13 Computational Models of Human Object and Scene Recognition

AUDE OLIVA

ABSTRACT To solve object and scene recognition tasks, the human brain has developed a particular cortical topology within its ventral and dorsal streams, recruiting regions in cascades to build a representation of what we see. A class of models, termed *artificial deep neural networks*, has been shown to learn a hierarchical representation of images, akin to the primate visual system, revealing internal representations that resemble the hierarchical topography in both the ventral and the dorsal visual streams of the human brain. Deep neural network models provide a hypothesis-testing framework to predict human brain responses as well as to give insights into how a network, natural or artificial, can learn and represent the visual world.

The High-Level Brain Regions Involved in Object and Scene Recognition

Brains are optimized to compute meaningful patterns from sensory inputs and to solve tasks fitted to their environment: while echolocation in dolphins is powerful under water, vision is the most dominant sense in primates. Yet visual object and scene recognition are difficult computational problems to solve given an almost infinite space of variation within our environment: Objects appear in different places, with different orientations, shapes, colors, and textures, and many can be made of different materials. In the real world, objects can be embedded in clutter, are often occluded, and can be observed from different distances and viewpoints. Despite these challenges, when we look at the world, a "feat of neural engineering" delivers a representation of what we see within only a few hundred milliseconds. How does the human brain instantiate visual recognition?

Entering the retina, visual information initially reaches the primary visual cortex (area V1, in the calcarine sulcus) before being transmitted to a series of retinotopically organized regions in visual occipital areas (areas V2, V3, human V4) and then higher-level regions of the cortex (for a review, see Grill-Spector and Weiner, 2014), which are responsible for visual perception and recognition. The types of computations and image features represented in different cortical sites have been extensively studied in primates in the past decades. For instance, low-level image properties, such as contrast, edges, and orientations, are processed within V1, V2, and V3; object form, shape, and physical size elicit stronger responses in the lateral occipital cortex, or LOC (Grill-Spector et al., 1999; Konkle & Oliva, 2012; Malach et al., 1995); images representing a layout, such as a scene or space, elicit stronger responses in the parahippocampal gyrus (Epstein & Kanwisher, 1998).

The temporal dynamics reflecting how these regions respond to an object or place suggest that the neural representation quickly passes through serial computations (Grill-Spector & Weiner, 2014). Neural responses first emerge in the occipital pole (V1, V2, V3) within 80–100 ms of viewing. Responses then spread rapidly and progressively in the anterior direction along the ventral stream (i.e., the LOC, ventral occipital cortex, temporal occipital cortex, and parahippocampal cortex) and the dorsal stream (intraparietal sulcus regions) within 110–170 ms after image onset (Cichy, Pantazis, & Oliva, 2016a).

These processing stages, particularly those in the ventral stream pathway, have inspired a class of computational models named deep (convolutional) neural networks (deep CNNs or DNNs). These models are quickly gaining popularity as a tool for the hypothesis testing of brain computations by providing simplified *artificial network streams*. What are DNNs and how can they be used to evaluate the computations performed within a biological brain?

Artificial Neural Networks

What is an artificial DNN? Artificial neural networks are a class of models that learn to recognize patterns from input data. One of their main properties is that they learn to progressively improve performance on a specific task without being explicitly programmed (for a review, see LeCun, Bengio, & Hinton, 2015). A given network trained in natural images is taught to achieve high performance on a specific task (e.g., a *detection* task, like finding the location of an object, or a *recognition* task, like identifying a place). Note that training these networks for a particular task, such as object classification (Krizhevsky, Sutskever, & Hinton, 2012) or scene classification (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014), generally requires thousands of examples per class, due to a very large space of parameters that define the inner workings of the network.

While the basic operators of these networks are performed by interconnected units (i.e., artificial neurons), the entire network can be described as a high-dimensional mathematical function with many parameters (i.e., weights connecting units), which are tuned during a training (i.e., a learning) phase. In a supervised mode of training, an artificial network is taught associations between an input (i.e., an image) and an output (i.e., a label describing that image). It can adjust its parameters by iteratively reducing the errors between an output-input pair (i.e., the word *orange* and a specific picture of an orange fed, for instance, to the first layer as pixel values over a small zone of the image) across many presentations of the pair. The error is calculated based on the difference between the network's output value and the desired target value. It is then sent backward through the network to iteratively calculate error values for each layer, which are then used to update the network parameters.

