Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future

Grace W. Lindsay

Abstract

Convolutional neural networks (CNNs) were inspired by early findings in the study of biological vision. They have since become successful tools in computer vision and state-of-the-art models of both neural activity and behavior on visual tasks. This review highlights what, in the context of CNNs, it means to be a good model in computational neuroscience and the various ways models can provide insight. Specifically, it covers the origins of CNNs and the methods by which we validate them as models of biological vision. It then goes on to elaborate on what we can learn about biological vision by understanding and experimenting on CNNs and discusses emerging opportunities for the use of CNNs in vision research beyond basic object recognition.

INTRODUCTION

Computational models serve several purposes in neuroscience. They can validate intuitions about how a system works by providing a way to test those intuitions directly. They also offer a means to explore new hypotheses in an ideal experimental testing ground, wherein every detail can be controlled and measured. In addition models open the system in question up to a new realm of understanding through the use of mathematical analysis. In recent years, convolutional neural networks (CNNs) have performed all of these roles as a model of the visual system.

This review covers the origins of CNNs, the methods by which we validate them as models of the visual system, what we can find by experimenting on them, and emerging opportunities for their use in vision research. Importantly, this review is not intended to be a thorough overview of CNNs or extensively cover all uses of deep learning in the study of vision (other reviews may be of use to the reader for this; Kietzmann, McClure, & Kriegeskorte, 2019; Kriegeskorte & Golan, 2019; Serre, 2019; Storrs & Kriegeskorte, 2019; Yamins & DiCarlo, 2016, Kriegeskorte, 2015). Rather, it is meant to demonstrate the strategies by which CNNs as a model can be used to gain insight and understanding about biological vision.

According to Kay (2018), “a functional model attempts only to match the outputs of a system given the same inputs provided to the system, whereas a mechanistic model attempts to also use components that parallel the actual physical components of the system.” Using these definitions, this review is concerned with the use of CNNs as “mechanistic” models of the visual system. That is, it will be assumed and argued that, in addition to an overall match between outputs of the two systems, subparts of a CNN are intended to match subparts of the visual system.

WHERE CNNS CAME FROM

The history of CNNs threads through both neuroscience and artificial intelligence. Like artificial neural networks in general, they are an example of brain-inspired ideas coming to fruition through an interaction with computer science and engineering.

Origins of the Model

In the mid twentieth century, Hubel and Wiesel discovered two major cell types in the primary visual cortex (V1) of cats (Hubel & Wiesel, 1962). The first type—the simple cells—responds to bars of light or dark when placed at specific spatial locations. Each cell has an orientation of the bar at which it fires most, with its response falling off as the angle of the bar changes from this preferred orientation (creating an orientation “tuning curve”). The second type—complex cells—has less strict response profiles; these cells still have preferred orientations but can respond just as strongly to a bar in several different nearby locations. Hubel and Wiesel concluded that these complex cells are likely receiving input from several simple cells, all with the same preferred orientation but with slightly different preferred locations (Figure 1, left).

In 1980, Fukushima transformed Hubel and Wiesel’s findings into a functioning model of the visual system (Fukushima, 1980). This model, the Neocognitron, is the predecessor to modern CNNs. It contains two main cell
The S-cells are named after simple cells and replicate their basic features: Specifically, a 2-D grid of weights is applied at each location in the input image to create the S-cell responses. A "plane" of S-cells thus has a retinotopic layout with all cells sharing the same preferred visual features and multiple planes existing at a layer. The response of the C-cells (named after complex cells) is a nonlinear function of several S-cells coming from the same plane but at different locations.

After a layer of simple and complex cells representing the basic computations of V1, the Neocognitron simply repeats the process again. That is, the output of the first layer of complex cells serves as the input to the second simple cell layer, and so on. With several repeats, this creates a hierarchical model that mimics not just the operations of V1 but the ventral visual pathway as a whole. The network is "self-organized," meaning weights change with repeated exposure to unlabeled images.

By the 1990s, many similar hierarchical models of the visual system were being explored and related back to data (Riesenhuber & Poggio, 2000). One of the most prominent ones, HMAX, used the simple "max" operation over the activity of a set of simple cells to determine the response of the C-cells and was very robust to image variations. Because these models could be applied to the same images used in human psychophysics experiments, the behavior of a model could be directly compared with the ability of humans to perform rapid visual categorization. Through this, a correspondence between these hierarchical models and the first 100–150 msec of visual processing was found (Serre et al., 2007). Such models were also fit to capture the responses of V4 neurons to complex shape stimuli (Cadieu et al., 2007).

**CNNs in Computer Vision**

The CNN as we know it today comes from the field of computer vision, yet the inspiration from the work of Hubel and Wiesel is clearly visible in it (Figure 1; Rawat & Wang, 2017). Modern CNNs start by convolving a set of filters with an input image and rectifying the outputs, leading to "feature maps" akin to the planes of S-cells in the Neocognitron. Max pooling is then applied, creating complex cell-like responses. After several iterations of this pattern, nonconvolutional fully connected layers are added, and the last layer contains as many units as number of categories in the task to output a category label for the image (Figure 2, bottom).

