Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Deep neural network models of sensory systems: windows onto the role of task constraints

Alexander JE Kell[1,2] and Josh H McDermott[1,2,3,4]

Sensory neuroscience aims to build models that predict neural responses and perceptual behaviors, and that provide insight into the principles that give rise to them. For decades, artificial neural networks trained to perform perceptual tasks have attracted interest as potential models of neural computation. Only recently, however, have such systems begun to perform at human levels on some real-world tasks. The recent engineering successes of deep learning have led to renewed interest in artificial neural networks as models of the brain. Here we review applications of deep learning to sensory neuroscience, discussing potential limitations and future directions. We highlight the potential uses of deep neural networks to reveal how task performance may constrain neural systems and behavior. In particular, we consider how task-optimized networks can generate hypotheses about neural representations and functional organization in ways that are analogous to traditional ideal observer models.

Addresses
[1] Department of Brain and Cognitive Sciences, MIT, United States
[2] Center for Brains, Minds, and Machines, United States
[3] McGovern Institute for Brain Research, MIT, United States
[4] Program in Speech and Hearing Biosciences and Technology, Harvard University, United States

Corresponding authors: Kell, Alexander JE (alexkell@mit.edu), McDermott, Josh H (jhm@mit.edu)

Current Opinion in Neurobiology 2019, **55**:121–132

This review comes from a themed issue on **Machine learning, big data, and neuroscience**

Edited by **Maneesh Sahani** and **Jonathan Pillow**

https://doi.org/10.1016/j.conb.2019.02.003

Introduction

A longstanding goal of sensory neuroscience is to build models that reproduce behavioral and neural responses. Models have historically originated from a range of sources, including experimental observation [1–5], a combination of biological inspiration and engineering principles [6–9], and normative criteria (e.g. efficient coding) applied to representations of natural sensory signals [10–15].

Models have also been inspired by the idea that they should be able to perform tasks that organisms perform. One use of tasks is to derive ideal observer models — models that perform a task optimally under certain assumptions [16]. Such models provide hypotheses for biological systems based on the notion that biological systems may be near-optimal for ecologically important tasks. Behavioral predictions from ideal observer models can also provide normative explanations of otherwise puzzling perceptual phenomena, for instance by showing how 'illusions' can be viewed as optimal inferences given the statistics of the natural world [17].

Ideal observer models are provably optimal, but they are typically derived analytically and are often restricted to relatively simple domains where the task structure can be precisely specified. An alternative approach is to learn solutions to tasks from data. Supervised learning approaches take a set of input-output pairs (e.g. images and object labels or sounds and word labels) and modify a system's parameters to minimize the error between the system's output and the desired output. The resulting models are usually not provably optimal because the task is specified with training data — generalization performance must be estimated empirically rather than derived analytically. However, supervised learning allows models to be constructed for a wide range of tasks, including some that organisms perform in their everyday environments (for which the derivation of ideal observed models may be intractable).

Supervised learning approaches were adopted in neurally inspired models as early as the 1960s [18]. They were then adapted to multi-layer networks in the 1980s, and the resulting wave of neural network research led to optimism that learned representations could be used to generate hypothesis about actual neural computation [19–21]. However, neural network models at the time were limited to relatively small-scale tasks and networks. The advent of inexpensive GPU-based computing along with assorted technical advances [22–24] led to a resurgence of interest in neural networks in the engineering world in the 2010s. For the first time, computing systems attained human levels of performance on a handful of challenging classification tasks in vision and in speech recognition [25,26]. These successes caused many neuroscientists to reassess the relevance of such networks for the brain. In this paper, we discuss the recent developments in this domain along with reasons for skepticism.

Deep neural networks

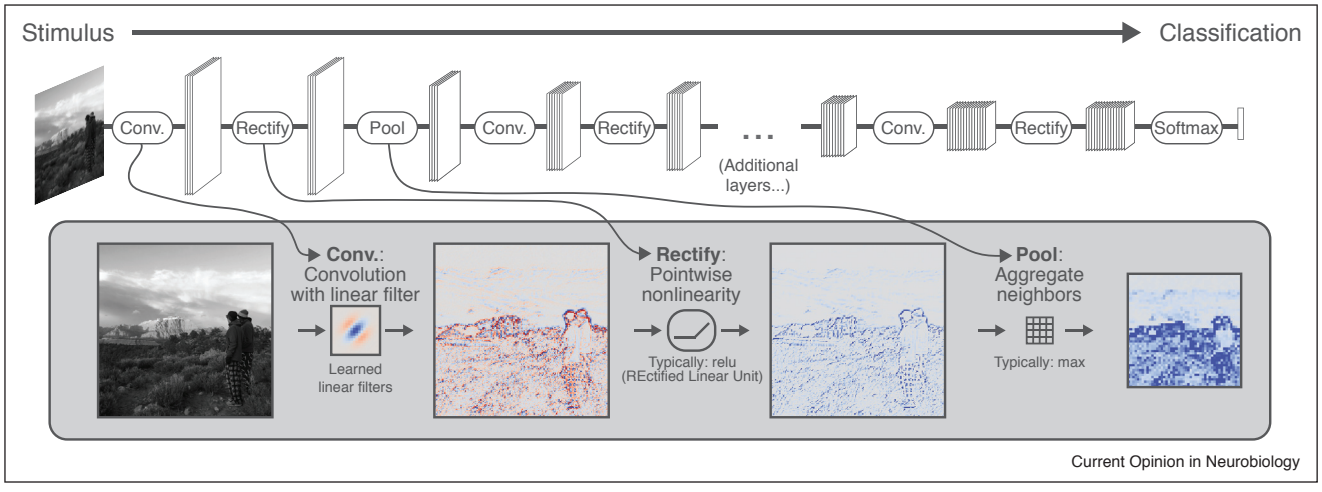
Artificial neural networks consist of sets of units with connections defined by weights. The units and weights are loosely modeled on neurons and synaptic efficacies, respectively. A unit's activation is computed by multiplying its inputs (the activations of other units) by the associated weights, summing the results, and passing the sum through a simple pointwise nonlinear function (e.g. a sigmoid or, more commonly in recent years, a rectifying function [22]). The input is usually some sort of sensory signal (e.g. an image, sound waveform, or spectrogram) and the output units are interpreted as probabilities of target classes (e.g. digits, object identities, or phonemes). Because the output activations are differentiable functions of the network weights, the weights can be adjusted via gradient descent to cause the output activations to approach target values [27]. Given a training set of signals and class labels, a network can thus be optimized to minimize classification errors.

The most recent wave of neural networks add a few more ingredients to this broader recipe (Figure 1). The first is that the weights for subsets of units in a particular layer are often constrained to implement convolution operations with a filter that is small relative to the input dimensionality [28]. Units in a layer, therefore, apply the same dot-product operation at different locations in a signal, analogous to similarly structured visual receptive fields at different retinotopic locations. A single layer of a deep network will often implement dozens or hundreds of such filters. The second ingredient is the incorporation of pooling operations, in which the responses of nearby units are aggregated in some way. Pooling operations downsample the preceding representation, and thus can be related to classical signal processing, but were also in part inspired by 'complex' cells in primary visual cortex (that are thought to combine input from multiple 'simple' cells) [8,29]. Convolution and pooling were both introduced to artificial neural networks several decades ago [28], but have become widely used in the last decade. Recent networks have begun to incorporate additional architectural motifs, such as 'skip' and 'residual' connections that violate feedforward organization in various ways [30,31].

