Deep neural network models of sensory systems: advances and limitations

Alexander J.E. Kell[1,2] and Josh H. McDermott[1,2,3,4]

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

Abstract

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 recent 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.

Highlights

- Deep neural networks (DNNs) now reach human-level performance on some perceptual tasks.
- They show human-like error patterns and predict sensory cortical responses.
- Like ideal observer models, DNNs reveal how tasks shape neural systems and behavior.
- DNNs offer hypotheses of intermediate representations and functional organization.
- DNNs remain incomplete models of neural circuitry, learning, and perceptual inference.

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 therefore 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 (Fig. 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 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] (Fig. 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 (Fig. 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] (Fig. 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 interpretion, 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 neural 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. Prior to 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 [50-53] (Fig. 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] (Fig. 2D).

The ability of DNN features to generate good predictions of neural responses raises questions about the purpose of 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 [50-53] (Fig. 2C and 2E). 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] (Fig. 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.

Based in part on their utility in the visual and auditory systems, deep networks have recently begun to be employed in analogous fashion in other domains, including the somatosensory system [57], as well as the grid and place cell systems of the medial temporal lobe [58-60].

Future directions

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, analogous to saccadic eye movements. 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 (Fig. 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.

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 experimental findings, including whether a training tasks yields changes at earlier or later stages of sensory processing and how a task’s precision alters generalization to new stimuli. 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, the proposed 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 (Fig. 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 therefore 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 forcing 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] (Fig. 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] (Fig. 3D).

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 [86]. 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 not always [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.

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 DOE Computational Science Graduate Fellowship (DE-FG02-97ER25308) to A.J.E.K., a McDonnell Scholar Award to J.H.M., and NSF grant BCS-1634050.

Conflict of Interest
The authors declare no conflict of interest.

References

1. Heeger, D.J. (1992). Normalization of cell responses in cat striate cortex. Visual Neuroscience *9*, 181-197.

2. Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network *12*, 289-316.

3. Pillow, J.W., Paninski, L., Uzzell, V.J., Simoncelli, E.P., and Chichilnisky, E.J. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. Journal of Neuroscience *25*, 11003-11013.

4. Rust, N.C., Mante, V., Simoncelli, E.P., and Movshon, J.A. (2006). How MT cells analyze the motion of visual patterns. Nature Neuroscience *9*, 1421-1431.

5. David, S.V., Mesgarani, N., Fritz, J.B., and Shamma, S.A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. Journal of Neuroscience *29*, 3374-3386.

6. Adelson, E.H., and Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A *2*, 284-299.

7. Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. *102*, 2892-2905.

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

9. Chi, T., Ru, P., and Shamma, S.A. (2005). Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. *118*, 887-906.

10. Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607-609.

11. Schwartz, O., and Simoncelli, E.P. (2001). Natural signal statistics and sensory gain control. Nature Neuroscience *4*, 819-825.

12. Smith, E.C., and Lewicki, M.S. (2006). Efficient auditory coding. Nature *439*, 978-982.

13. Karklin, Y., and Lewicki, M.S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. Nature *457*, 83-86.

14. Carlson, N.L., Ming, V.L., and DeWeese, M.R. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. PLoS Comp. Biol. *8*, e1002594.

15. Mlynarski, W., and McDermott, J.H. (2018). Learning mid-level auditory codes from natural sound statistics. Neural Computation *30*, 631-669.

16. Geisler, W.S. (2011). Contributions of ideal observer theory to vision research. Vision Research *51*, 771-781.

17. Weiss, Y., Simoncelli, E.P., and Adelson, E.H. (2002). Motion illusions as optimal percepts. Nature Neuroscience *5*, 598-604.

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

19. Rumelhart, D., and McClelland, J. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, (MIT Press).

20. Lehky, S.R., and Sejnowski, T.J. (1988). Network model of shape-from-shading: neural function arises from both receptive and projective fields. Nature *333*, 452-454.

21. Zipser, D., and Andersen, R.A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature *331*, 679-684.

22. Nair, V., and Hinton, G.E. (2010). Rectified linear units improve restricted boltzmann machines. In 27th International Conference on Machine Learning.

