Check for updates

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
2019, Vol. 28(1) 34–39
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0963721418801342
www.psychologicalscience.org/CDPS
SAGE

Comparing the Visual Representations and Performance of Humans and Deep Neural Networks

Robert A. Jacobs and Christopher J. Bates
Department of Brain and Cognitive Sciences, University of Rochester

Abstract
Although deep neural networks (DNNs) are state-of-the-art artificial intelligence systems, it is unclear what insights, if any, they provide about human intelligence. We address this issue in the domain of visual perception. After briefly describing DNNs, we provide an overview of recent results comparing human visual representations and performance with those of DNNs. In many cases, DNNs acquire visual representations and processing strategies that are very different from those used by people. We conjecture that there are at least two factors preventing them from serving as better psychological models. First, DNNs are currently trained with impoverished data, such as data lacking important visual cues to three-dimensional structure, data lacking multisensory statistical regularities, and data in which stimuli are unconnected to an observer's actions and goals. Second, DNNs typically lack adaptations to capacity limits, such as attentional mechanisms, visual working memory, and compressed mental representations biased toward preserving task-relevant abstractions.

Keywords
perception, vision, artificial intelligence, deep neural networks

Deep neural networks (DNNs) are state-of-the-art artificial intelligence (AI) systems providing impressive performance in a wide range of domains, such as visual perception, speech recognition, text-to-text language translation, and product recommendation. For example, in the domain of computer vision, a subclass of DNNs known as convolutional DNNs has consistently won the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015) in recent years. Although these networks have advanced the field of AI, it is an open question as to what insights, if any, these DNNs provide about human intelligence. Do they provide new insights into the nature of human thought and cognition? In this article, we address this question as it pertains to the domain of visual perception. Our main conclusion is that DNNs (more precisely, convolutional DNNs) provide a good starting point for the development of comprehensive accounts of human visual perception. To date, however, at least two factors—an impoverished set of training experiences and a lack of adaptations to capacity limits—prevent them from serving as better psychological models.

Overview of DNNs

Neural networks consist of interconnected sets of units (LeCun, Bengio, & Hinton, 2015). Some units are designated as input units, other units are output units, and still other units are "hidden" units. The goal of a network is to map patterns of input-unit "activations" to target or desired patterns of output-unit activations. For instance, a network might map patterns representing visual images (e.g., images of vehicles) to patterns representing category labels (e.g., a vehicle might be a car, truck, or bus). If an image contains M pixels, then the

Corresponding Author:
Robert A. Jacobs, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627
E-mail: robbie@bcs.rochester.edu

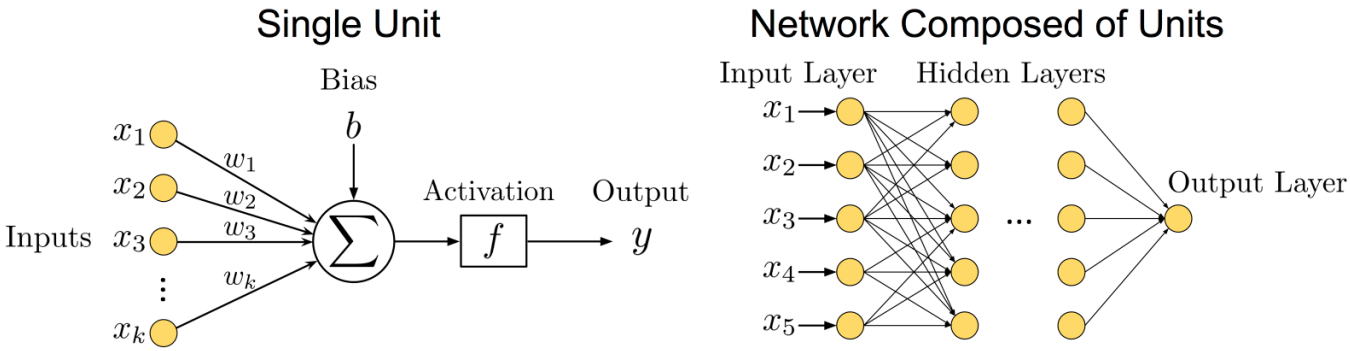

Fig. 1. A single unit of a neural network and a network composed of several units. An individual hidden or output unit of a network (left) computes its activation in two stages. First, it computes the weighted sum of the activations of the units that connect to it (these activations are denoted $x_1, \ldots x_k$; the weights are denoted $w_1, \ldots w_k$; and the symbol Σ denotes summation). Second, it uses a nonlinear function f to map the weighted sum to an activation value y. Input, hidden, and output units are organized into input, hidden, and output layers, respectively, which form a network (right). Researchers need to make many choices when designing networks. How many layers should a network have? How many units should be in each layer? What should be the pattern of connectivity between units in one layer and units in subsequent layers? How should units map their weighted sums to their activations? What learning rule should be used to modify a network's weights? To date, there are few mathematically principled ways of addressing these questions, and thus researchers rely primarily on intuitions gained through experience.