Each of the operations is defined by hyperparameters that specify the network architecture, including the filter size, the pooling region size, the pooling operation (e.g. taking the maximum value within the pooling region), and the order of operations. The cascade of these operations instantiate sets of progressively more complex features through the course of the network. If the network is appropriately optimized through the selection of hyperparameters and via gradient descent on the network weights, it may achieve good performance on the task on which it was trained.

What might one learn about the brain from such a system? The structure of an artificial neural network can in some cases be mapped in a loose sense onto the structure of

Figure 1



Schematic of a typical deep convolutional neural network.
The stimulus (e.g. an image for a visual task or a spectrogram for auditory task) is passed through a cascade of simple operations, in which the output of one stage of operations is the input to the next. This cascade culminates in a discriminative classification (e.g. of the object category present in the image, or the spoken word present in the sound signal). Because of downsampling, units in later layer have access to a greater portion of the stimulus (i.e. a larger 'receptive field'). Concurrently, the feature maps (represented in the schematic by the stacked panels at each stage) tend to decrease in size at deeper network stages, again due to the downsampling that happens over the course of the network. The number of feature maps per stage is typically made to increase at deeper network stages, yielding a greater diversity of unit response properties. Bottom: Insets of schematics of typical operations, including convolution with a linear filter (left), a pointwise nonlinearity such as rectification (center), and pooling over a local neighborhood (right), with their effect illustrated on an example image.

sensory systems, which are also often conceptualized as a sequence of hierarchically organized distributed stages. It is thus natural to wonder whether an artificial network trained on an ecologically important task might exhibit representations like those in biological sensory systems, offering hypotheses about their inner workings. On the other hand, although modern-day DNNs produce remarkable levels of task performance, they differ in many respects from actual neural circuits. Moreover, the means by which they achieve good performance is often resistant to interpretation. Here we will review recent work comparing trained DNNs to brain and behavior data, and we will consider what we can learn from such comparisons.

Behavioral and brain responses predicted by deep neural networks

One of the main motivations for considering deep neural networks as models of perceptual systems is that they attain (or exceed) human-level performance on some object and speech recognition tasks. But for DNNs to serve as models of biological sensory systems, they should arguably also match detailed patterns of performance. There are now several demonstrations of similar performance characteristics for human observers and DNNs. The most comprehensive comparisons have occurred for visual object recognition, where DNNs trained to recognize objects match human error patterns across object categories [32–34] and viewpoint variations [35], exhibit similar sensitivity to object shape [36], and predict object similarity judgments [37] (Figure 2a). Despite the similarity with human perception when analyzed in terms of object categories, fine-grained discrepancies are evident. In the one case where it has been measured, behavioral similarity breaks down somewhat at the image-by-image level – humans and deep networks make errors on different images (Figure 2a) [34]. Some of these discrepancies may reflect algorithmic differences. For instance, deep networks may rely more on texture to classify images than humans do [38–40]. Nonetheless, at the level of object categories, the similarity in behavioral recognition is strong. Such similarities appear in the auditory domain as well, where speech recognition performance in different types of background noise is likewise highly correlated across humans and a trained DNN [41••] (Figure 2b). Notably, the network models in these cases are not fit to best match human behavior – they are optimized only to perform visual or auditory tasks. The similarities to human behavior arise simply as a consequence of learning to perform the task.

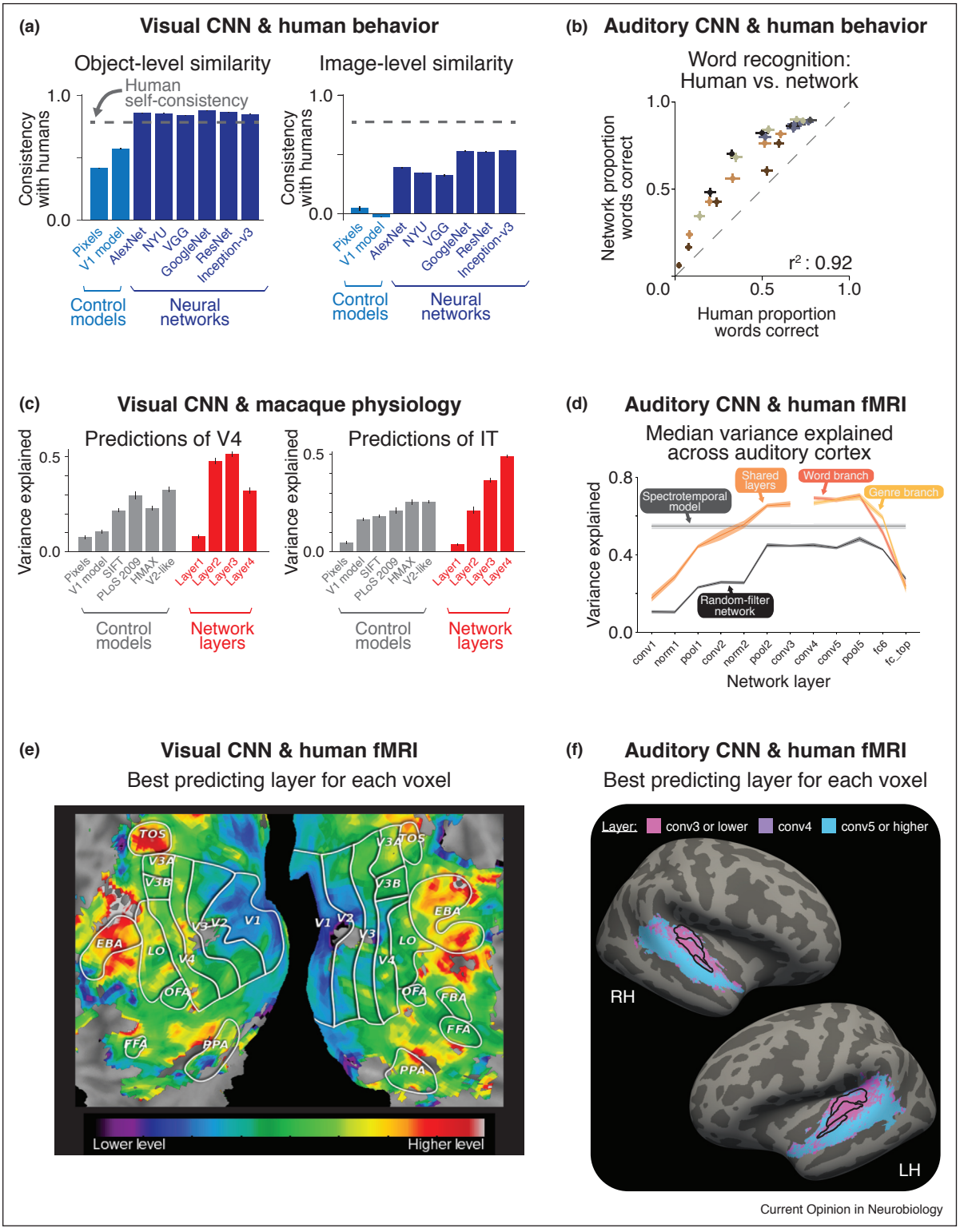
What do these behavioral similarities reveal? One possibility is that they simply reflect the limits of optimal performance, such that any system attaining human levels of overall performance would exhibit performance characteristics resembling those of humans. It is also possible that the behavioral similarity depends on similarity in the internal representational transformations instantiated by the DNN and human sensory systems. This second possibility would imply that alternative systems could produce comparable overall task performance but exhibit detailed performance characteristics distinct from those of humans. These possibilities are difficult to distinguish at present given that we lack alternative model classes that produce human-level performance on real-world classification tasks.

