.LG]*.

Kreiman, G., & Serre, T. (2020). Beyond the feedforward sweep: Feedback computations in the visual cortex *<https://doi.org/10.1111/nyas.14320>*. *Annals of the New York Academy of Sciences*, *1464*(1), 222–241.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience relating representations in brains and models. *Neuroscience*, *2*(November), 1–28.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 1097–1105). Curran Associates, Inc.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4), e1004896.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images *<http://gureckislab.org/papers/LakeZarembaFergusGureckis.CogSci2015.pdf>*. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society*. Cognitive Science Society, 1243–1248.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, *92*(5), 2704–2713.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1096.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition *<https://doi.org/10.1162/neco.1989.1.4.541>*. *Neural Computation*, *1*(4), 541–551.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., & DiCarlo, J. J. (2020, July 10). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network *<https://doi.org/10.1101/2020.07.09.185116>*. *bioRxiv*.

Lee, K., Zung, J., Li, P., Jain, V., & Seung, H. S. (2017, May 31). Superhuman Accuracy accuracy on the SNEMI3D connectomics challenge *<http://arxiv.org/abs/1706.00120>*. *arXiv*.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, *6*(1), 9–16.

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

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex *<http://arxiv.org/abs/1604.03640v1>*. *arXiv [cs.LG]*.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews. Neuroscience*, *21*(6), 335–346.

Lindsay, G. W. (2020a). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, *14*, 29.

Lindsay, G. W. (2020b). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 1–15.

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model <*https://doi.org/10.7554/eLife.38105*>. *eLife*, *7*.

Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., & Serre, T. (2017). What are the visual features underlying human versus machine vision? *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2706–2714.

Linsley, D., Kim, J. K., Ashok, A., & Serre, T. (2020). Recurrent neural circuits for contours detection. *International Conference on Learning Representations (ICLR)*.

Linsley, D., Kim, J. K., Veerabadran, V., Windolf, C., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units <*https://papers.nips.cc/paper/2018/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf*>. *Neural Information Processing Systems (NIPS)*.

Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2018). Learning what and where to attend. arXiv preprint arXiv: 1805.08819.

Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2019). Learning what and where to attend. *International Conference on Learning Representations*.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.

Ma, W. J., & Peters, B. (2020, May 2). A neural network walks into a lab: Ttowards using deep nets as models for human behavior <*http://arxiv.org/abs/2005.02181*>. *arXiv [cs.AI]*.

Maheswaranathan, N., Kastner, D. B., Baccus, S. A., & Ganguli, S. (2018). Inferring hidden structure in multilayered neural circuits. *PLoS Computational Biology*, *14*(8), e1006291.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, *2*, 416–423.

Martinez-Trujillo, J. C., & Treue, S. (2005). The Feature Similarity Gain model of attention: Unifying multiplicative effects of spatial and feature-based attention. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 300–304). Academic Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725.

Morcos, A. S., Raghu, M., & Bengio, S. (2018, October 23). Insights on representational similarity in neural networks with canonical correlation <*http://arxiv.org/abs/1806.05759*>. *arXiv [stat.ML]*.

Mordvintsev, A., Pezzotti, N., Schubert, L., & Olah, C. (2018). Differentiable image parameterizations <https://distill.pub/2018/differentiable-parameterizations/>. *Distill*, *3*(7).

Mozafari, M., Reddy, L., & VanRullen, R. (2020, December 7). Reconstructing natural scenes from fMRI pPatterns using BigBiGAN <http://arxiv.org/abs/2001.11761>. *arXiv [cs.CV]*.

Murray, R. F. (2011). Classification images: A review <https://doi.org/10.1167/11.5.2>. *Journal of Vision*, *11*(5).

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4), e1003553.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *13*(1), 87–108.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020a). An overview of early vision in inceptionV1. *Distill*, *5*(4), e00024.002.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020b). Zoom iIn: An iIntroduction to cCircuits <https://doi.org/10.23915/distill.00024.001>. *Distill*, *5*(3).