network will contain M input units. When a particular image is presented to the network, the activation of each input unit is set to its corresponding pixel value. If there are N possible category labels, then the network will have N output units. The activation of an output unit might be an estimate of the probability that an image maps to the unit's corresponding category.

Typically, input units are not directly connected to output units. Instead, input units connect to one or more layers of hidden units that, in turn, connect to output units. The activation of each hidden or output unit is computed in two steps. First, a unit computes the "weighted sum" of its inputs—it multiplies the activation of each unit that connects to it by an input-specific weight value and then sums these products. Next, the unit maps its weighted sum to an activation value. This mapping is nonlinear. Additional details regarding DNNs are provided in Figure 1, and the subclass of convolutional DNNs is described in Figure 2.

The power of DNNs is that they are capable of learning and generalization. Networks learn by adapting the values of their units' weights. Learning is typically supervised, meaning that a "teacher" has specified the target output activation pattern for each input activation pattern. During training, a network's weights are modified to minimize its error or difference between the target output pattern and its actual output pattern. Learning rules for adapting networks' weights often resemble Hebbian rules governing learning in biological neural networks (Marblestone, Wayne, & Kording, 2016). Following training, it is hoped that a network is capable of generalization, meaning that in addition to producing the target output pattern for each input pattern in the training set, it can also produce approximately correct output patterns for novel input patterns that are similar to the training set's input patterns.

Comparison of Human and DNN Visual Perception

In this section, we provide a brief overview of some recent comparisons between people and DNNs, focusing on their visual-processing strategies and performances. Although there are important articles highlighting similarities in visual processing between people and DNNs (e.g., Battleday, Peterson, & Griffiths, 2017; Kubilius, Bracci, & Op de Beeck, 2016; Peterson, Abbott, & Griffiths, 2016), here we emphasize articles pointing out differences because we believe that these differences are more enlightening. For brevity, we do not include neuroscientific comparisons between biological nervous systems and DNNs (e.g., Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).

Several articles have demonstrated differences in people's and DNNs' visual representations through the use of *adversarial examples*. These are images that people and DNNs classify correctly but, when perturbed in special ways that are imperceptible to people, are misclassified by DNNs. For example, Szegedy et al. (2014) found that a small perturbation of an image of a school bus caused a DNN to misclassify the image as depicting an ostrich. Adversarial examples illustrate that DNNs have nonhumanlike discontinuities in the space of their visual representations. (Interested readers should also see Elsayed et al., 2018.)

Other patterns of errors also indicate that people and DNNs use very different visual strategies and