Regardless of the interpretation, the observed behavioral similarities between DNN models and humans motivate comparisons of their internal processing stages. A natural means of comparison is to test how well the features learned by a network can be used to predict brain responses. Although deep learning has also been used to directly optimize models to predict empirically measured responses [42–45], the amount of neural data needed to constrain a complex model may limit the extent to which models can be built entirely from the constraints of predicting neural responses. Here, we focus instead on the use of neural predictions to evaluate DNN models whose structure is determined exclusively by task optimization. The most visible applications of deep neural networks to neuroscience have come from efforts along these lines to predict neural responses in the ventral visual stream. Before the advent of high-performing DNNs, models of sensory systems were able to account for neural responses of early stages of sensory processing reasonably well [2,5], but were less successful for intermediate or higher-level cortical stages.

Deep neural networks optimized to classify images of objects provided the first models that could generate good predictions of neural responses in high-level sensory areas. One standard approach is to model the responses of individual neurons, or of voxels measured with fMRI, with linear combinations of the features from a particular layer of a trained neural network [46,47]. The weights of the linear mapping are fit to best predict responses to a subset of stimuli, and the quality of the fit is evaluated by comparing actual and predicted responses to left-out stimuli [48,49]. When evaluated in this way, DNN models provide far better predictions of responses in inferotemporal cortex than any previous model [50•,51–53] (Figure 2c), as well as better predictions in early visual areas [45,53]. Alternative types of brain-model comparisons, such as representational similarity analysis [54], also find that DNN models best replicate the representational structure evident in brain measurements from IT [55•,56]. This success is not limited to the visual system — DNNs optimized for speech and music recognition tasks also produce better predictions of responses in auditory cortex than previous models [41••] (Figure 2d).

The ability of DNN features to generate good predictions of neural responses raises questions about the purpose of

Figure 2



Task-optimized deep neural networks predict visual and auditory cortical responses and recapitulate real-world behavior.
(a) Deep networks exhibit human-like errors at the scale of visual object categories (left), but not at the scale of single images (right). *Y*-axis plots the consistency of the network's performance with that of humans, quantified with a modified correlation coefficient (see original paper for details in Ref. [34]). Dashed gray indicates the noise ceiling (the test–retest consistency of the human data). Each bar plots the consistency for a different model. Light blue bars are for control models: linear classifiers operating on a pixel array or a standard model of visual area V1 [102]. Dark blue

the modeling enterprise. Although DNNs predict neural responses, their inner workings are typically difficult to describe or characterize, at least at the level of individual units. However, DNNs can have well-defined structure at the scale of layers: in 'feedforward' networks, each stage of processing provides the input to the next, such that successive stages instantiate compositions of increasing numbers of operations. When trained, this hierarchical structure appears to recapitulate aspects of hierarchical structure in the brain. Early stages of the ventral visual stream (V1) are well predicted by early layers of DNNs optimized for visual object recognition [45,52,53], whereas intermediate stages (V4) are best predicted by intermediate layers, and late stages (IT) best predicted by late layers [50•,51–53] (Figure 2c and e). This result is consistent with the idea that the hierarchical stages of the ventral stream result from the constraints imposed by biological vision tasks.

The organization of the ventral visual stream into stages was uncontroversial before this modeling work was done, and these results thus largely provide a validation of the idea that a task-optimized hierarchical model can replicate aspects of hierarchical organization in biological sensory systems. However, they raise the possibility that one use of DNN models could be to probe for hierarchical organization in domains where it is not yet well established. We recently adopted this approach in the auditory system, showing that intermediate layers of a DNN optimized for speech and music recognition best predicted fMRI voxel responses around primary auditory cortex, whereas deeper layers best predicted voxel responses in non-primary cortex [41••] (Figure 2f). This result was not merely a reflection of the scale of the features computed at different network stages: networks with identical architectures but random (untrained) weights did not produce this correspondence between cortical regions and network layers. The results provided evidence for a division of the auditory cortex into at least two stages, with one stage potentially providing input into the next.

Deep networks have recently also been employed in analogous fashion in other domains, including the somatosensory system [57], as well as the hippocampal and entorhinal systems of the medial temporal lobe [58–60].

Using deep learning to reveal how tasks constrain neural systems and behavior

Because deep learning provides a means to optimize systems for some real-world tasks, it may hold promise for understanding the role of such tasks in shaping neural systems and behavior. Specifically, deep neural networks may be useful as stand-ins for ideal observer models in domains for which an actual ideal observer is either intractable to derive analytically, or unknowable (i.e. in cases where the task structure is not well understood in theoretical terms). Like ideal observers, deep networks may help reveal how task constraints shape brains and behavior, but could enable such insights for a larger range of tasks.

In one recent example that illustrates this potential, a neural network was trained to perform a simple visual search task using a 'retinal' receptor lattice [61••]. This lattice could be translated across an input image, in much the same way that saccadic eye movements shift an image across the retina. Each receptor on the lattice was parameterized by its position and spread, and these parameters were optimized during training along with the rest of the network. The result of the optimization procedure was a receptor lattice that qualitatively replicated the organization of the primate retina, with a high resolution 'fovea' surrounded by a low resolution periphery (Figure 3a). Notably, this result did not occur when the system was allowed to use additional actions, like 'zooming', that are not present in the primate visual system. These results are consistent with the possibility that the arrangement of receptors on the retina may result from an evolutionary optimization of the sampling of the visual world conditioned on the use of eye movements.

(Figure 2 Legend Continued) bars are for various artificial neural networks: AlexNet [25], NYU [103], VGG [104], GoogLeNet [105], Resnet [30], and Inception-v3 [106]. From Rajalingham *et al.* [34].

(b) Speech recognition by deep networks and humans are similarly affected by background noise. *X*-axis plots human performance and *Y*-axis plots network performance. Each point represents speech recognition performance in a particular type of background noise at a particular SNR. From Kell *et al.* [41••].