**Figure 1.** The relationship between components of the visual system and the base operations of a CNN. Hubel and Wiesel (1962) discovered that simple cells (left, blue) have preferred locations in the image (dashed ovals) wherein they respond most strongly to bars of particular orientation. Complex cells (green) receive input from many simple cells and thus have more spatially invariant responses. These operations are replicated in a CNN (right). The first convolutional layer (blue) is produced by applying a convolution operation to the image. Specifically, the application of a small filter (gray box) to every location in the image creates a feature map. A convolutional layer has as many feature maps as filters applied (only one shown here). Taking the maximal activation in a small section of each feature map (gray box) downsamples the image and leads to complex cell-like responses (green). This is known as a "max-pooling" operation, and the downsamplied feature maps produced by it comprise a pooling layer.

**Figure 2.** Validating CNNs as a model by comparing their representations to those of the brain. A CNN trained to perform object recognition (bottom) contains several convolutional layers (Conv-1 and Conv-2 shown here, each followed by max-pooling layers Pool-1 and Pool-2) and ends with a fully connected layer with a number of units equal to the number of categories. When the same image is shown to a CNN and an animal (top), the response from each can be compared. Here, the activity of single units recorded from a visual area in the brain can be predicted from the activity of artificial units at a particular layer.
The first major demonstration of the power of CNNs came in 1989 when it was shown that a small CNN trained with supervision using the backpropagation algorithm could perform handwritten digit classification (LeCun et al., 1989). However, these networks did not really take off until 2012, when an eight-layer network (dubbed “AlexNet,” “Standard Architecture” in Figure 3) trained with backpropagation far exceeded state-of-the-art performance on the ImageNet challenge. The ImageNet data set is composed of over a million real-world images, and the challenge requires classifying an image into one of a thousand object categories. The success of this network demonstrated that the basic features of the visual system found by neuroscientists were indeed capable of supporting vision; they simply needed appropriate learning algorithms and data.

In the years since this demonstration, many different CNN architectures have been explored, with the main parameters varied including network depth, placement of pooling layers, number of feature maps per layers, training procedures, and whether or not residual connections that skip layers exist (Rawat & Wang, 2017). The goal of exploring these parameters in the computer vision community is to create a model that performs better on standard image benchmarks, with secondary goals of making networks that are smaller or train with less data. Correspondence with biology is not a driving factor.

VALIDATING CNNS AS A MODEL OF THE VISUAL SYSTEM

The architecture of a CNN has (by design) direct parallels to the architecture of the visual system. Images fed into these networks are usually first normalized and separated into three different color channels (red, green, blue), which captures certain computations done by the retina. Each stacked bundle of convolution–nonlinearity–pooling can then be thought of as an approximation to a single visual area—usually the ones along the ventral stream such as V1, V2, V4, and IT—each with its own retinotopy and feature maps. This stacking creates receptive fields for individual neurons that increase in size deeper in the network and the features of the image that they respond to become more complex. As has been mentioned, when trained to, these architectures can take in an image and output a category label in agreement with human judgment.

All of these features make CNNs good candidates for models of the visual system. However, all of these features have been explicitly built into CNNs. To validate that CNNs are performing computations similar to those of the visual system, they should match the visual system in additional, nonengineered ways. That is, from the assumptions put into the model, further features of the data should fall out. Indeed, many further correspondences have been found at the neural and behavioral levels.

Comparison at the Neural Level

One of the major causes of the recent resurgence of interest in artificial neural networks among neuroscientists is the finding that they can recapitulate the representation of visual information along the ventral stream. In particular, when CNNs and animals are shown the same

---

**Figure 3.** Example architectures. The standard architecture shown here is similar to the original AlexNet, where an red/green/blue image is taken as input and the activity at the final layer determines the category label. In the second architecture, local recurrence is shown at each convolutional layer and feedback recurrence goes from Layer 5 to Layer 3. This network would be trained with the backpropagation-through-time algorithm. The third network is an autoencoder. Here, the goal of the network is to recreate the input image. Crucially, the fully connected layer is of significantly lower dimensionality than the image, forcing the network to find a more compact representation of relevant image statistics. Finally, an architecture for reinforcement learning is shown. Here, the aim of the network is to turn an input image representing information about the current state into an action for the network to take. This architecture is augmented with a specific type of local recurrent units known as LSTMs, which embue it with memory.
image (Figure 2), the activity of the artificial units can be used to predict the activity of real neurons, with accuracy beyond that of any previous methods.

One of the early studies to show this (Yamins et al., 2014), published in 2014, recorded extracellular activity in macaque during the viewing of complex object images. Regressing the activity of a real V4 or IT neuron onto the activity of the artificial units in a network (and cross-validating the predictive ability with a held-out test set), the authors found that networks that performed better on an object recognition task also better predicted neural activity (a relationship also found using video classification; Tacchetti, Isik, & Poggio, 2017). Furthermore, the activity of units from the last layer of the network best predicted IT activity and the penultimate layer best predicted V4. This relationship between models and the brain wherein later layers in the network better predict higher areas of the ventral stream has been found in several other studies, including using human fMRI (Güçlü & van Gerven, 2015), MEG (Seeleiger et al., 2018), and with video instead of static images as the stimuli (Eickenberg, Gramfort, Varoquaux, & Thirion, 2017).