23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research *15*, 1929-1958.

24. Ioffe, S., and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv, 1502.03167.

25. Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems.

26. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine *29*, 82-97.

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

28. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. (1990). Handwritten digit recognition with a back-propagation network. In Advances in Neural Information Processing (NIPS 1989), Volume 2, D. Touretsky, ed. (Denver, CO: Morgan Kauffman).

29. Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology *160*, 106-154.

30. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In The IEEE Conference on Computer Vision and Pattern Recognition. pp. 770-778.

31. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Deep connected convolutional networks. In 2017 IEEE Conference on Pattern Recognition and Computer Vision (CVPR). pp. 4700-4708.

32. Rajalingham, R., Schmidt, K., and DiCarlo, J.J. (2015). Comparison of object recognition behavior in human and monkey. Journal of Neuroscience *35*, 12127–12136.

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

34. Rajalingham, R., Issa, E., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, 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. Journal of Neuroscience *38*, 7255-7269.

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

36. Kubilius, J., Bracci, S., and de Beeck, H.P.O. (2016). Deep neural networks as a computational model for human shape sensitivity. PLoS Comp. Biol. *12*, e1004896.

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

38. Baker, N., Lu, H., Erlikhman, G., and Kellman, P.J. (2018). Deep convolutional networks do not classify based on global object shape. PLoS Comp. Biol. *14*, e1006613.

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

40. Gatys, L.A., Ecker, A.S., and Bethge, M. (2017). Texture and art with deep neural networks. Curr. Opin. Neurobiol. *46*, 178-186.

** 41. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630-644.

42. Sussillo, D., Churchland, M.M., Kaufman, M.T., and Shenoy, K.V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. Nature Neuroscience *18*, 1025-1033.

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

44. Oliver, M., and Gallant, J.L. (2016). A deep convolutional energy model of V4 responses to natural movies. Journal of Vision *16*, 876.

45. Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., and Ecker, A.S. (2017). Deep convolutional models improve predictions of macaque V1 responses to natural images. bioRxiv.

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

47. Klindt, D., Ecker, A.S., Euler, T., and Bethge, M. (2017). Neural system identification for large populations separating "what" and "where". In Advances in Neural Information Processing Systems. pp. 3508-3518.

48. Wu, M.C.K., David, S.V., and Gallant, J.L. (2006). Complete functional characterization of sensory neurons by system identification. Annual Review of Neuroscience *29*, 477-505.

49. Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage *56*, 400-410.

* 50. Yamins, D., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619-8624.

51. Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., and DiCarlo, J.J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comp. Biol. *10*, e1003963.

52. Güçlü, U., and 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*, 10005-10014.

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

54. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience *2*.

* 55. Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comp. Biol. *10*, e1003915.

56. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific Reports *6*, 27755.

57. Zhuang, C., Kubilius, J., Hartmann, M.J., and Yamins, D.L. (2017). Toward goal-driven neural network models for the rodent whisker-trigeminal system. In Advances in Neural Information Processing Systems (NIPS), Volume 30. pp. 2555-2565.

58. Kanitscheider, I., and Ila, F. (2017). Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems. In Advances in Neural Information Processing Systems (NIPS 30), U.V.L. I.. Guyon, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, ed., pp. 4529-4538.

59. Cueva, C.J., and Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In International Conference on Learning Representations.

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

** 61. Cheung, B., Weiss, E., and Olshausen, B.A. (2017). Emergence of foveal image sampling from learning to attend in visual scenes. In International Conference on Learning Representations.

62. Lee, R., and Saxe, A. (2014). Modeling perceptual learning with deep networks. In Annual Meeting of the Cognitive Science Society.

63. Wenliang, L.K., and Seitz, A.R. (2018). Deep neural networks for modeling visual perceptual learning. Journal of Neuroscience *38*, 6028-6044.

** 64. Lindsay, G.W., and Miller, K.D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. eLife *7*.

65. Treue, S., and Martinez Trujillo, J.C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. Nature *399*, 575–579.

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

67. Rauschecker, J.P., and Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. Proc. Natl. Acad. Sci. USA *97*, 11800-11806.