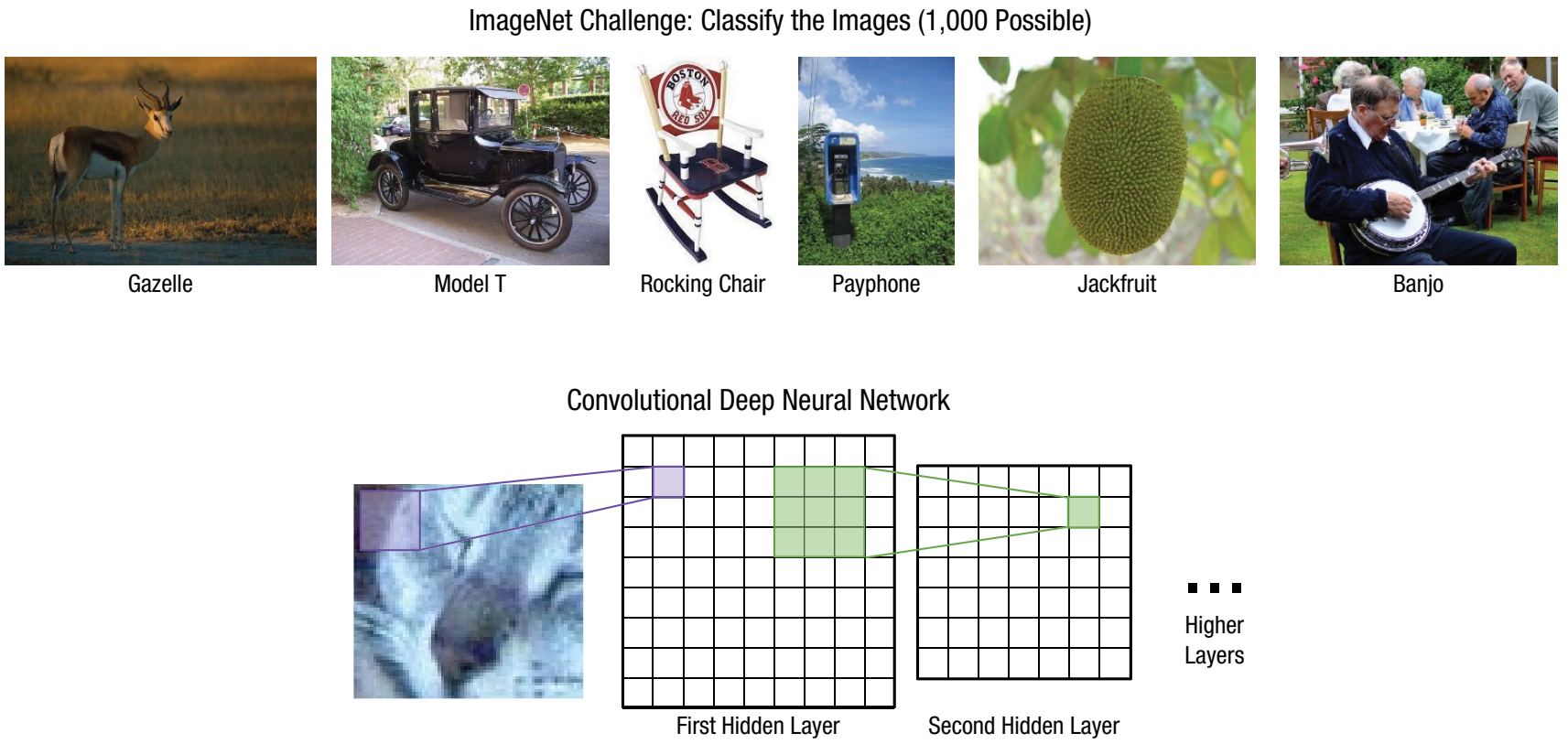

Fig. 2. The ImageNet Large Scale Visual Recognition Challenge and convolutional deep neural network (DNN). The ImageNet Large Scale Visual Recognition Challenge is a prominent competition in the computer-vision community that has been won by convolutional DNNs in recent years. The images at the top represent six sample data items that have been used in this competition. Each data item is a static image that has been assigned to one of a thousand possible categories. As can be seen in the bottom image, convolutional DNNs have early layers of hidden units with local connectivity, meaning that each unit in a layer receives inputs from a small number of units in the previous layer. For example, a unit in the first hidden layer may receive inputs from a small local patch of the input image. Subsequent layers use hidden units with larger or more global connectivity.

representations. Rajalingham et al. (2018) compared people's and DNNs' performance in visual-object categorization and found that they are similar when considered at a coarse-scale category level but markedly different when considered at a finer-scale image level. Lake, Salakhutdinov, and Tenenbaum (2015) reported that people were much better than DNNs at classifying images of letters on the basis of very few training exemplars. Ricci, Kim, and Serre (2018) found that people are very good at evaluating visual relations (e.g., Are the two objects depicted in an image the same or different?), whereas DNNs struggle at learning to make such evaluations. Erdogan and Jacobs (2017) reported that people and DNNs make different shape-similarity judgments on an image set depicting novel three-dimensional objects lacking semantics.

People and DNNs seem to show different visual performances under unusual or impoverished viewing conditions. Dodge and Karam (2017) found differences in people's and DNNs' responses to visual-object categories when images were distorted by added noise or blur. Geirhos et al. (2017) reported that people respond more robustly than DNNs to several types of image distortions. Hosseini, Xiao, Jaiswal, and Poovendran (2017) showed that people, but not DNNs, were good at classifying negative images that have the same structure and semantics as regular images but with reversed brightness (i.e., in a negative image, bright pixels in an image appear dark, and dark pixels appear bright). Ullman, Assif, Fetaya, and Haran (2016) showed that people and DNNs have different "minimal recognizable configurations" (p. 2744), which are the smallest image patches that still permit an object to be recognized.