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability <https://doi.org/10.23915/distill.00010>. *Distill*, *3*(3).

Oliva, A., Hospital, W., & Ave, L. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. *Frontiers in Psychology*, *4*(April), 1–14.

Ostrovsky, Y., Meyers, E., Ganesh, S., Mathur, U., & Sinha, P. (2009). Visual parsing after recovery from blindness. *Psychological Science*, *20*(12), 1484–1491.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America, A, Optics, Image & Science*, *10*(3), 405–411.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, *22*(9), 794–809.

O'Toole, B., & Wenderoth, P. (1977). The tilt illusion: Repulsion and attraction effects in the oblique meridian. *Vision Research*, *17*(3), 367–374.

Papernot, N., McDaniel, P., Goodfellow, I., & Jha, S. (2017). Practical black-box attacks against machine learning <https://dl.acm.org/doi/abs/10.1145/3052973.3053009?casa_token=-DwxUM3IaFoAAAAA:PlxG9S4GlTU9dwKfT4t8umYl1N1Sx5Z19NSJMiGLmJbinYKmU2q5knHARErGlxS-jlUr-ZGr8a4>. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 506–519.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations <https://doi.org/10.1111/cogs.12670>. *Cognitive Science*, *42*(8), 2648–2669.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6171–6176.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999–1009.

Pramod, R. T., & Arun, S. P. (2016). Do computational models differ systematically from human object perception? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1601–1609.

Qin, Z., Yu, F., Liu, C., & Chen, X. (2018, May 31). How convolutional neural network see the world: —A survey of convolutional neural network visualization methods <http://arxiv.org/abs/1804.11191>. *arXiv [cs.CV]*.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (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. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(33), 7255–7269.

Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(35), 12127–12136.

Raman, R., & Hosoya, H. (2020). Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. *Communications Biology*, *3*(1), 221.

Ricci, M., Cadene, R., & Serre, T. (2020). Same-different conceptualization: A machine vision perspective. *Current Opinion in Behavioral Sciences*, *37*, 47–55.

Ricci, M., & Serre, T. (2020). Hierarchical models of the visual system <https://doi.org/10.1007/978-1-4614-7320-6_345-2>. In D. Jaeger& R. Jung (Eds.), *Encyclopedia of computational neuroscience* (pp. 1–14). Springer.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (2000). The implementation of visual routines. *Vision Research*, *40*(10–12), 1385–1411.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rosenfeld, A., Solbach, M. D., & Tsotsos, J. K. (2018). *Totally looks like: -Hhow humans compare, compared to machines* <http://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w39/Rosenfeld_Totally_Looks_Like_CVPR_2018_paper.pdf>. CVPR Workshop Paper.

Rosenfeld, A., Zemel, R., & Tsotsos, J. K. (2018, August 9). The elephant in the room <*http://arxiv.org/abs/1808.03305*>. *arXiv [cs.CV]*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Saleh, B., Elgammal, A. M., Feldman, J., & Farhadi, A. (2016). Toward a taxonomy and computational models of abnormalities in images. *AAAI*, *30*, 3588–3596.

Scholl, B., Tan, A. Y. Y., Corey, J., & Priebe, N. J. (2013). Emergence of orientation selectivity in the mammalian visual pathway. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(26), 10616–10624.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018, January 2). Brain-Scorescore: Which artificial neural network for object recognition is most brain-like? <*https://doi.org/10.1101/407007*>*bioRxiv*.

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, *13*(5), 402–409.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. A. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785.

Serre, T. (2016). Models of visual categorization. *Wiley Interdisciplinary Reviews. Cognitive Science*, *7*(3), 197–213.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, *5*, 399–426.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, *165*, 33–56.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6424–6429.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013, December 20). Deep inside convolutional networks: Visualising image classification models and saliency maps <*http://arxiv.org/abs/1312.6034v1*>. *arXiv [cs.CV]*.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., & Charest, I. (2020, March 26). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition <*https://www.biorxiv.org/content/10.1101/677237v4.abstract*>. *bioRxiv*.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020, May 8). Diverse deep neural networks all predict human IT well, after training and fitting <*https://doi.org/10.1101/2020.05.07.082743*>. *bioRxiv*.