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

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

70. Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. Distill.

71. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE *10*, e0130140.

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

73. Cheung, B., Livezey, J.A., Bansal, A.K., and Olshausen, B.A. (2015). Discovering hidden factors of variation in deep networks. In International Conference on Learning Representations.

74. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In International Conference on Learning Representations. (Toulon, France).

75. Hong, H., Yamins, D., Majaj, N.J., and DiCarlo, J.J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. Nature Neuroscience *19*, 613-622.

76. Norman-Haignere, S.V., and McDermott, J.H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and non-primary auditory cortex. PLoS Biology *16*, e2005127.

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

** 78. Kar, K., Kubilius, J., Schmidt, K. M., Issa, E. B., & DiCarlo, J. J. (2018). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. bioRxiv, 354753.

79. Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J.O., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. Proc. Natl. Acad. Sci. USA *115*, 8835-8840.

80. Abbott, L.F., DePasquale, B., and Memmesheimer, R.M. (2016). Building functional networks of spiking model neurons. Nature Neuroscience *19*, 350-355.

81. Nicola, W., and Clopath, C. (2017). Supervised learning in spiking neural networks with FORCE training. Nature Communications *8*, 2208.

82. Zenke, F., and Ganguli, S. (2018). SuperSpike: Supervised learning in multilayer spiking neural networks. Neural Computation *30*, 1514-1541.

83. Miconi, T., Rawal, A., Clune, J., and Stanley, K.O. (2019). Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity. In International Conference on Learning Representations.

* 84. Guerguiev, J., Lillicrap, T.P., and Richards, B.A. (2017). Towards deep learning with segregated dendrites. eLIFE *6*, e22901.

85. Bartunov, S., Santoro, A., Richards, B.A., Hinton, G.E., and Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. arXiv, 1807.04587.

86. Schrimpf, M., Kubilius, J., and DiCarlo, J.J. (2018). Brain-score: Which artificial neural network best emulates the brain's neural network? In Computational Cognitive Neuroscience. (Philadelphia, PA).

87. Henaff, O.J., and Simoncelli, E.P. (2016). Geodesics of learned representations. In International Conference on Learning Representations.

** 88. Azulay, A., and Weiss, Y. (2018). Why do deep convolutional networks generalize so poorly to small image transformations. arXiv.

89. Berardino, A., Balle, J., Laparra, V., and Simoncelli, E.P. (2017). Eigen-distortions of hierarchical representations. In Advances in Neural Information Processing Systems (NIPS 30), Volume 30. pp. 1-10.

90. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In International Conference on Learning Representations. (Banff, Canada), p. 1312.6199.

91. Goodfellow, I.J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations. (San Diego, CA).

92. Elsayed, G.F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. (2018). Adversarial examples that

fool both computer vision and time-limited humans. In Neural Information Processing Systems.

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

94. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). DeCAF: A deep convolutional activation feature for generic visual recognition. In The 31st International Conference on Machine Learning, Volume 32. pp. 647-655.

95. Kornblith, S., Shlens, J., and Le, Q.V. (2018). Do better ImageNet models transfer better? arXiv, 1805.08974.

96. Siegel, M.H. (2018). Compositional Simulation in Perception and Cognition. In Brain and Cognitive Sciences, Volume PhD. (Massachusetts Institute of Technology).

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

98. Cusimano, M., Hewitt, L., Tenenbaum, J.B., and McDermott, J.H. (2018). Auditory scene analysis as Bayesian inference in sound source models. In Computational Cognitive Neuroscience. (Philadelphia, PA).

* 99. Yildirim, I., Freiwald, W., and Tenenbaum, J.B. (2018). Efficient inverse graphics in biological face processing. bioRxiv.

100. Dayan, P., Hinton, G.E., Neal, R.M., and Zemel, R.S. (1995). The Helmholtz machine. Neural Computation *7*, 889-904.

101. Hinton, G.E., Dayan, P., Frey, B.J., and Neal, R.M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. Science *268*, 1158-1161.