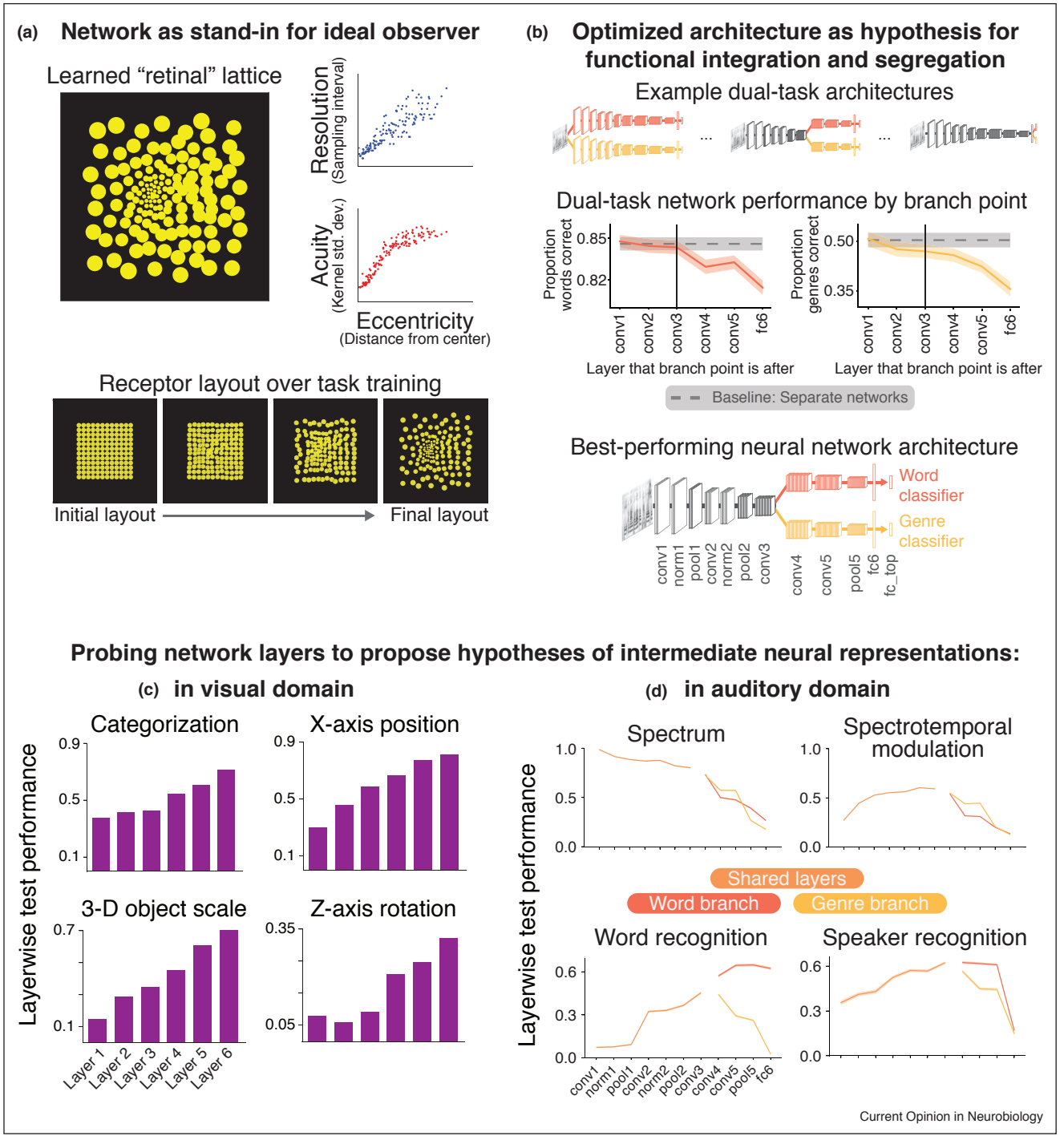
(c) Deep networks predict multi-unit neuronal activity recorded from macaque visual areas V4 (left) and IT (right) better than comparison models. *Y*-axis plots cross-validated prediction accuracy. Gray bars plot results for control models: linear classifiers operating on pixel arrays, a model of visual area V1 [102], SIFT features [107], an untrained neural network [108], HMAX [109], and a set of V2-like features [110]. Red bars are generated from different layers of a trained neural network (the HMO model from Ref. [50•]). Intermediate network layers best predict intermediate visual area V4, while later layers best predict later visual area IT. From Yamins *et al.* [50•].

(d) Response prediction accuracy of an audio-trained DNN used to predict responses to natural sounds. A deep network trained to recognize words and musical genres predicted fMRI responses in auditory cortex better than a baseline spectrotemporal filter model [9] (gray line). *Y*-axis plots prediction accuracy for different network layers (displayed along the *X*-axis). From Kell *et al.* [41••].

(e) Map of the best-predicting DNN layer across human visual cortex. Human fMRI responses in early and late stages of the visual cortical hierarchy are best predicted by early and late network layers, respectively. White outlines indicate functionally localized regions of interest: retinotopic visual areas (V1, V2, V3, V3A, V3B, V4), transverse occipital sulcus (TOS), parahippocampal place area (PPA), extrastriate body area (EBA), occipital face area (OFA), and fusiform face area (FFA). From Eickenberg *et al.* [53].

(f) Map of the best-predicting DNN layer across human auditory cortex. Black outlines denote anatomical parcellations of primary auditory cortex. Early and intermediate layers best predict primary auditory cortical responses; later layers best predict non-primary auditory cortical responses. From Kell *et al.* [41••].

Figure 3



Neural networks as hypothesis generators for neuroscience.
(a) A neural network optimized to identify digits in a cluttered visual scene learns a retinal-like lattice with fine acuity within a 'fovea' and decreased acuity in the periphery. Left: resulting lattice; circles indicate pooling regions of individual receptors. Right: Resolution (top) and acuity (bottom) as a function of distance from center of lattice. Bottom: Receptor layout over training. From Cheung *et al.* [61••].
(b) Branched neural networks used to generate hypotheses about functional segregation and integration in the brain. Top: Example dual-task architectures, ranging from one with two totally separate pathways on the left to an entirely shared single pathway on the right. Middle: Performance on word recognition (left) and musical genre recognition (right) tasks as a function of number of shared stages. Bottom: Resulting network architecture that shares as much processing as possible without producing a performance decrement. From Kell *et al.* [41••].
(c) Hypotheses for intermediate stages of neural computation generated from decoding. The decoding of a variety of category-orthogonal variables (horizontal position, object scale, *Z*-axis rotation) improves as one moves deeper into a network trained to recognize visual object

Task-optimized neural networks have also been used to understand perceptual learning experiments in which participants are trained on psychophysical tasks (e.g. orientation discrimination) [62,63]. Deep networks trained on the same tasks used in laboratory experiments have been shown to recapitulate a diverse set of neurophysiological and psychophysical findings. For instance, some training tasks yield changes at either earlier or later stages of sensory processing, and similar changes occur in deep networks trained on these tasks. The precision of the training task also alters network generalization to new stimuli in ways that match results in humans. The results suggest that the outcomes of perceptual learning experiments can be understood as the consequences of optimizing representations for tasks, even though the mechanisms that instantiate learning in DNNs are likely to be different than those in humans (see 'Limitations and Caveats' section below).