This *back-propagation algorithm* (LeCun & Bengio, 1995) is able to find a solution to high-dimensional discrimination tasks—that is, learning the most useful features that recur across all the examples of pictures labeled *orange* in order to distinguish an orange from 1,000 other object categories. While it remains contentious whether a biological network like the brain uses a back-propagation function to learn feature matching between sensory inputs and classes of concepts, supervision signals can be triggered from different external and internal sources of reinforcement (e.g., context provided by different sensory modalities, direct verbal or tactile supervision, or comparison with mnemonic traces; for a review, see Kriegeskorte & Douglas, 2018).

How do artificial DNN models work? Let us look at a typical example of an artificial neural network, a CNN. A CNN is made of multiple layers, each implementing signal- and image-processing functions. Figure 13.1 illustrates the original architecture of a CNN for visual object recognition, nicknamed AlexNet and proposed by Krizhevsky, Sutskever, & Hinton (2012). AlexNet is composed of eight layers, some implementing one function only (e.g., layers three and four) and some implementing different functions in succession (e.g., layers one and two). The key function of a CNN is its convolution, which is implemented in most of the layers through the whole network (five convolutional layers in AlexNet): each layer is made by feature detection received through input from only a small part of the image. The convolution operation performs a summarization of each region of the image. For example, imagine a flashlight covering a region of five by five pixels directed over the top-left corner of an image. The flashlight can slide across the entire image, covering a new region with each movement. In this example, the flashlight is known as a filter (i.e., a neuron or kernel), which is an array of numbers (i.e., weights), and the region it is projecting light over is called a *receptive field*. The sliding motion of the flashlight filter over the whole image is called a *convolution*: as the filter is convolving, it is multiplying the values in the filter with the original pixel values of the image and summing these values. Therefore, for example, for each five-by-five-pixel image region, a single number is produced. With the flashlight filter moving through all regions in the image, a new feature map, of a smaller size than the original image, is created. Therefore, as more convolutional layers are added, the region of the input image analyzed by a given neuron (i.e, its receptive field) increases, giving to the CNN an important property for visual recognition: tolerance to the spatial position of a feature, or pattern, in the image (see the HMAX original model; Riesenhuber and Poggio, 1999; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007).

What are the advantages of DNNs? CNNs have been shown to yield several desirable outcomes observed in biological networks (for a review, see Bengio, 2009; Kriegeskorte, 2015). Akin to the observation that biological neurons exhibit receptive fields that increase in size between early visual areas and ventral stream regions, the "neurons" in later layers of an artificial neural network also have larger receptive fields. Larger receptive fields provide tolerance to the spatial translation of feature or shape in the image. This means that later layers will be able to discriminate a pattern (i.e., a circle or a corner shape) independently of the location of that shape in the original image input. As an added benefit, the shared parameters of the receptive fields that feed into each feature map allow for the same kernel function to be performed in different areas of the input to each layer. This results in increased robustness of a CNN to shifts and distortions in the data. This robustness is then compounded by the greater expressive power of deep architectures (Bengio & LeCun, 2007) where the features of each layer are combined to form a higher level of abstraction in the succeeding layer. This increasing abstraction from layer to layer allows deep CNNs to produce strong generalizations for highly variable functions. Additionally, this architecture supports the hypothesis that the network can disentangle at each step the factors of variation underlying the input–output data. As a result, the neurons in the network learn to