Another method to check for a correspondence between different populations is representational similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008). This method starts by creating a matrix for each population that represents how dissimilar the responses of that population are for every pair of images. This serves as a signature of the population’s representation properties. The similarity between two different populations is then measured as the correlation between their dissimilarity matrices. This method was used in 2014 (Khaligh-Razavi & Kriegeskorte, 2014) to demonstrate that the later layers of an AlexNet network trained on ImageNet match multiple higher areas of the human visual system, along with monkey IT, better than previously used models.

Because dissimilarity matrices can be created from any kind of responses, including behavioral outputs, RSA is widely applicable as a means to compare data from different experimental methods and models. It is also a straightforward way to incorporate and compare full population responses, whereas regression techniques focus on a single neuron or voxel at a time. On the other hand, the regression techniques allow for selective weighting of the model features most relevant for fitting the data, which may be more informative than the more “unsupervised” RSA approach. In all cases, details of the methodology and interpretation should be carefully considered (Kornblith, Norouzi, Lee, & Hinton, 2019; Thompson, Bengio, Formisano, & Schönwiesner, 2018).

Many of the studies comparing CNNs to biological data have highlighted their ability to explain later visual areas such as V4 and IT. This is a notable feat of CNNs because the complex response properties of these areas have made them notoriously difficult to fit compared with primary visual cortex. Recent work, however, has shown that early-to-middle layers of task-trained CNNs can also predict V1 activity beyond the ability of more traditional V1 models (Cadena, Denfield, et al., 2019).

Beyond predicting neural activity or matching overall representations, CNNs can also be compared with neural data using more specific features traditionally used in systems neuroscience. Similarities have been found between units in the network and individual neurons in terms of response sparseness and size tuning, but differences have been found for object selectivity and orientation tuning (Tripp, 2017). Other studies have explored tuning to shape and categories (Zeman, Ritchie, Bracci, & Op de Beeck, 2019) and how responses change with changes in pose, location, and size (Murty & Arun, 2017, 2018; Hong, Yamins, Majaj, & DiCarlo, 2016).

Generally, the similarities with real neural activity that emerge from training a CNN to perform object recognition suggest that this architecture and task do indeed have some similarity to the architecture and purpose of the visual system.

Comparison at the Behavioral Level

Insofar as CNNs outperform any previous model of the visual system on real-world image classification, they can be considered a good match to human behavior. However, overall accuracy on the standard ImageNet task is only one measure of CNN behavior, and it is the one for which the network is explicitly optimized.

A deeper comparison can be made by looking at the errors these networks make. Although there is only one way to be correct, there are many ways humans and models can make mistakes in classification, and this can provide rich insight into their respective workings. Confusion matrices, for example, are used to represent how frequently images from one category are classified as belonging to another and can be compared between models and animal behavior. A large-scale study showed that several different deep CNN architectures show similar matches to human and monkey object classification, though the models are not predictive down to the image level (Rajalingham et al., 2018).

Other studies have explicitly asked participants to rate similarity between images and compared human judgment to model representations (King, Groen, Steel, Kravitz, & Baker, 2019; Jozwik, Kriegeskorte, Storrs, & Mur, 2017). In the study of Rosenfeld, Solbach, and Tsotsos (2018), for example, a large data set was taken from the Web site “Totally Looks Like.” Although many similarity studies find good matches from CNNs, this study demonstrated more challenging elements of similarity that CNNs struggled to replicate.

In addition to similarity, other psychological concepts that have been tested in CNNs include typicality (Lake, Zaremba, Fergus, & Gureckis, 2015), Gestalt principles (Kim, Reif, Wattcnberg, & Bengio, 2019), and animacy (Bracci, Ritchie, Kalfas, & Op de Beeck, 2019). In Jacob, Pramod, Katti, and Arun (2019), a battery of tests inspired
by findings in visual psychophysics were applied to CNNs, and CNNs were found to be similar to biological vision according to roughly half of them.

Another way to probe animal and CNN behavior is to make the classification task more challenging by degrading image quality. Several studies have added various types of noise, occlusion, or blur to standard images and observed a decrease in classification performance (Geirhos, Temme, et al., 2018; Tang et al., 2018; Geirhos et al., 2017; Wichmann et al., 2017; Ghodrati, Farzamahdi, Rajaei, Ebrahimpour, & Khaligh-Razavi, 2014). Importantly this performance decrease is usually more severe in the CNNs than it is in humans, suggesting biological vision has mechanisms for overcoming degradation. These points of mismatch are thus important to identify to steer future research (e.g., see Alternative Data Sets section). A particular CNN architecture, known as a capsule network (Roy, Ghosh, Bhattacharya, & Pal, 2018), was shown to be more robust to degradation; however, it is not exactly clear how to relate the “capsules” in this architecture to parts of the visual system.