Deep learning has also been used to explore how visual attention mechanisms may affect task performance [64••]. The 'feature similarity gain' model of visual attention proposes that attention scales a neuron's activity in proportion to its preference for the attended stimulus [65]. To test this theory, this type of scaling was applied to unit activations from a deep neural network optimized to classify visual objects [64••]. The authors found that the scaling led to behavioral performance improvements similar to those previously observed psychophysically under conditions of directed attention. However, this result was only observed at later layers of the network — applying the scaling to early and intermediate network layers did not produce comparable behavioral differences. This result illustrates how deep neural networks can provide hypotheses about the effect of internal representational changes on behavioral performance.

Using optimized networks as stand-ins for ideal observers may also reveal normative constraints on the integration and segregation of function in sensory systems. One approach is to train a single system to perform multiple tasks, and to examine the amount of processing that can be shared without producing a detriment in task performance relative to that obtained with a single-task system. The resulting model offers a hypothesis for how a sensory system may be functionally organized, under the assumption that sensory systems evolve or develop to perform well subject to a resource constraint (e.g. the number of neurons). We recently employed this approach to examine the extent to which speech and music processing might be expected to functionally segregate in the brain [41••]. We found that a network jointly optimized for speech and music recognition could share roughly the first half of its processing stages across tasks without seeing a performance decrement (Figure 3b). This result was consistent with fMRI evidence for segregated pathways for music and speech processing in non-primary auditory cortex [66], and suggested a computational justification for this organization. The methodology could be more broadly applied to address current controversies over domain specificity and functional segregation [67,68].

Another potential application of deep neural networks is to suggest hypotheses for intermediate sensory representations. Intermediate sensory stages have long posed a challenge for sensory neuroscience because they are often too nonlinear for linear systems tools to be applicable, and yet too distant from task read-out for neural tuning to directly reflect behaviorally relevant variables. Model-driven hypotheses of intermediate stages could thus be particularly useful. Individual units of deep networks are typically challenging to interpret, but could become more accessible with new developments in visualization [69–72], or from constraints on models that may aid interpretability, such as cost functions that bias units within a layer to be independent [73,74].

Alternatively, insight into intermediate representations might be best generated at the population level, by assessing the types of information that can be easily extracted from different stages of a network. A standard approach is to train linear classifiers on a layer's activations, and then measure performance on a validation set. One recent application of this methodology tested whether invariance to object position is a prerequisite for object recognition. In DNNs trained to categorize visual objects, later layers provided better estimates than earlier layers of various 'category-orthogonal' variables, such as the position of an object within an image or its overall scale [75] (Figure 3c). Notably, a similar pattern of results was found in the primate visual system, with position and scale more accurately decoded from IT than V4 [75]. Decoding also reveals biologically relevant representational transformations in audio-trained networks. For instance, in a DNN trained to recognize spoken words and musical genres, the frequency spectrum of a sound was best estimated from the earliest layers, whereas spectrotemporal modulations were best estimated from intermediate layers [41••], consistent with their hypothesized role in primary auditory cortex [9,76] (Figure 3d).

(Figure 3 Legend Continued) categories. From Hong *et al.* [75].
(d) Different stimulus properties are best decoded from different layers of a network trained to recognize words and musical genre. Top left: Decoding of the spectrum peaks early. Top right: Decoding of spectrotemporal modulation power peaks in intermediate layers. Bottom right: Word recognition performance increases over the course of the network for the task-relevant branch, but decreases in task-irrelevant (genre) branch. Bottom left: Decoding of a task-irrelevant feature (speaker identity) peaks in late-to-intermediate layers. From Kell *et al.* [41••].

Limitations and caveats

The renaissance of deep neural networks in neuroscience has been accompanied by skepticism regarding the extent to which DNNs could be relevant to the brain. Most obviously, current DNNs are at best loosely analogous to actual neural circuits, and so at present do not provide circuit-level models of neural computation. These limitations alone render them inappropriate for many purposes. Moreover, if the details of neural circuitry place strong constraints on neural representations and behavior, DNNs could be limited in their ability to predict even relatively coarse-scale phenomena like neural firing rates and behavior.

Some of the discrepancies between artificial neural networks and human sensory systems can be addressed with modifications to standard DNN architectures. For instance, recent work has incorporated recurrent connections to the feedforward neural networks often used to model the ventral visual pathway [77]. Such recurrent connections may be important for predicting responses to natural images that are not well accounted for by feedforward models [78••], including those with occlusion [79]. However, it is less obvious how to incorporate other aspects of biological neural circuits, even those as fundamental as action potentials and neuromodulatory effects [80–83].

As it currently stands, deep learning is also clearly not an account of biological learning. Most obviously, biological organisms do not require the millions of labeled examples needed to train contemporary deep networks. Moreover, whatever learning algorithms are employed by the brain may not have much similarity to the standard backpropagation algorithm [84•,85], which is conventionally considered biologically implausible for a variety of reasons (e.g. the need to access the weights used for feedforward computation in order to compute learning updates).

Another challenge for the general notion that task-driven training can reveal neural computation is that as DNN systems have increased in size, they have begun to exceed human levels of performance, at least on particular computer vision tasks. Moreover, neural predictions from these very high-performing networks has plateaued or even declined in accuracy, as if the networks have begun to diverge from biologically relevant solutions [86]. This divergence could reflect differences between the specific tasks used to optimize current DNNs and those that may have constrained biological systems over the course of evolution and development. Alternatively, additional constraints could be needed to obtain brain-like systems under task optimization. Possibilities include a resource limitation (e.g. on the number of neurons or on metabolic activity) or constraints imposed by the historical trajectory of the brain's evolution.

Some of the differences between DNNs and human observers may be due to violations of traditional signal processing principles by DNNs. The sampling theorem dictates that if signals are not lowpass filtered before downsampling, they will be 'aliased' — low frequencies will be corrupted by high frequencies present in the signal before downsampling. Because contemporary deep networks typically employ downsampling operations (max pooling and/or strided convolution) without the constraint of a preceding lowpass filter, aliasing is likely to occur [87,88••]. It is perhaps remarkable that aliasing apparently does not prevent good classification performance, but it may impair generalization [88••] and produce representations that diverge from those of biological systems [89].