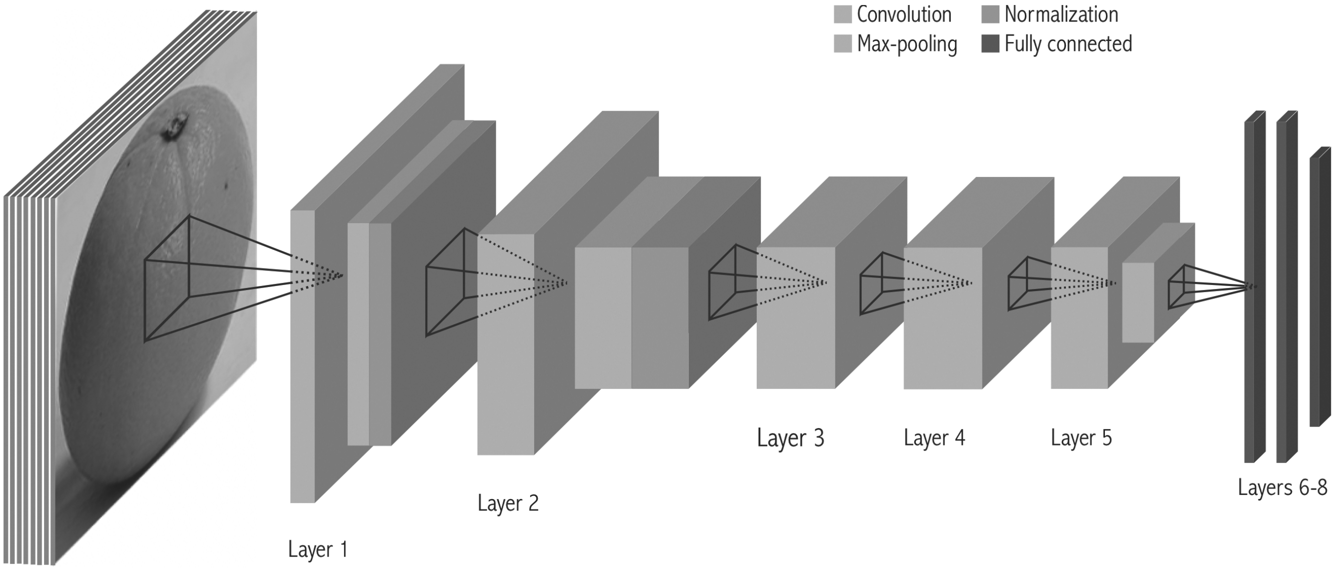

FIGURE 13.1 Deep CNN architecture AlexNet comprises eight layers. Each of layers one through five contains a convolution, some layers with max-pooling and normalization stages. The last three layers are fully connected, with the last layer acting as the output label. The network takes pixel values as inputs and propagates information in a feedforward manner through the layers, activating artificial units with particular weight values successively at each layer. (See color plate 15.)

represent increasingly complex visual patterns as layers deepen (see figure 13.2*B*). For example, in layer one of AlexNet, neurons are tuned to features such as contours and edges, while in layer two, neurons are tuned to curvature and repetitive patterns. In deeper layers, such as layers four and five, neurons become tuned to parts of objects and images, such as shape and form (Zeiler & Fergus, 2014; Zhou, Khosla, Oliva, & Torralba, 2015). This architecture of layers affords the artificial network the ability to learn many variations of a given object class, allowing the network to start resolving one of the most difficult challenges of object recognition: its diversity of views. In the real world, objects of the same category can be seen from different viewpoints, in different backgrounds, with different levels of clutter and occlusion, and at different retinal image sizes. While these models do not yet reach levels of human object recognition, CNNs have solved some of these issues better than models that are not hierarchically organized.

What do DNNs Learn? Different methods for visualizing the artificial units' receptive fields and their selectivity to specific patterns have been proposed as a means toward understanding what a CNN has learned after training for a particular task (Bau, Zhou, Khosla, Oliva, & Torralba, 2017; Simonyan, Vedaldi, & Zisserman, 2014; Zeiler and Fergus, 2014).

In Zhou et al. (2015), the authors describe a method to estimate and quantify the selectivity of every single unit within the five convolutional layers of AlexNet as a function of the task (recognizing objects or scenes) the network was trained to perform. Specifically, a first AlexNet CNN (Object-Net) was trained to discriminate between classes of objects (is it a dog, a cat, or another animal, using the ImageNet data set; Deng et al., 2009). A second AlexNet (Scene-Net) was trained to discriminate between classes of scenes (is it a kitchen, a highway, or another object, using the Places data set; Zhou et al., 2014, 2017). To check what each artificial unit learned, Zhou et al. (2015) designed a pipeline inspired by neurophysiological experiments (illustrated in figure 13.2*A*): the two trained networks were shown many versions of the same set of testing images in which a small patch of random pixels occluded each image version. For each tested image, 5,000 stimulus versions were created by sliding the localized occluder over the image. Feeding all these occluded stimulus versions into a network and recording the change in activation compared to the original image (without any occluded region) allowed the calculation of a "discrepancy map" for each unit in the network. Examples of discrepancy maps for a particular unit for the ten images to which the unit was most responsive are shown in figure 13.2*A*. To consolidate the information, the discrepancy maps are centered around the spatial location of the unit that provided the maximum activation (see figure 13.2*A*, *bottom row*) and then averaged to build a final receptive field for that unit (figure 13.2*A*, *right*).