Another emerging finding in the behavioral analysis of CNNs is their reliance on texture. Although an argument has been made for CNNs as a model of human shape sensitivity (Kubilius, Bracci, & Op de Beeck, 2016), other studies have demonstrated that CNNs rely too much on texture and not enough on shape when classifying images (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos, Rubisch, et al., 2018).

Interestingly, it is possible for certain deep networks to outperform humans on some tasks (Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016). This highlights an important tension between the goals of computer vision and those of neuroscience: In the former exceeding human performance is desirable, in the latter it is still counted as a mismatch between the model and the data.

Something to keep in mind when studying CNN behavior is that standard feedforward CNN architectures are believed to represent the very initial stages of visual processing, before various kinds of recurrent processing can take place. Therefore, when comparing the behavior of CNNs to animal behavior, fast stimulus presentation and backward masking are advised, as these are known to prevent many stages of recurrent processing (Tang et al., 2018).

**Other Forms of Validation**

Although neural and behavioral comparisons are the main methods by which CNNs are validated as models of the visual system, other approaches can provide further support.

Methods for visualizing the image features that drive units at different layers in the network (Figure 4; Olah, Mordvintsev, & Schubert, 2017; Zeiler & Fergus, 2014) have revealed preferred visual patterns that align with those found in neuroscience. For example, the first layer in a CNN has filters that look like Gabors, whereas later layers respond to partial object features and eventually fuller features such as faces. This supports the idea that the processing steps in CNNs match those in the ventral stream.

CNNs have also been used to produce optimal stimuli for real neurons. Starting from the above-mentioned procedure for predicting neural activity with CNNs, stimuli that were intended to maximally drive a neuron’s firing rate were produced (Bashivan, Kar, & DiCarlo, 2019). The fact that the resulting stimuli were indeed effective at driving the neuron beyond its normal rates despite being unnatural and unrelated to the images the network was trained on further supports the notion that CNNs are capturing something fundamental about the visual processing stream.

---

**Figure 4.** Visualizing preferences of different network components. The general method for creating these images is shown at the top. Specifically, a component of the model is selected for optimization and gradient calculations sent back through the network indicate how pixels in an initially noisy image should change to best activate the chosen component. Components can be individual units, whole feature maps, layers (using the DeepDream algorithm), or units representing a specific class in the final layer. Feature visualizations taken from Olah et al. (2017).
In a related vein, studies have also used CNNs to decode neural activity to recreate the stimuli that are being presented to participants (Shen, Horikawa, Majima, & Kamitani, 2019).

**WHAT WE LEARN FROM VARYING THE MODEL**

The above-mentioned studies have focused mainly on standard feedforward CNN architectures, trained via supervised learning to perform object recognition. Yet, with full control over these models, it is possible to explore many variations in data sets, architectures, and training procedures. By observing how these changes make the model a better or worse fit to data, we can gain insight into how and why certain features of biological visual processing exist.

**Alternative Data Sets**

The ImageNet data set has proven very useful for learning a set of basic visual features that can be adapted to many different tasks in a way that previous smaller and simpler data sets such as MNIST and CIFAR-10 were not. However, it contains images where objects are the focus and as a result is best suited for studying object recognition pathways. To study scene processing areas such as the occipital place area, several studies have instead trained on scene images. In the study of Bonner and Epstein (2018), the responses of occipital place area could be predicted using a network trained to classify scenes. What’s more, the authors were able to relate the features captured by the network to navigational affordances in the scene. In the study of Cichy, Khosla, Pantazis, and Oliva (2017), scene-trained networks were able to capture representations of scene size collected using MEG. Such studies can in general help to identify the evolutionarily or developmentally determined role of different brain areas based on the extent to which their representations align with networks trained on different specialized data sets. For an open data set of fMRI responses to different image sets, see Chang et al. (2019).

Data sets have also been developed with the explicit intent of correcting ways in which CNNs do not match human behavior. For example, training with different image degradations can make networks more robust to those degradations (Geirhos, Temme, et al., 2018). However, this only works for the noise model specifically trained and does not generalize to new noise types. In addition, in the study of Geirhos, Rubisch, et al. (2018), a data set that keeps abstract shapes but varies low-level textural elements was shown to decrease a CNN’s texture bias. Whether animals have good visual performance due to an exposure to similarly diverse data during developmental or due to built-in priors instead remains to be determined.

**Alternative Architectures**

A variety of purely feedforward architectures have been explored both in the machine learning community and by neuroscientists. In comparison to data, AlexNet has been commonly used and performs well. But it can be beat by somewhat deeper architectures such as VGG models and ResNets (Schrömpf et al., 2018; Wen, Shi, Chen, & Liu, 2018). With very deep networks—though they may perform better on image tasks—the relationship between layers in the network and areas in the brain may break down, making them a worse fit to the data. However, in some cases, the processing in very deep networks can be thought of as equivalent to the multiple stages of recurrent processing in the brain (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019).