One example of such divergences can be found in demonstrations that DNNs can be fooled by 'adversarial' stimuli [90,91]. These stimuli are derived by using the gradients of the output units of a network with respect to its input to generate small perturbations to an input signal that cause it to be misclassified. In principle, such adversarial stimuli could be generated for a human perceptual system if one had the complete description of the system necessary to derive the perturbations — obviously beyond reach for the moment. But if the network were a correct description of a biological perceptual system, then its adversarial stimuli should also be perceived differently by humans. In practice, the perturbations generated in this way for high-performing DNNs are typically imperceptible to humans (though in some cases humans exhibit some sensitivity to such perturbations [92]). One potential explanation could be that the exact perturbations needed to produce this effect depend on minor idiosyncrasies of a model, such that adversarial perturbations for one system would not generalize to other systems. However, adversarial examples tend to have similar effects on networks trained from different initial conditions, and with different architectures, suggesting there may be a more fundamental and consistent difference with biological systems. Notably, adversarial images are not specific to DNNs — they are observed even for linear classifiers [91]. One speculative possibility is that they may reveal a limit of models exclusively trained on classification tasks [93].

The most fundamental difference between current DNNs and human perceptual systems may lie in the relative inflexibility of artificial networks — a trained network is typically limited to performing the tasks on which it is trained. Representations learned for one task can transfer to others [75,94,95], but usually require training a new classifier with many new training examples. This rigidity seems at odds with the fact that humans can answer a wide range of queries when presented with a novel auditory or visual scene, even questions that they may not have ever previously been asked [96]. Observations along these lines have led some to suggest that humans have an internal model of the world, and infer generative parameters of this model

when presented with a stimulus, allowing them to perform a wide range of tasks [97].

Many of these limitations could be addressed by combining DNNs with generative models of how structures in the world give rise to sensory data. Such internal models could in principle explain the flexibility of our perceptual abilities, but inferring the parameters needed to explain a stimulus is often hugely computationally expensive. One appealing idea is to leverage DNNs to generate initial estimates of generative variables that can accelerate inference — given a generative model, a DNN can be trained to map samples (e.g. images) to their underlying parameters (e.g. 3D shape descriptors) [98,99•]. This approach raises the question of how the generative model itself would be acquired, but in principle a feedforward recognition network could be jointly trained in parallel with a generative model [100,101]. Such marriages are appealing directions to explore, both for next-generation AI systems and models of biological perception.

Conflict of interest statement

Nothing declared.

Acknowledgements

The authors thank Jenelle Feather, Andrew Francl, and Rishi Rajalingham for comments on the manuscript, and Brian Cheung, Rishi Rajalingham, Bertrand Thirion, and Dan Yamins for contributions to subpanels of Figures 2 and 3. This work was supported by a Department of Energy Computational Science Graduate Fellowship (DE-FG02-97ER25308) to A.J.E.K., a McDonnell Scholar Award to J.H.M., and National Science Foundation grant BCS-1634050.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Heeger DJ: **Normalization of cell responses in cat striate cortex**. *Vis Neurosci* 1992, **9**:181-197.
2. Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL: **Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli**. *Network* 2001, **12**:289-316.
3. Pillow JW, Paninski L, Uzzell VJ, Simoncelli EP, Chichilnisky EJ: **Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model**. *J Neurosci* 2005, **25**:11003-11013.
4. Rust NC, Mante V, Simoncelli EP, Movshon JA: **How MT cells analyze the motion of visual patterns**. *Nat Neurosci* 2006, **9**:1421-1431.
5. David SV, Mesgarani N, Fritz JB, Shamma SA: **Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli**. *J Neurosci* 2009, **29**:3374-3386.
6. Adelson EH, Bergen JR: **Spatiotemporal energy models for the perception of motion**. *J Opt Soc Am A* 1985, **2**:284-299.
7. Dau T, Kollmeier B, Kohlrausch A: **Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers**. *J Acoust Soc Am* 1997, **102**:2892-2905.
8. Riesenhuber M, Poggio T: **Hierarchical models of object recognition in cortex**. *Nat Neurosci* 1999, **2**:1019-1025.
9. Chi T, Ru P, Shamma SA: **Multiresolution spectrotemporal analysis of complex sounds**. *J Acoust Soc Am* 2005, **118**:887-906.
10. Olshausen BA, Field DJ: **Emergence of simple-cell receptive field properties by learning a sparse code for natural images**. *Nature* 1996, **381**:607-609.
11. Schwartz O, Simoncelli EP: **Natural signal statistics and sensory gain control**. *Nat Neurosci* 2001, **4**:819-825.
12. Smith EC, Lewicki MS: **Efficient auditory coding**. *Nature* 2006, **439**:978-982.
13. Karklin Y, Lewicki MS: **Emergence of complex cell properties by learning to generalize in natural scenes**. *Nature* 2009, **457**:83-86.
14. Carlson NL, Ming VL, DeWeese MR: **Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus**. *PLoS Comput Biol* 2012, **8**:e1002594.
15. Mlynarski W, McDermott JH: **Learning mid-level auditory codes from natural sound statistics**. *Neural Comput* 2018, **30**:631-669.
16. Geisler WS: **Contributions of ideal observer theory to vision research**. *Vis Res* 2011, **51**:771-781.
17. Weiss Y, Simoncelli EP, Adelson EH: **Motion illusions as optimal percepts**. *Nat Neurosci* 2002, **5**:598-604.
18. Rosenblatt F: **The perceptron: a probabilistic model for information storage and organization in the brain**. *Psychol Rev* 1958, **65**:386-408.
19. Rumelhart D, McClelland J: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press; 1986.
20. Lehky SR, Sejnowski TJ: **Network model of shape-from-shading: neural function arises from both receptive and projective fields**. *Nature* 1988, **333**:452-454.
21. Zipser D, Andersen RA: **A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons**. *Nature* 1988, **331**:679-684.
22. Nair V, Hinton GE: **Rectified linear units improve restricted Boltzmann machines**. *27th International Conference on Machine Learning*. 2010.
23. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R: **Dropout: a simple way to prevent neural networks from overfitting**. *J Mach Learn Res* 2014, **15**:1929-1958.
24. Ioffe S, Szegedy C: **Batch normalization: accelerating deep network training by reducing internal covariate shift**. *arXiv* 2015. 1502.03167.
25. Krizhevsky A, Sutskever I, Hinton G: **ImageNet classification with deep convolutional neural networks**. *Advances in Neural Information Processing Systems*. 2012.
26. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN *et al.*: **Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups**. *IEEE Signal Process Mag* 2012, **29**:82-97.
27. Rumelhart DE, Hinton GE, Williams RJ: **Learning representations by back-propagating errors**. *Nature* 1986, **323**:533-536.
28. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD: **Handwritten digit recognition with a back-propagation network**. In *Advances in Neural Information Processing (NIPS 1989)*, , vol 2. Edited by Touretsky D. Denver, CO: Morgan Kauffman; 1990.
29. Hubel DH, Wiesel TN: **Receptive fields, binocular interaction and functional architecture in the cat's visual cortex**. *J Physiol* 1962, **160**:106-154.
30. He K, Zhang X, Ren S, Sun J: **Deep residual learning for image recognition**. *The IEEE Conference on Computer Vision and Pattern Recognition*. 2016:770-778.
31. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ: **Deep connected convolutional networks**. *2017 IEEE Conference on Pattern Recognition and Computer Vision (CVPR)*. 2017:4700-4708.