Figure 13.2*B* shows a visualization of the discrepancy map of a few such units: specifically, the visualization

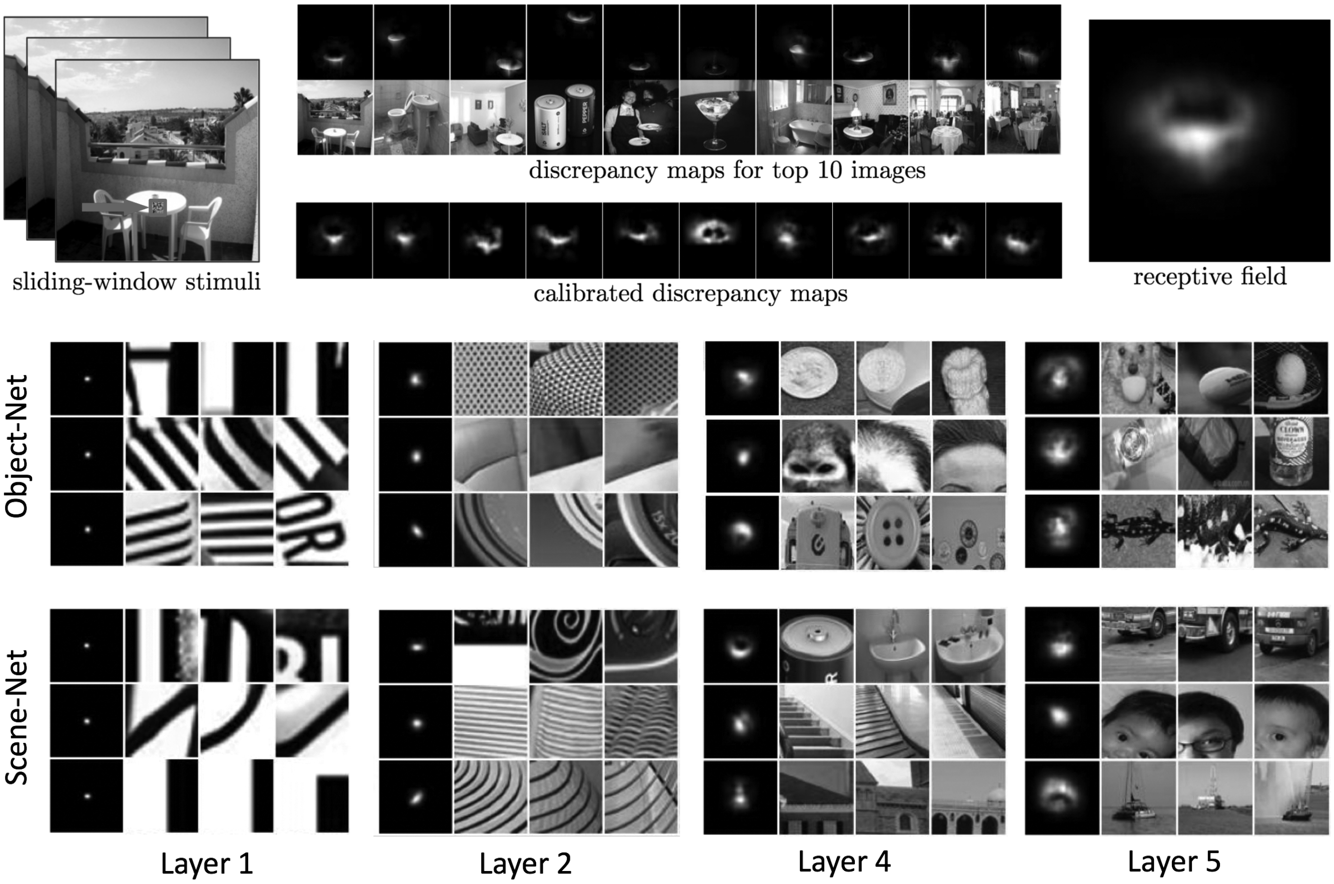

FIGURE 13.2 *A*, The pipeline for estimating the receptive field of each artificial unit of a DNN. Each sliding-window stimulus contains a small, randomized occluding patch (*red arrow*) at different spatial locations. By comparing the activation response of each sliding-window occluder with the activation response of the original image, a discrepancy map for each image can be built (*middle top*). By summing up the calibrated discrepancy maps (*middle bottom*) for the top-ranked images (i.e. 10), we can visualize the receptive field of that unit (*right*). *B*, The receptive fields of three units of pool1, pool2, conv4, and pool5 layers, respectively, for Object-Net (AlexNet trained with the ImageNet data set) and Scene-Net (AlexNet trained with Places data set), along with three images that correspond to particular images that activated these units the most. As the layers go deeper, the receptive field size gradually increases. (See color plates 16 and 17.)