In primate vision, the first two processing stages—retina and the lateral geniculate nucleus—contain cells that respond preferentially to on-center or off-center light patterns, and only in V1 does orientation tuning become dominant. Yet, Gabor filters are frequently learned at the first layer of a CNN, creating a V1-like representation. Using a modified architecture wherein the connections from early layers are constrained as they are biologically by the optic nerve creates a CNN with on- and off-center responses at early stages and orientation tuning only later (Lindsey, Ocko, Ganguli, & Deny, 2019). Thus, the pattern of selectivity seen in primate vision is potentially a consequence of anatomical constraints.

In the study of Kell, Yamins, Shook, Norman-Haignere, and McDermott (2018), various architectures were trained to perform a pair of auditory tasks (speech and music recognition). Through this, the authors found that the two tasks could share three layers of processing before the network needed to split into specialized streams to perform well on both tasks. Recently, a similar procedure was applied to visual tasks and used to explain why specialized pathways for face processing arise in the visual system (Dobs, Kell, Palmer, Cohen, & Kanwisher, 2019). A similar question was also approached by training a single network to perform two tasks and looking for emergent subnetworks in the trained network (Scholte, Losch, Ramakrishnan, de Haan, & Bohte, 2018). Such studies can explain the existence of dorsal and ventral streams and other details of visual architectures.

Taking further inspiration from biology, many studies have explored the beneficial role of both local and feedback recurrence (Figure 3). Local recurrence refers to horizontal connections within a single visual area. By adding these connections to CNNs, studies have found that these recurrent connections make networks better at more challenging tasks (Hasani, Soleymani, & Aghajan, 2019; Montobbio, Bonnasse-Gahot, Citti, & Sarti, 2019; Spoerer, Kietzmann, & Kriegeskorte, 2019; Tang et al., 2018). These connections can also help make the CNN representations a better match to neural data, particularly for challenging images and at later time points in the response (Kar et al., 2019; Kubilius et al., 2019; Rajaei, Molsenzedeh, Ebrahimpour, & Khaligh-Razavi, 2019; Shi, Wen, Zhang, Han, & Liu, 2018; McIntosh, Maheswaranathan, Nayebi, Ganguli, & Baccus, 2016). These studies make a strong argument for the computational role of these local connections.
Feedback connections go from frontal or parietal areas to regions in the visual system or from higher visual areas back to lower areas (Wyatte, Jilk, & O’Reilly, 2014). Like horizontal recurrence, they are known to be common in biological vision. Connections from frontal and parietal areas are believed to implement goal-directed selective attention. Such feedback has been added to network models to implement cued detection tasks (Thorat, van Gerven, & Peelen, 2019; Wang, Zhang, Song, & Zhang, 2014). Feedback from higher visual areas back to lower ones are thought to implement more immediate and general image processing such as denoising. Some studies have added these connections in addition to local recurrence and found that they can aid performance as well (Kim, Linsley, Thakkar, & Serre, 2019; Spoerer, McClure, & Kriegeskorte, 2017). A study comparing different feedback and feedforward architectures suggests that feedback can help in part by increasing the effective receptive field size of cells (Jarvers & Neumann, 2019).

Alternative Training Procedures

Supervised learning using backpropagation is the most common method of training CNNs; however, other methods have the potential to result in a good model of the visual system. The 2014 study that initially showed a correlation between performance on object recognition and ability to capture neural responses (Yamins et al., 2014), for example, did not use backpropagation but rather a modular optimization procedure.

Unsupervised learning, wherein networks aim to capture relevant statistics of the input data rather than match inputs to outputs, can also be used to train neural networks (Fleming & Stors, 2019; Figure 3). These methods may help identify a low-dimensional set of features that underlie high-dimensional visual inputs and thus allow an animal to make better sense of the world and possibly build useful causal models (Lake, Ullman, Tenenbaum, & Gershman, 2017). Furthermore, because of the large amount of labeled data points required for supervised training, it is assumed that the brain must make use of unsupervised learning. However, as of yet, unsupervised methods do not produce models that capture neural representations as well as supervised methods. Behaviorally, a generative model was shown to perform much worse than supervised models on capturing human image categorization (Peterson, Abbott, & Griffiths, 2018). In addition, a model trained based on the concept of predictive coding (Lotter, Kreiman, & Cox, 2016) was able to predict object movement and replicate motion illusions (Watanabe, Kitaoka, Sakamoto, Yasugi, & Tanaka, 2018). Overall, the limits and benefits of unsupervised training for vision models need further exploration.

Interestingly, a recent study found a class-conditioned generative model to be in better agreement with human judgment on controversial stimuli than models trained to perform classification (Golan, Raju, & Kriegeskorte, 2019). Although this generative model still relied on class labels for training and was thus not unsupervised, its ability to replicate aspects of human perception supports the notion that the visual system aims in part to capture a distribution of the visual world rather than merely funnel images into object categories.

A compromise between unsupervised and supervised methods is “semisupervised” learning, which has recently been explored as a means of making more biologically realistic networks (Zhuang, Yan, Nayebi, & Yamins, 2019; Zhuang, Zhai, & Yamins, 2019).