32. Rajalingham R, Schmidt K, DiCarlo JJ: **Comparison of object recognition behavior in human and monkey**. *J Neurosci* 2015, **35**:12127-12136.

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

34. Rajalingham R, Issa E, Bashivan P, Kar K, Schmidt K, DiCarlo J: **Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks**. *J Neurosci* 2018, **38**:7255-7269.

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

36. Kubilius J, Bracci S, de Beeck HPO: **Deep neural networks as a computational model for human shape sensitivity**. *PLoS Comput Biol* 2016, **12**:e1004896.

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

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

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

40. Gatys LA, Ecker AS, Bethge M: **Texture and art with deep neural networks**. *Curr Opin Neurobiol* 2017, **46**:178-186.

41. •• Kell AJE, Yamins DLK, Shook EN, Norman-Haignere S, McDermott JH: **A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy**. *Neuron* 2018, **98**:630-644.

This paper demonstrates the use of deep networks in a domain outside of the ventral visual stream. It shows that deep networks optimized for speech and music recognition exhibit human-like behavior, predict auditory cortical responses, and provide evidence for hierarchical organization in the human auditory system. It also introduces the use of multi-task networks with different branches as a means to propose hypotheses about functional organization in brain systems.

42. Sussillo D, Churchland MM, Kaufman MT, Shenoy KV: **A neural network that finds a naturalistic solution for the production of muscle activity**. *Nat Neurosci* 2015, **18**:1025-1033.

43. McIntosh L, Maheswaranathan N, Nayebi A, Ganguli S, Baccus S: *Deep Learning Models of the Retinal Response to Natural Scenes*. 2016:1369-1377.

44. Oliver M, Gallant JL: **A deep convolutional energy model of V4 responses to natural movies**. *J Vis* 2016, **16**:876.

45. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, Ecker AS: **Deep convolutional models improve predictions of macaque V1 responses to natural images**. *BioRxiv* 2017:64.

46. Yamins DLK, DiCarlo JJ: **Using goal-driven deep learning models to understand sensory cortex**. *Nat Neurosci* 2016, **19**:356-365.

47. Klindt D, Ecker AS, Euler T, Bethge M: **Neural system identification for large populations separating "what" and "where"**. *Advances in Neural Information Processing Systems*. 2017:3508-3518.

48. Wu MCK, David SV, Gallant JL: **Complete functional characterization of sensory neurons by system identification**. *Annu Rev Neurosci* 2006, **29**:477-505.

49. Naselaris T, Kay KN, Nishimoto S, Gallant JL: **Encoding and decoding in fMRI**. *Neuroimage* 2011, **56**:400-410.

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

This paper was among the first to show that task-optimized deep networks predict multi-unit responses from macaque V4 and IT. Moreover, it shows that aspects of the ventral visual hierarchy are recapitulated by deep networks: intermediate network layers best predict V4, while later layers best predict IT.

51. Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ: **Deep neural networks rival the representation of primate IT cortex for core visual object recognition**. *PLoS Comput Biol* 2014, **10**:e1003963.

52. Güçlü U, van Gerven MAJ: **Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream**. *J Neurosci* 2015, **35**:10005-10014.

53. Eickenberg M, Gramfort A, Varoquaux G, Thirion B: **Seeing it all: Convolutional network layers map the function of the human visual system**. *Neuroimage* 2017, **152**:184-194.

54. Kriegeskorte N, Mur M, Bandettini P: **Representational similarity analysis – connecting the branches of systems neuroscience**. *Front Syst Neurosci* 2008, **2**.

55. • Khaligh-Razavi S-M, Kriegeskorte N: **Deep supervised, but not unsupervised, models may explain IT cortical representation**. *PLoS Comput Biol* 2014, **10**:e1003915.

This paper was among the first to show similarity between the representations of neural networks and the responses in inferotemporal cortex, as measured both with human fMRI and with macaque electrophysiology.

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

57. Zhuang C, Kubilius J, Hartmann MJ, Yamins DL: **Toward goal-driven neural network models for the rodent whisker-trigeminal system**. *Advances in Neural Information Processing Systems (NIPS)* 2017, **vol 30**:2555-2565.

58. Kanitscheider I, Ila F: **Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems**. In *Advances in Neural Information Processing Systems (NIPS 30).* Edited by Luxburg UV, Guyon I, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R. 2017:4529-4538.

59. Cueva CJ, Wei X-X: **Emergence of grid-like representations by training recurrent neural networks to perform spatial localization**. *International Conference on Learning Representations*. 2018.

60. Banino A, Barry C, Uria B, Blundell C, Lillicrap T, Mirowski P, Pritzel A, Chadwick MJ, Degris T, Modayil J *et al.*: **Vector-based navigation using grid-like representations in artificial agents**. *Nature* 2018, **557**:429-433.

61. •• Cheung B, Weiss E, Olshausen BA: **Emergence of foveal image sampling from learning to attend in visual scenes**. *International Conference on Learning Representations*. 2017.

This paper optimizes a relatively small neural network with a trainable 'retinal' front-end to perform a simple visual search task. It finds that the resulting retinal lattice exhibits features of the primate retina, with a densely sampled fovea and more coarsely sampled periphery.

62. Lee R, Saxe A: **Modeling perceptual learning with deep networks**. *Annual Meeting of the Cognitive Science Society*. 2014.

63. Wenliang LK, Seitz AR: **Deep neural networks for modeling visual perceptual learning**. *J Neurosci* 2018, **38**:6028-6044.

64. •• Lindsay GW, Miller KD: **How biological attention mechanisms improve task performance in a large-scale visual system model**. *eLife* 2018, **7**.

This paper uses a task-optimized neural network as a stand-in for the visual system, and asks a series of questions about feature-based attention. The authors observe different effects on performance depending on where in the network simulated feature-based attention is applied. The paper concludes by proposing neural experiments motivated by their modeling results.

65. Treue S, Martinez Trujillo JC: **Feature-based attention influences motion processing gain in macaque visual cortex**. *Nature* 1999, **399**:575-579.

66. Norman-Haignere S, Kanwisher N, McDermott JH: **Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition**. *Neuron* 2015, **88**: 1281-1296.

67. Rauschecker JP, Tian B: **Mechanisms and streams for processing of "what" and "where" in auditory cortex**. *Proc Natl Acad Sci U S A* 2000, **97**:11800-11806.

68. Kanwisher N: **Functional specificity in the human brain: a window into the functional architecture of the mind**. *Proc Natl Acad Sci U S A* 2010, **107**:11163-11170.

69. Mahendran A, Vedaldi A: **Understanding deep image representations by inverting them**. *IEEE Conference on Computer Vision and Pattern Recognition*. 2015:5188-5196.

70. Olah C, Mordvintsev A, Schubert L: *Feature Visualization*. Distill; 2017.

71. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W: **On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation**. *PLoS One* 2015, **10**: e0130140.

72. Nagamine T, Mesgarani N: **Understanding the representation and computation of multilayer perceptrons: a case study in speech recognition**. *International Conference on Machine Learning*. 2017:2564-2573.