shows the particular image patterns that drive the response of selected units in layers one, two, four, and five of AlexNet. For each unit's receptive field, three examples are shown that correspond to particular images that activated this neuron the most. Note that to robustly evaluate the true selectivity of a particular artificial unit, many images need to be passed through the network (Zhou et al., 2015).

As illustrated in figure 13.2*B*, as the layers deepen we can see that the units are tuned to increasingly complex features and sensitive to larger regions of the visual field—that is, they exhibit a larger receptive field, akin to the neurons found in the ventral visual-processing stream of primate brains.

Furthermore, as illustrated in figure 13.2*B*, the Object network and the Scene network learn highly similar patterns of oriented edges, curves, and textures in their two first layers. These results are most likely due to the similar distribution of image statistics across the two image data sets, as both are made of natural images (Torralba & Oliva, 2003). However, the properties of units in subsequent, higher layers of the network differ (see also Zeiler & Fergus, 2014): while a network trained to categorize objects (Object-Net in figure 13.2*B*) showed a predominance of units tuned to shapes and object parts, the network trained to classify scenes (Scene-Net in figure 13.2*B*) showed units selective to patterns that resemble whole objects, as well as more elaborate spatial layout patterns. This is expected given the hierarchical structure and the cost function the CNNs are programmed to solve: finding the discriminant or diagnostic information that maximizes performance on a specific task. As scene classes are mainly differentiated from the objects they are made of (i.e., a sofa makes a living room, a stove makes a kitchen), the network learned that patterns that we call "objects" are diagnostic features of scene classes. An emerging property is that both Object-Net and Scene-Net show units that seem specific to faces, people, and body parts, even as these networks did not explicitly learn to associate images to labels of faces, people, or body parts. Yet the emergence of such units suggests that the networks may have implicitly learned faces and body parts as diagnostic features that are contextually related to objects—for example, that baby faces are contextually related to *nursery*, or heads are contextually related to *hats*.

Human-interpretable concepts (i.e., a hat, a face, or an oval shape) emerge as *latent* variables in DNNs trained to solve detection or recognition tasks (Bau et al., 2017). These networks are not forced to decompose the classification problem in any interpretable way: the units do not have to specialize for particular patterns nor specialize for patterns that make sense to us (like object shapes, parts of objects, and more). However, these networks show interpretable units (figure 13.2*B*), which suggests they are able to spontaneously learn representations that are *disentangled* (Bau et al., 2017). A disentangled representation learns separate variables for separate meaningful features (Bengio, Courville, & Vincent, 2013)—for example, a concept or word that a person would use to describe a scene. While a network can learn an efficient encoding that makes economical use of hidden variables to distinguish between inputs, an internal structure with disentangled representation is an interesting property for comparing deep network representation to neural brain representations.

Applications of Deep Neural Networks to Investigate Cortical Regions of Object and Scene Processing

One can derive an algorithmically informed view on visual processing in the human brain by comparing, for the same set of images, the responses of a cortical region to the representation calculated from models. Because DNNs are made of a series of layers, a representation can be extracted from each layer for each image input. As shown in figure 13.2*B*, units in early layers of the DNN are more responsive to simple features like lines and corners, whereas units in later layers can be selective to complex patterns like shapes and object parts. This hierarchical representation of objects within the DNN resembles the hierarchical neural representation found in the primate visual brain (Grill-Spector & Weiner, 2014).