Reinforcement learning is the third major class of training in machine learning. In these systems, an artificial agent must learn to produce action outcomes in response to information from the environment, including rewards. Several such artificial systems have used convolutional architectures on the front end to process visual information about the world (Figure 3; Merel et al., 2018; Paine et al., 2018; Zhu et al., 2018). It would be interesting to compare the representations learned in the context of these models to those trained by other mechanisms, as well as to data.

A simple way to understand the importance of training for a network is to compare to a network with the same architecture but random weights. In Kim, Reif, et al. (2019), the ability of a network to perform perceptual closing existed only in networks that had been trained with natural images, not random networks.

The above methods rely on training data, such as a set of images, that do not come from neural recordings. However, it is possible to train these architectures to replicate neural activity directly. Doing so can help identify the features of the input image most responsible for a neuron’s firing (Cadena, Denfield, et al., 2019; Kindel, Christensen, & Zylberberg, 2019; Sinz et al., 2018). Ideally, the components of the model can be also related back to anatomical features of the circuit in question (Günthner et al., 2019; Maheswaranathan et al., 2018) as is done when networks are trained on classification tasks.

A final hybrid option is to train on a classification task while also using neural data to constrain the intermediate representations to be brain-like (Fong, Scheirer, & Cox, 2018), resulting in a network that better matches neural data and can perform the task.

HOW TO UNDERSTAND CNNS

Varying a network’s structure and training is one way to explore how it functions. Additionally, trained networks can be probed directly using techniques normally applied to brains or those only available in models (Samek & Müller, 2019). In either case, if we believe CNNs have been validated as a model of the visual system, any insight gained from probing how they work may apply to biological vision as well.

Empirical Methods

The tools of the standard neuroscientific toolbox—lesions, recordings, anatomical tracings, stimulations, silencing, and
so forth—are all readily available in artificial neural networks (Barrett, Morcos, & Macke, 2019) and can be used to answer questions about the workings of these networks.

“Ablating” individual units in a CNN, for example, can have an impact on classification accuracy; however, the impact of ablating a particular unit does not have a strong relationship with the unit’s selectivity properties (Morcos, Barrett, Rabinowitz, & Botvinick, 2018; Zhou, Sun, Bau, & Torralba, 2018).

In the study of Bonner and Epstein (2018), images were manipulated or occluded to determine which features were responsible for the CNN’s response to scenes, akin to how the function of real neurons is explored.

Other studies (Lindsay et al., 2019; Spoerer et al., 2019) have analyzed the connectivity properties of trained networks to see if they replicate features of visual cortical anatomy.

One conceptual framework to describe what the stages of visual processing are doing is that of “untangling.” High-level concepts that are intertwined in the pixel or retinal representation get pulled apart to form easily separable clusters in later representations. This theory has been developed using biological data (DiCarlo & Cox, 2007); however, recently, techniques for describing the geometry of these clusters (or manifolds) have been developed and used to understand the untangling process in deep neural networks (Cohen, Chung, Lee, & Sompolsky, 2019; Chung, Lee, & Sompolsky, 2018). This work highlights the relevant features of these manifolds for classification and how they change through learning and processing stages. This can help identify which response features to look for in data.

Mathematical Analyses

Given our full access to the equations that define a CNN, many mathematical techniques not currently applicable to brains can be performed. One such common tool is the calculation of gradients. Gradients indicate how certain components in the network affect others, potentially far away. They are used to train these networks by determining how a weight at one layer can be changed to decrease error at the output. However, they can also be used to visualize the preferred features of a unit in a network. In that case, the gradient calculation goes all the way through to the input image, where it can be determined how individual pixels should change to increase the activity of a particular unit (Olah et al., 2017; Figure 4). Multiple variants of feature visualization have also been used to probe neural network functions (Nguyen, Yosinski, & Clune, 2019), for example, by showing which invariances exist at different layers (Cadena, Weis, Gatys, Bethge, & Ecker, 2018).

Gradients can also be used to determine the role an artificial neuron plays in a classification task. In the study of Lindsay and Miller (2018), it was found that the role a unit plays in classifying an image as being of a certain category is not tightly correlated with how strongly it responds to images of that category. Much like the ablation study above, this demonstrates a disassociation between tuning and function that should give neuroscientists pause about how informative a tuning-based analysis of the brain is.

Machine learning researchers and mathematicians also aim to cast CNNs in a light that makes them more amenable to traditional mathematical analysis. The concepts of information theory (Tishby & Zaslavsky, 2015) and wavelet scattering, for example, have been used toward this end (Mallat, 2016). Another fruitful approach to mathematical analysis has been to study deep linear neural networks as this approximation makes more analyses possible (Saxe, McClelland, & Ganguli, 2013).

Several studies in computational neuroscience have trained simple recurrent neural networks to perform tasks and then interrogated the properties of the trained networks for clues as to how they work (Barak, 2017; Foerster, Gilmer, Sohl-Dickstein, Chorowski, & Sussillo, 2017; Sussillo & Barak, 2013). Going forward, more sophisticated techniques for understanding the learned representations and connectivity patterns of CNNs should be developed. This can both provide insight into how these networks work as well as indicate which experimental techniques and data analysis methods would be fruitful to pursue.

Are They Understandable?