Taken as a whole, there are several factors accounting for the differences in visual performance between people and DNNs. People tend to be highly sensitive to three-dimensional shape features, whereas DNNs are more sensitive to two-dimensional image features. In addition, people's responses to image distortions or reduced viewing conditions are more robust because they are better at using global information, such as image context or top-down knowledge.

Factors Limiting the Use of DNNs as Psychological Models

Consistencies between the visual performances of people and DNNs suggest that DNNs are a computational framework providing a good starting point for the development of comprehensive accounts of human visual perception. Discrepancies, however, suggest that there is much more work that needs to be done. To date, at least two factors prevent DNNs from serving as better psychological models.

First, relative to people, DNNs receive impoverished training experiences. Whereas natural environments provide people with perceptually rich and dynamic experiences from which they learn to perceive the world, AI researchers typically train DNNs in a supervised manner using data sets of labeled static images. From a psychological perspective, there are at least two shortcomings with training in this manner. First, whereas DNNs receive explicit supervision from a teacher, people in natural environments typically learn in a manner that involves no or little explicit supervision. Second, whereas DNNs are trained with static images, people learn in perceptually rich, dynamic, and interactive environments. The end result is that people and DNNs often learn different information. People tend to learn low-dimensional statistical regularities of visual environments that give rise to visual stimuli and regularities that are informative for decision making and action selection. In contrast, DNNs tend to learn image features that distinguish images of one category from images of other categories. In the future, DNNs will be better psychological models if they are trained in a more humanlike manner with more realistic data items.

More humanlike training can take place in at least three different ways. First, researchers can create data sets with a variety of visual cues, including cues to three-dimensional structure, such as motion parallax and binocular disparities. Scientists working with video already have data that include motion cues. A benefit of working with video is that it provides an opportunity to combine supervised and unsupervised learning—there is no need for a teacher to label every video frame because a DNN could learn in an unsupervised manner by taking advantage of the temporal coherency of the three-dimensional structure across frames.

Second, researchers can create multisensory data sets, which could include both visual and auditory information. Again, scientists working with video already have access to this information. And again, this would provide new opportunities for unsupervised learning—a DNN could learn statistical regularities that occur across modalities, using information from one modality to disambiguate information in another modality.

The presence of visual-auditory data sets raises an interesting question: Can an agent (biological or artificial) learn to perceive the world by watching television (which provides both visual and auditory information)? The fact that the answer is probably "no" motivates a third approach, one in which researchers develop data sets in which stimuli and actions interact in a continuous loop—the current stimulus influences an agent's actions, which, in turn, influence the next stimulus, and so on. Researchers developing software for virtual-reality environments already have the means to generate data sets with perception-action loops. Although

virtual-reality applications would provide simulated (as opposed to real) data, they would serve as useful starting points. Moreover, because objects in virtual reality are simulated, the objects could be fully labeled, thereby providing data for supervised training. Recent research using video games has explored perception and action learning when an agent interacts with a complex environment (Mnih et al., 2015). Interestingly, the DNNs used in this research were trained via reinforcement learning, not supervised learning, a training paradigm that is often regarded as more psychologically realistic.

A second factor limiting the usefulness of DNNs as psychological models is that DNNs rarely contain adaptations to capacity limits. People's visual systems have limited processing powers. Consequently, we have evolved or developed mechanisms or strategies to compensate for our own limitations. For example, because we cannot visually perceive and represent all aspects of a scene at all levels of detail, we often represent a summary, or gist, of the scene (Oliva, 2005). This representation contains abstractions needed to grasp the scene's meaning, to recognize a few objects and other salient properties, and to facilitate object detection and the deployment of attention. A second adaptation to our inability to simultaneously perceive the entirety of a scene is visual attention. Instead of attempting to perceive and represent all aspects of a scene at the same time, we often use a sequential strategy in which we perceive different task-relevant subsets of a scene's properties at different moments in time. Consistent with this strategy, visual working memory is used to preserve and integrate task-relevant information obtained at earlier moments to aid perception and action at the current moment (Ballard, Hayhoe, Pook, & Rao, 1997; Sims, Jacobs, & Knill, 2012).