Using different neural-imaging techniques (i.e., electrophysiology, functional magnetic resonance imaging [fMRI], magneto/electroencephalography [M/EEG]), several recent studies have shown a systematic hierarchical relationship between convolutional neural network layers, such as in AlexNet (figure 13.1) and the processing cascade of information in both ventral and dorsal visual pathways in the human brain (Cadieu et al., 2014; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016b; Khaligh-Razavi, Cichy, Pantazis, & Oliva, 2018; Khaligh-Razavi & Kriegeskorte, 2014; Scholte, Losch, Ramakrishnan, de Haan, & Bohte, 2017; Seeliger et al., 2017; Yamins et al., 2014, among others). The relationship between model and brain data can be estimated using a technique termed *representational similarity analysis*, or RSA (Kriegeskorte, Mur, & Bandettini, 2008). Briefly, RSA builds dissimilarity distance matrices (referred to as a representational dissimilarity matrices, or RDMs) between every pair of images for which data are produced. For instance, for a data set of 100 images, an RDM can be built as a 100-by-100 matrix of distances between each pair of images. The RDMs can be built from the artificial units of a DNN layer, or from fMRI voxels activity in a brain region, or from the neuromagnetic signals captured from MEG sensors. The distributed nature of neural network data, where the response to an image input is best characterized as a pattern across artificial units within a layer, makes RSA an efficient analysis framework for relating the brain and models.

In the case of object processing, most studies have focused on comparing neural networks with early visual cortical (EVC) areas and inferior-temporal (IT) regions of the brain (Cadieu et al., 2014; Cichy et al. 2016b; Khaligh-Razavi & Kriegeskorte, 2014; Khaligh-Razavi et al., 2018; Yamins et al., 2014). Figure 13.3 illustrates results from such a study by Cichy et al. (2016b): the authors compared the RDMs from fMRI responses to object images to those from Object DNN (specifically, an AlexNet architecture trained on hundreds of object categories). This yielded eight brain maps (one for each of the eight layers of AlexNet; see figure 13.1) identifying the cortical regions where representations in the object DNN most correlated with cortical brain responses. As expected, for early model layers, similarities of visual representations were confined to the occipital lobe (i.e., the low-level and midlevel visual regions), and for late model layers, similarities of visual representations were found to correlate with more anterior regions in both the ventral and dorsal visual streams.

The comparison of DNNs for scene processing follows a similar pattern of results, with studies focusing on comparing scene models with brain regions previously identified to be more selective to images of places and scene layout than to other images (i.e., the parahippocampal place area, or PPA; Epstein & Kanwisher, 1998; the occipital place area, or OPA; Dilks, Julian, Paunov, & Kanwisher, 2013). For instance, Bonner and Epstein (2018)

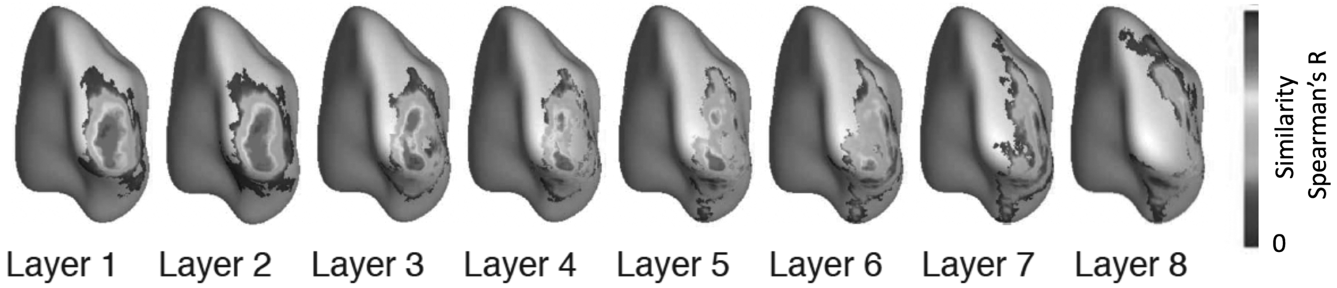

FIGURE 13.3 Spatial maps of visual representations common to human brain and object DNNs. The colors and the color bar show the strength of the brain-DNN similarity (similarity measured as Spearman's correlation, from 0 in *blue*, to a significant value in *red*, the maximum correlation varying between 0.07 and 0.14 depending on the layers). Low layers (one through two) have significant representational similarities with the occipital lobe of the brain—that is, low-level and midlevel visual regions. Higher layers have significant representational similarities with more anterior regions in the temporal and parietal lobe, with layers seven and eight reaching far into the inferior temporal cortex and inferior parietal cortex (from Cichy et al., 2016b). (See color plate 18.)

found that responses of a scene CNN to place images (Zhou et al., 2014, 2017) were highly predictive of human fMRI responses in the OPA, a scene-selective region of the dorsal occipitoparietal cortex. A visualization (as in figure 13.2***B***) of the feature selectivities in the scene network showed layout-type patterns, junctions, and large surfaces diagnostic of pictures of environments (see also Cichy, Khosla, Pantazis, & Oliva, 2017).