For some researchers, CNNs represent the unavoidable trade-off between complexity and interpretability. To have a model complex enough to perform real-world tasks, we must sacrifice the desire to make simple statements about how each stage of it works—a goal inherent in much of systems neuroscience. For this reason, an alternative way to describe the network that is compact without relying on simple statements about computations has been proposed (Lillicrap & Kording, 2019; Richards et al., 2019); this viewpoint focuses on describing the architecture, optimization function, and learning algorithm of the network—instead of attempting to describe specific computations—because specific computations and representations can be seen as simply emerging from these three factors.

It is true that the historical aim of language-based descriptions of the roles of individual neurons or groups of neurons (e.g., “tuned to orientation” or “face detectors”) seems woefully incomplete as a way to capture the essential computations of CNNs. Yet, it also seems that there are still more compact ways to describe the functioning of these networks and that finding these simpler descriptions could provide a further sense of understanding. Only certain sets of weights, for example, allow a network to perform well on real-world classification tasks. As of yet, the main thing we know about these “good” weights is that they can be found through optimization procedures. But are there essential properties we could
identify that would result in a more condensed description of the network? The “lottery ticket” method of training can produce networks that work as well as dense networks using only a fraction of the weights; a recent study noted that as many as 95% of model weights could be guessed from the remaining 5% (Frankle & Carbin, 2018; Denil, Shakibi, Dinh, Ranzato, & de Freitas, 2013). Such findings suggest that more condensed (and thus potentially more understandable) descriptions of the computations of high-performing networks are possible.

A similar analysis of architectures could be done to determine which broad features of connectivity are sufficient to create good performance. Recent work using random weights, for example, helps to isolate the role of architecture versus specific weight values (Gaier & Ha, 2019). A more compact description of the essential features of a high-performing network is a goal for both machine learning and neuroscience.

One undeniably noncompact component of deep learning is the data set. As has been mentioned, large, real-world data sets are required to match the performance and neural representations of biological vision. Are we doomed to carry around the full ImageNet data set as part of our description of how vision works? Or can a set of sufficient statistics be defined? Natural scene statistics have been a historically large part of both computational neuroscience and computer vision (Simoncelli & Olshausen, 2001). Although much of that work has focused on lower order correlations that are insufficient to capture the relevant features for object recognition, the time seems ripe for explorations of higher order image statistics, particularly as advances in generative modeling (Salakhutdinov, 2015) point to the ability to condense full and complex features of natural images into a model.

In any case, nearly all the critiques against the interpretability of CNNs could equally apply to biological networks. Therefore, CNNs make a good testing ground for deciding what understanding looks like in neural systems.

**BEYOND THE BASICS**

Since 2014, an explosion of research has answered many questions about how varying different features of a CNN can change its properties and its ability to match data. These findings, in turn, have aided progress in understanding both the “how” and “why” of biological vision. Beyond this, having access to an “image computable” model of the visual system opens the door to many more modeling opportunities that can explore more than just core vision.

**Exploring Cognitive Tasks**

The relationship between image encoding and memorability was explored in Jaegle et al. (2019). The authors showed that the overall magnitude of the response of later layers of a CNN correlated with how memorable images were found to be experimentally.

In Devereux, Clarke, and Tyler (2018), an attractor network based on semantic features was added to the end of a CNN architecture. This additional processing stage was able to account for perirhinal cortical activity during a semantic task.

Visual attention is known to enhance performance on challenging visual tasks. Applying the neuromodulatory effects of attention to the units in a CNN was shown to increase performance in these networks as well, more so when applied at later rather than earlier layers (Lindsay & Miller, 2018). This use of task-performing models has also led to better theories of how attention can work than those stemming solely from neural data (Thorat et al., 2019).

Finally, CNNs were also able to recapitulate several behavioral and neural effects of fine grain perceptual learning (Wenliang & Seitz, 2018).

**Adding Biological Details**

Some of the architectural variants described above—particularly the addition of local and feedback recurrence—are biologically inspired and assumed to enhance the ability of the network to perform challenging tasks. Other brain-inspired details have been explored by the machine learning community in the hopes that these additions would be useful for difficult tasks. This includes foveation and saccading (Mnih, Heess, & Graves, 2014) for image classification.

Another reason to add biological detail would be out of a belief that it may make the network “worse.” The fact that CNNs lack many biological details does not make them a poor model of the visual system necessarily; it simply makes them an abstract one. However, it should be an ultimate aim to bring abstract and detailed models together to show how the high-level computations of a CNN are implemented using the machinery available to the brain. To this end, work has been done on spiking CNNs (Cao, Chen, & Khosla, 2015) and CNNs with stochastic noise added (Mcnintosh et al., 2016). These attempts can also identify aspects of these biological details that are useful for computation.