102. Pinto, N., Cox, D.D., and DiCarlo, J.J. (2008). Why is real-world visual object recognition hard? PLoS Comp. Biol. *4*, e27.

103. Zeiler, M.D., and Fergus, R. (2014). Visualizing and understanding convolutional networks. In European Conference on Computer Vision. (Springer International), pp. 818-833.

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

105. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with

convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1-9.

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

107. Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision *60*, 91-110.

108. Pinto, N., Doukhan, D., DiCarlo, J.J., and Cox, D.D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comp. Biol. *5*, e1000579.

109. Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. USA *104*, 6424-6429.

110. Freeman, J., Ziemba, C.M., Heeger, D.J., Simoncelli, E.P., and Movshon, J.A. (2013). A functional and perceptual signature of the second visual area in primates. Nature Neuroscience *16*, 974-981.

Recommended:

Guerguiev, J., Lillicrap, T. P., Richards, B. A. (2017) Towards deep learning with segregated dendrites. eLife; 6:e22901. DOI: https://doi.org/10.7554/eLife.22901

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

Yildirim, I., Freiwald, W., Tenenbaum, J.B. (2018) Efficient inverse graphics in biological face processing. bioRxiv. http://dx.doi.org/10.1101/282798.

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.

Khaligh-Razavi, S.-M. & Kriegeskorte, N. (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Computational Biology, 10, e1003915, doi:10.1371/journal.pcbi.1003915.

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.

Yamins, D., 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. Proc. Natl. Acad. Sci. USA 111, 8619-8624, doi:10.1073/pnas.1403112111.

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.

Highly recommended:

Azulay, A. & Weiss, Y. (2018) Why do deep convolutional networks generalize so poorly to small image transformations. arXiv.

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.

Cheung, B., Weiss, E. & Olshausen, B. A. (2017) Emergence of foveal image sampling from learning to attend in visual scenes. In International Conference on Learning Representations.

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.

Kar, K., Kubilius, J., Schmidt, K. M., Issa, E. B., & DiCarlo, J. J. (2018). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. bioRxiv, 354753.

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.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. & McDermott, J. H. (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 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.

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

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.

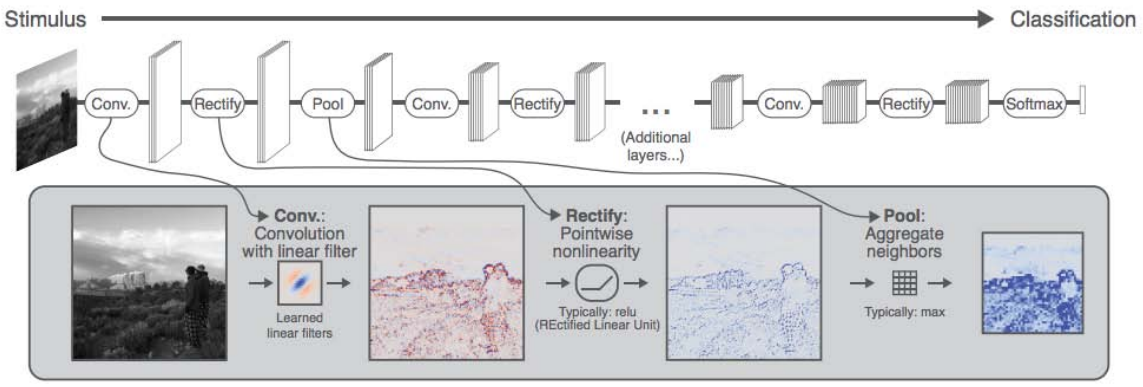

Figure 1. Schematic of a typical deep convolutional neural network.
The stimulus (e.g., an image for a visual task or a spectrogram for auditory task) is passed through a cascade of simple operations, in which the output of one stage of operations is the input to the next. This cascade culminates in a discriminative classification (e.g., of the object category present in the image, or the spoken word present in the sound signal). Due to 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.