To better account for human perception, DNNs will need to include capacity limits as well as mechanisms to compensate for these limits. Within the AI community, this is beginning to happen. For example, researchers are developing DNNs that acquire low-dimensional or compressed representations (Kingma & Welling, 2014), DNNs that use visual attention to perform tasks in a sequential manner (Eslami et al., 2016), and DNNs with recurrent connections or external memory to integrate previously acquired information with new information (Graves et al., 2016; Hochreiter & Schmidhuber, 1997). These are early and promising steps toward more psychologically realistic DNNs.

Conclusion

DNNs have achieved impressive performance on important visual tasks. However, such performance has often been obtained using simplified training procedures and data sets that were laboriously labeled by people. We believe that future progress will require AI researchers to use training procedures in which DNNs take actions in perceptually rich environments to achieve goals. We also speculate that future progress will require AI researchers to make DNNs more humanlike, including adding capacity limits and adaptations to these limits (e.g., visual attention, memory, and abstraction). For AI researchers, this research strategy will provide sophisticated solutions to difficult perceptual and decision-making problems. For psychologists, this strategy will provide advanced computational frameworks for implementing and evaluating psychological theories in large-scale realistic settings.

Recommended Reading

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). (See References). A thorough investigation of changes to people's and deep neural networks' (DNNs') visual performances when images are distorted.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. A broad commentary and analysis of how deep neural networks can help us understand neurophysiological, brain imaging (functional MRI), and behavioral data regarding visual perception.

Ullman, S., Assif, L., Fetaya, E., & Haran, D. (2016). (See References). An innovative approach to uncovering the smallest patches or "atoms" that can be used by people and DNNs during visual object recognition.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365. An insightful and highly accessible overview of how DNNs are being used to study visual cortex.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object detectors emerge in deep scene CNNs. Retrieved from Cornell University Library arXiv.org website: https://arxiv.org/abs/1412.6856. Demonstrates that—although the inner workings of DNNs tend to be opaque—networks trained to categorize scenes develop units that are object detectors for the objects that compose those scenes.

Action Editor

Randall W. Engle served as action editor for this article.

Acknowledgments

We thank Ilker Yildirim for helpful discussions.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by National Science Foundation (NSF) Grants BCS-1400784 and DRL-1561335. C. J. Bates was also supported by NSF Research Traineeship Grant NRT-1449828 and NSF Graduate Research Fellowship Grant DGE-1419118.

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral & Brain Sciences*, *20*, 723–767.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2017). Modeling human categorization of natural images using deep feature representations. Retrieved from Cornell University Library arXiv.org website: https://arxiv.org/abs/1711.04855

Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. Retrieved from Cornell University Library arXiv.org website: https://arxiv.org/abs/1705.02498

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. Retrieved from Cornell University Library arXiv .org website: https://arxiv.org/abs/1802.08195

Erdogan, G., & Jacobs, R. A. (2017). Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychological Review*, *124*, 740–761.

Eslami, S. M. A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Kavukcuoglu, K., & Hinton, G. E. (2016). Attend, infer, repeat: Fast scene understanding with generative models. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29*. Retrieved from https://papers.nips .cc/paper/6230-attend-infer-repeat-fast-scene-understanding-with-generative-models.pdf

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. Retrieved from Cornell University Library arXiv.org website: https://arxiv.org/abs/1706.06969

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwinska, A., . . . Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*, 471–476.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hosseini, H., Xiao, B., Jaiswal, M., & Poovendran, R. (2017). On the limitation of convolutional neural networks in recognizing negative images. Retrieved from Cornell University Library arXiv.org website: https://arxiv.org/abs/1703.06857

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, *10*(11), Article e1003915. doi:10.1371/journal.pcbi.1003915

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. Retrieved from Cornell University Library arXiv.org website: https://arxiv.org/abs/1312.6114

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology*, *12*(4), Article e1004896. 10.1371/journal.pcbi.1004896

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*, 1332–1338.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience, 10*, Article 94. doi:10.3389/fncom.2016.00094

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier Academic Press.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. Retrieved from Cornell University Library arXiv .org website: https://arxiv.org/abs/1608.02164

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *The Journal of Neuroscience*, *38*, 7255–7269.

Ricci, M., Kim, J., & Serre, T. (2018). Same-different problems strain convolutional neural networks. Retrieved from Cornell University Library arXiv.org website: https://arxiv .org/abs/1802.03390

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*, 211–252.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*, 807–830.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. Retrieved from Cornell University Library arXiv.org website: https://arxiv.org/abs/1312.6199

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

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