The long history of modeling the circuitry of visual cortex can provide more ideas about what details to incorporate and how. A preexisting circuit model of V1 anatomy and function (Rubin, Van Hooser, & Miller, 2015), for example, was placed into the architecture of a CNN and used to replicate effects of visual attention (Lindsay, Rubin, & Miller, 2019). A more extreme approach to adding biological detail can be found in the study of Tschopp, Reiser, and Turaga (2018), where the connectome of the fly visual system defined the architecture of the model, which was then trained to perform visual tasks. Beyond this, having access to an “image computable” model of the visual system opens the door to many more modeling opportunities that can explore more than just core vision (Figure 5).
LIMITATIONS AND FUTURE DIRECTIONS

As with any model, the current limitations and flaws of CNNs should point the way to future research directions that will bring these models more in line with biology. The basic structure of a CNN assumes weight sharing. That is, a feature map is the result of the exact same filter weights applied at each location in the layer below. Although selectivity to visual features like orientation can appear all over the retinotopic map, it is clear that this is not the result of any sort of explicit weight sharing. Either genetic programming ensures the same features are detected throughout space, or this property is learned through exposure. Studies on “translation tolerance” have shown the latter may be true (Dill & Fahle, 1998). Weight sharing makes CNNs easier to train; however, ideally the same results could be found using a more biologically plausible way of fitting filters.

Furthermore, in most CNNs, Dale’s law is not respected. That is, the same neuron can be weighted by both inhibitory (negative) and excitatory (positive) weights. In the visual system, connections between areas tend to come only from excitatory cells. To be consistent with biology, a negative feedforward weight could be interpreted as an excitatory feedforward connection that acts on local inhibitory neurons. But this relationship between excitatory feedforward connections and the need for local inhibitory recurrence points to a complication of adding biological details to these networks: Some of these biological details may only function well or make sense in light of others and thus need to be added together.

To some, the way in which these networks are trained also poses a problem. The backpropagation algorithm is not considered biologically plausible enough to be an approximation of how the visual system actually learns. However, most methods for model fitting in computational neuroscience do not intend to mimic biological learning, and backpropagation could be thought of as just another parameter fitting technique. That being said, several researchers are investigating means by which the brain could perform something like backpropagation (Whittington & Bogacz, 2019; Bartunov et al., 2018; Sacramento, Costa, Bengio, & Senn, 2018; Roelfsema, van Ooyen, & Watanabe, 2010). Comparing models trained using more biologically plausible techniques to standard supervised learning (as well as to the unsupervised and reinforcement learning approaches discussed above) could offer insights as to the role of learning in determining representations.

The vast majority of studies comparing CNNs to biological vision have used data from humans or nonhuman primates. Where attempts have been made to compare CNNs to one of the most commonly used animal groups in neuroscience research—rodents—results are not nearly as strong as they are for primates (Cadena, Sinz, et al., 2019; de Vries et al., 2019; Matteucci, Marotti, Riggi, Rosselli, & Zoccolan, 2019). Understanding what can turn CNNs into a good model of rodent vision would go a long way in understanding the difference between primate and rodent vision and would open rodent vision up to the exploration tactics described here. CNNs have also been compared with the behavioral patterns of pigeons on a classification task (Lowe, Guest, Slomka, Navarro, & Wasserman, 2017).

Even in the context of primate vision, simple object or scene classification tasks only represent a small fraction of what visual systems are capable of and used for naturally. More ethologically relevant and embodied tasks such as navigation, object manipulation and visual reasoning, may be needed to capture the full diversity of visual processing and its relation to other brain areas. Early versions of this idea are already being explored (Cichy, Kriegeskorte, Jozwik, van den Bosch, & Charest, 2019; Dwivedi & Roig, 2018). The study of insect vision has historically taken this more holistic approach and may make for useful inspiration (Turner, Giraldo, Schwartz, & Rieke, 2019).

Conclusions

The story of CNNs started with a study on the tuning properties of individual neurons in primary visual cortex. Yet, one of the impacts of using CNNs to study the visual system has been to push the field away from focusing on interpretable responses of single neurons and toward population-level descriptions of how visual information is represented and transformed to perform visual tasks. The shift toward models that actually “do” something has forced a reshaping of the questions around the study of the visual system. Neuroscientists are adapting to this new style of explanation and the different expectations that come with it (Hasson, Nastase, & Goldstein, 2019).

Importantly, these models have also made it possible to reach some of the preexisting goals in the study of...
vision. In 2007 (Pinto, Cox, & DiCarlo, 2008), for example, a perspective piece on the study of object recognition claimed that “Progress in understanding the brain’s solution to object recognition requires the construction of artificial recognition systems that ultimately aim to emulate our own visual abilities, often with biological inspiration” and that “instantiation of a working recognition system represents a particularly effective method of success in understanding object recognition.” In this way, CNNs as a model of the visual system are a success.

Of course, nothing can be learned about biological vision through CNNs in isolation, but rather only through iteration. The insights gained from experimenting with CNNs should shape future experiments in the lab, which, in turn, should inform the next generation of models.

Acknowledgments

We thank SciDraw.io for providing brain and neuron drawings. Feature visualizations in Figure 4 taken from Olah et al. (2017) are licensed under Creative Commons Attribution 4.0 international license, and the insights gained from experimenting with CNNs should shape future experiments in the lab, which, in turn, should inform the next generation of models.

REFERENCES


How humans compare, compared to machines.