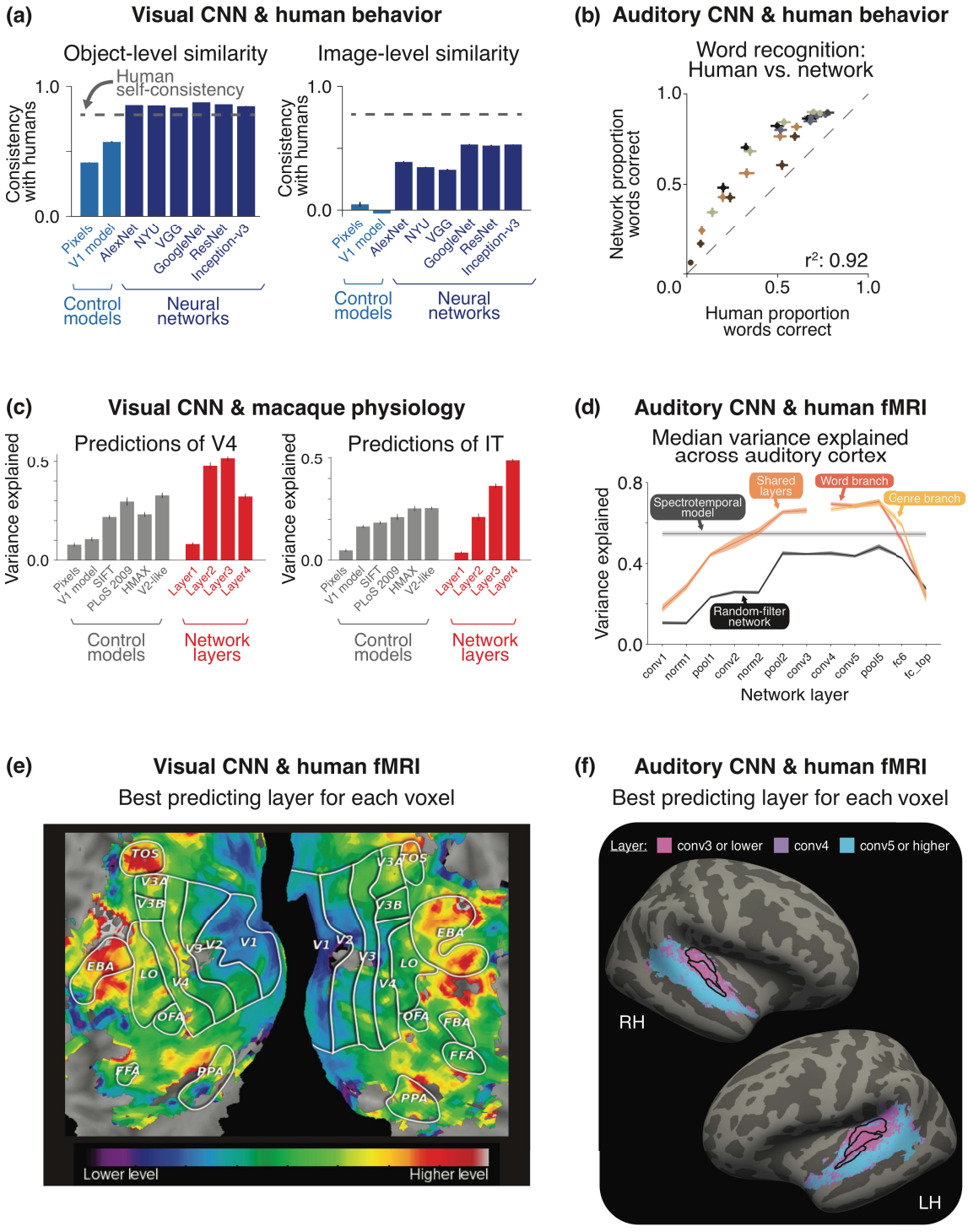

Figure 2. Task-optimized deep neural networks predict visual and auditory cortical responses and recapitulate real-world behavior.

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 [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 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., 2018.

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., 2018.

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 [50]). Intermediate network layers best predict intermediate visual area V4, while later layers best predict later visual area IT. From Yamins et al., 2014.

d. Response prediction accuracy of an audio-trained DNN used to predict responses to natural sound. 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., 2018.

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., 2017.

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., 2018.