Sturmfels, P., Lundberg, S., & Lee, S.-I. (2020). Visualizing the impact of feature attribution baselines *<https://doi.org/10.23915/distill.00022>*. *Distill*, *5*(1).

Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S., & Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: The structure of retinal prediction. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*, Vol. 32 (pp. 8537–8547). Curran Associates.

Tang, H., Buia, C., Madhavan, R., Crone, N. E., Madsen, J. R., Anderson, W. S., & Kreiman, G. (2014). Spatiotemporal dynamics underlying object completion in human ventral visual cortex *<https://doi.org/10.1016/j.neuron.2014.06.017>*. *Neuron*, *83*(3), 736–748.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., & Kreiman, G. (2018). Recurrent computations for visual pattern completion *<https://doi.org/10.1073/pnas.1719397115>*. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(35), 8835–8840.

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, *9*(4), 483–484.

Thorat, S., van Gerven, M., & Peelen, M. (2019, March 25). The functional role of cue-driven feature-based feedback in object recognition *<http://arxiv.org/abs/1903.10446>*. *arXiv [q-bio.NC]*.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, 4489–4497.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2744–2749.

Ullman, S., & Soloviev, S. (1999). Computation of pattern invariance in brain-like structures. *Neural Networks*, *12*, 1021–1036.

Van Den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, *48*, 1747–1756.

VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology/University of Finance and Management in Warsaw*, *3*(1–2), 167–176.

Volokitin, A., Roig, G., & Poggio, T. (2017, June 26). Do deep neural networks suffer from crowding?[*<http://arxiv.org/abs/1706.08616>arXiv [cs.CV]*.

Wang, J., Xie, C., Zhang, Z., Zhu, J., Xie, L., & Yuille, A. (2017, June 25). Detecting semantic parts on partially occluded objects *<http://arxiv.org/abs/1707.07819>*. *arXiv [cs.CV]*.

Wang, J., Zhang, Z., Xie, C., Zhou, Y., Premachandran, V., Zhu, J., Xie, L., & Yuille, A. (2017, November 13). Visual concepts and compositional voting *<http://arxiv.org/abs/1711.04451>*. *arXiv [cs.CV]*.

Wang, J., Zhang, Z., Xie, C., Zhou, Y., Premachandran, V., Zhu, J., Xie, L., & Yuille, A. (2018). Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications*, *3*(1), 151–188.

Wardle, S. G., & Baker, C. (2020). Recent advances in understanding object recognition in the human brain: Ddeep neural networks, temporal dynamics, and context <https://doi.org/10.12688/f1000research.22296.1>. *F1000Research*, *9*.

Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, *5*, 674.

Xiao, M., Kortylewski, A., Wu, R., Qiao, S., Shen, W., & Yuille, A. (2020). TDMPNet: Prototype network with recurrent top-down modulation for robust object classification under partial occlusion <https://openreview.net/pdf/74e55053fd95550993d69adae7a324ec49e9bdf9.pdf> [Paper presented]. European Conference on Computer Vision: —1st Visual Inductive Priors for Data-Efficient Deep Learning Workshop.

Xu, T., Garrod, O., Ince, R., & Schyns, P. (2019). Using psychophysics to reveal face identification information processing mechanisms in a deep neural network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1976–1984

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624.

Zamir, A. R., Wu, T., Sun, L., Shen, W. B., Shi, B. E., Malik, J., & Savarese, S. (2017). Feedback networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1808–1817.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision—ECCV 2014*, 818–833.

Related Articles

Implicit Memory and Cognitive Aging

The Group Dynamics of Interorganizational Collaboration

Brain Development