73. Cheung B, Livezey JA, Bansal AK, Olshausen BA: **Discovering hidden factors of variation in deep networks**. *International Conference on Learning Representations*. 2015.

74. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A: **beta-VAE: learning basic visual concepts with a constrained variational framework**. *International Conference on Learning Representations* 2016.

75. Hong H, Yamins D, Majaj NJ, DiCarlo JJ: **Explicit information for category-orthogonal object properties increases along the ventral stream**. *Nat Neurosci* 2016, **19**:613-622.

76. Norman-Haignere SV, McDermott JH: **Neural responses to natural and model-matched stimuli reveal distinct computations in primary and non-primary auditory cortex**. *PLoS Biol* 2018, **16**:e2005127.

77. Nayebi A, Bear D, Kubilius J, Kar K, Ganguli S, Sussillo D, DiCarlo JJ, Yamins D: **Task-driven convolutional recurrent models of the visual system**. *Neural Information Processing Systems* 2018, **31**.

78. •• Kar K, Kubilius J, Schmidt KM, Issa EB, DiCarlo JJ: **Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior**. *Nature Neuroscience* 2019. (in press).

This paper studies the role of recurrence in IT cortex, employing images that were poorly recognized by standard deep networks but correctly recognized by humans. Decoding performance from IT for these 'challenge' images peaked later in the response time course. Their results suggest that these kinds of images may require recurrent processing in order to be recognized.

79. Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Caro JO, Hardesty W, Cox D, Kreiman G: **Recurrent computations for visual pattern completion**. *Proc Natl Acad Sci U S A* 2018, **115**:8835-8840.

80. Abbott LF, DePasquale B, Memmesheimer RM: **Building functional networks of spiking model neurons**. *Nat Neurosci* 2016, **19**:350-355.

81. Nicola W, Clopath C: **Supervised learning in spiking neural networks with FORCE training**. *Nat Commun* 2017, **8**:2208.

82. Zenke F, Ganguli S: **SuperSpike: supervised learning in multilayer spiking neural networks**. *Neural Comput* 2018, **30**:1514-1541.

83. Miconi T, Rawal A, Clune J, Stanley KO: **Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity**. *International Conference on Learning Representations*. 2019.

84. • Guerguiev J, Lillicrap TP, Richards BA: **Towards deep learning with segregated dendrites**. *eLife* 2017, **6**:e22901.

This paper explores the potential benefits of incorporating multiple segregated compartments into each model 'neuron' in an artificial network. Such compartments may facilitate an implementation of back-propagation that may be more consistent with known neurobiology.

85. Bartunov S, Santoro A, Richards B, Marris L, Hinton GE, Lillicrap T: **Assessing the scalability of biologically-motivated deep learning algorithms and architectures**. *In Advances in Neural Information Processing Systems* 2018:9390-9400.

86. Schrimpf M, Kubilius J, DiCarlo JJ: **Brain-score: which artificial neural network best emulates the brain's neural network?** *Computational Cognitive Neuroscience* 2018.

87. Henaff OJ, Simoncelli EP: **Geodesics of learned representations**. *International Conference on Learning Representations*. 2016.

88. •• Azulay A, Weiss Y: **Why do deep convolutional networks generalize so poorly to small image transformations?** *arXiv preprint arXiv* 2018, **1805**:12177.

This paper demonstrates that convolutional neural networks are not translation-invariant, contrary to conventional wisdom. The authors suggest that the networks' sensitivity to small transformations is a result of strided convolution and pooling operations that ignore the sampling theorem.

89. Berardino A, Balle J, Laparra V, Simoncelli EP: **Eigen-distortions of hierarchical representations**. *Advances in Neural Information Processing Systems (NIPS 30)* 2017, **vol 30**:1-10.

90. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R: **Intriguing properties of neural networks**. *International Conference on Learning Representations* 2014. p. 1312.6199.

91. Goodfellow IJ, Shlens J, Szegedy C: **Explaining and harnessing adversarial examples**. *International Conference on Learning Representations* 2015.

92. Elsayed GF, Shankar S, Cheung B, Papernot N, Kurakin A, Goodfellow I, Sohl-Dickstein J: **Adversarial examples that fool both computer vision and time-limited humans**. *Neural Information Processing Systems*. 2018.

93. Schott L, Rauber J, Brendel W, Bethge M: **Robust perception through analysis by synthesis**. *arXiv* 2018. 1805.09190.

94. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T: **DeCAF: a deep convolutional activation feature for generic visual recognition**. *The 31st International Conference on Machine Learning* 2014, **vol 32**:647-655.

95. Kornblith S, Shlens J, Le QV: **Do better ImageNet models transfer better?** *arXiv* 2018. 1805.08974.

96. Siegel MH: **Compositional Simulation in Perception and Cognition**. *Brain and Cognitive Sciences, Volume PhD*. Massachusetts Institute of Technology; 2018.

97. Yuille A, Kersten D: **Vision as Bayesian inference: analysis by synthesis?** *Trends Cogn Sci* 2006, **10**:301-308.

98. Cusimano M, Hewitt L, Tenenbaum JB, McDermott JH: **Auditory scene analysis as Bayesian inference in sound source models**. *Computational Cognitive Neuroscience* 2018.

99. • Yildirim I, Freiwald W, Tenenbaum JB: **Efficient inverse graphics in biological face processing**. *bioRxiv* 2018.

This paper offers a modern take on the classic 'analysis-by-synthesis' approach to perception. It trains a neural network to efficiently invert a generative model of faces, and suggests that such a network accounts for human behavior and macaque physiology data.

100. Dayan P, Hinton GE, Neal RM, Zemel RS: **The Helmholtz machine**. *Neural Comput* 1995, **7**:889-904.

101. Hinton GE, Dayan P, Frey BJ, Neal RM: **The "wake-sleep" algorithm for unsupervised neural networks**. *Science* 1995, **268**:1158-1161.

102. Pinto N, Cox DD, DiCarlo JJ: **Why is real-world visual object recognition hard?** *PLoS Comput Biol* 2008, **4**:e27.

103. Zeiler MD, Fergus R: **Visualizing and understanding convolutional networks**. *European Conference on Computer Vision*. Springer International; 2014:818-833.

104. Simonyan K, Zisserman A: **Very deep convolutional networks for large-scale image recognition**. *arXiv* 2014. 1409.1556.

105. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: **Going deeper with convolutions**. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015:1-9.

106. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: **Rethinking the inception architecture for computer vision**. *The IEEE Conference on Computer Vision and Pattern Recognition*. 2016:2818-2826.

107. Lowe D: **Distinctive image features from scale-invariant keypoints**. *Int J Comput Vis* 2004, **60**:91-110.

108. Pinto N, Doukhan D, DiCarlo JJ, Cox DD: **A high-throughput screening approach to discovering good forms of biologically inspired visual representation**. *PLoS Comput Biol* 2009, **5**: e1000579.

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

110. Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA: **A functional and perceptual signature of the second visual area in primates**. *Nat Neurosci* 2013, **16**:974-981.