Conclusion

The isomorphism found between the human brain and DNNs suggests that hierarchical architectures are well suited for processing complex visual stimuli like real-world visual objects and scenes. This isomorphism also suggests that hierarchical systems of visual object and scene representations may emerge in both the human ventral and dorsal visual streams as the result of task constraints posed in everyday life (i.e., object categorization, place recognition for navigation).

While this chapter focused on the inner workings of only one class of DNN, neural network models simulating the neural dynamics of sensory information processing should have a better chance at representing the true spatiotemporal dynamics of vision. Decades of work in neuroscience suggest that vision starts as an initial feedforward phase, with signals going through the ventral stream (feedforward networks), followed by recurrent, long-range reverberating interactions between cortical regions (recurrent networks). Conceptually, feedforward and recurrent networks are sibling architectures, with the layered feedforward framework being a special case of a recurrent network in which some connections are missing (i.e., weights of connections between units are put at zero). Many neuroscience experiments suggest that performance at more attention-demanding tasks, such as object recognition in clutter, under occlusion, or with segmentation, are better approximated by a recurrent network. The basic assumption is that recurrent connections from higher to lower layers allow for looping back output to input, providing a temporal context to process the current input. Given that the state of the art in DNNs is evolving at an exponential pace, recurrent DNN models have a tremendous potential for explaining and operationalizing how biological brain networks learn diagnostic features, develop expertise, represent sensory information, solve recognition tasks, forecast the future, and never stop learning.

Acknowledgments

The author would like to thank Radoslaw Cichy, Matthew Lowe, Mathew Monfort, and Yalda Mohsenzadeh for thoughtful comments and feedback. This chapter was supported by the Vannevar Bush Faculty Fellowship program funded by the ONR grant No. N00014-16-1-3116 to A. Oliva.

REFERENCES

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. Proceedings of the 30th IEEE Conference on *Computer Vision and Pattern Recognition*, 3319–3327.

Bengio, Y. (2009). Learning deep architectures for AI. Hanover, MA: Now.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 1798–1828.

Bengio, Y., & Lecun, Y. (2007). Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, &

J. Weston (Eds.), *Large-scale kernel machines*. Cambridge, MA: MIT Press.

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *Plos Computational Biology*, 14(4), e1006111.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *Plos Computational Biology*, 10(12), e1003963.

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

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

Cichy, R. M., Pantazis, D., & Oliva, A. (2016a). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8), 3563–3579.

Deng, J., Dong, W., Socher, R., Li, L.-J., Lim, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Proceedings of the 22nd Conference on Computer Vision and Pattern Recognition,* 248–255.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, 33(4), 1331–1336.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, 24(1), 187–203.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Neuroscience Review*, 15, 536–548.

Khaligh-Razavi, S. M., Cichy, R. M., Pantazis, D., & Oliva, A. (2018). Tracking the spatiotemporal neural dynamics of object properties in the human brain. *Journal of Cognitive Neuroscience*, 30(11), 1–18.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *Plos Computational Biology*, 10(11), e1003915.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipito-temporal cortex. *Neuron*, 74(6), 1114–1124.

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

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21, 1148–1160.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 24(2), 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech and time series. In M. A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks* (pp. 255–258). Cambridge, MA: MIT Press.

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

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18), 8135–8139.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11), 1019–1025.

Scholte, S., Losch, M. M., Ramakrishnan, K., de Haan, E. H. F., & Bohte, S. M. (2017). Visual pathways from the perspective of cost functions and multi-task deep neural networks. *Cortex*, 98, 249–261.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J. M., Bosch, S. E., & van Gerven, M. A. J. (2017). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180, 243–256.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 411–426.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. Paper presented at the International Conference on Learning Representation, April 14–16, 2014, Banff, Canada.

Torralba, A., & Oliva, A. (2003). Statistics of natural images categories. *Network: Computation in Neural Systems*, 14, 391–412.

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*, 111, 8619–8624.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the 13th European Conference on Computer Vision* (pp. 818–833). New York: Springer.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object detectors emerge in deep scene CNNs. Paper presented at the International Conference on Learning Representations, May 7–9, 2015, San Diego, USA.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1452–1464.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. Advances in Neural Information Processing Systems (NIPS 2014), 27, Montreal, Canada.