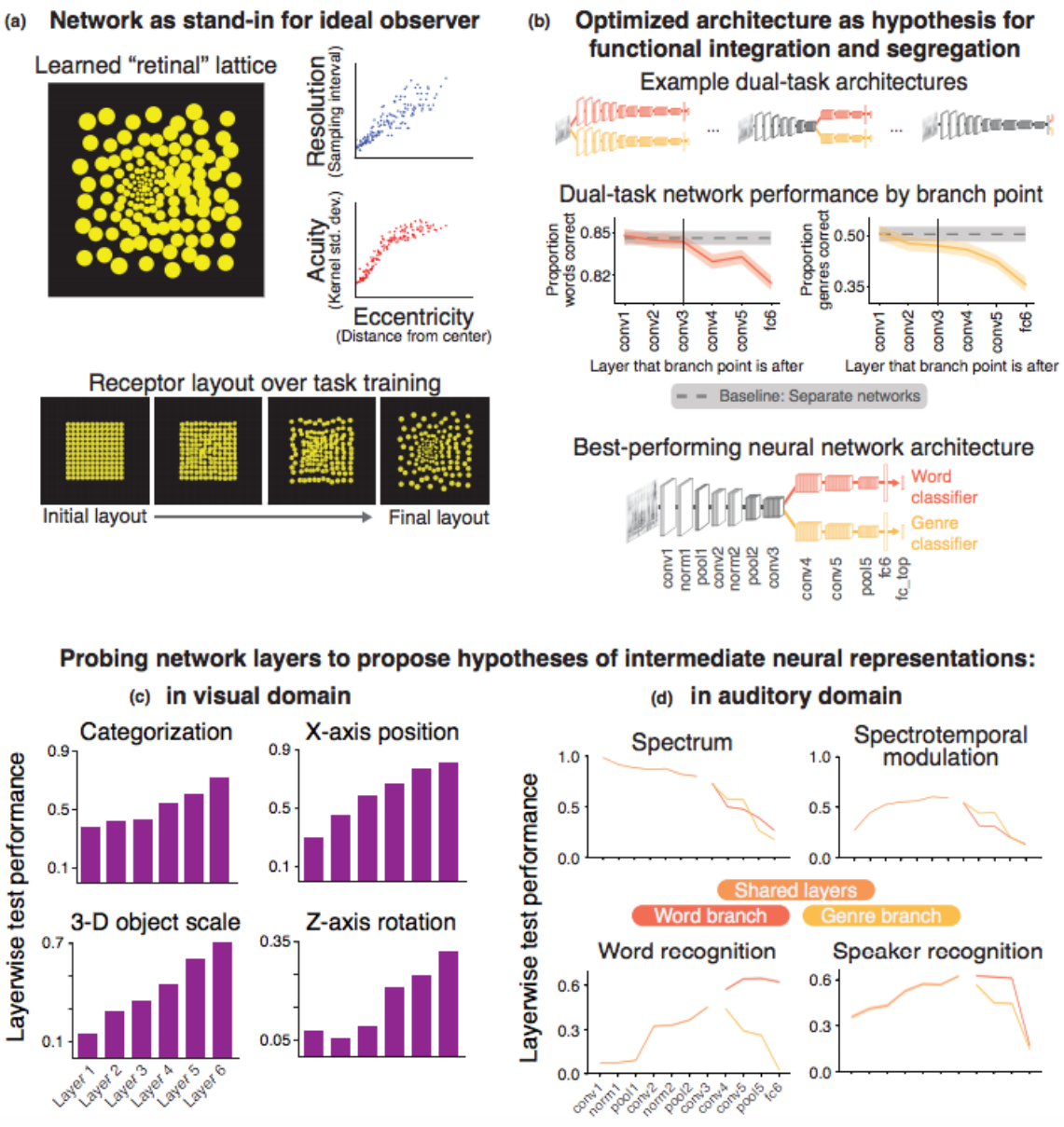

Figure 3. Neural networks as hypothesis generators for neuroscience.

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., 2016.

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., 2018.

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 categories. From Hong et al., 2016.

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., 2018.