Journal Club

Editor's Note: These short reviews of recent *JNeurosci* articles, written exclusively by students or postdoctoral fellows, summarize the important findings of the paper and provide additional insight and commentary. If the authors of the highlighted article have written a response to the Journal Club, the response can be found by viewing the Journal Club at www.jneurosci.org. For more information on the format, review process, and purpose of Journal Club articles, please see http://www.jneurosci.org/content/jneurosci-journal-club.

Do Primates and Deep Artificial Neural Networks Perform Object Categorization in a Similar Manner?

Prabaha Gangopadhyay[1] and Jhilik Das[2]

[1]Master of Science Program, Undergraduate Department and Centre for Neuroscience, and [2]Graduate Program in Neuroscience, Centre for Neuroscience, Indian Institute of Science, Bangalore 560012, India

Review of Rajalingham et al.

Building models is an important part of understanding neural systems. Models not only provide us with a plausible estimate of the computations underlying a system's functioning, but also help us predict the system's behavior when it is presented with novel inputs. In the quest to model object recognition and classification by the ventral visual pathway in the primate brain, deep convolutional neural networks (DNNs) have become leading tools, because DNNs roughly equal humans at the task of classifying objects based on images (He et al., 2015). Briefly, DNNs take an input image and pass it through multiple layers of computation to extract discriminative features from images. These features are then fed into a classifier that assigns a label to the input image. A typical layer takes the output of the previous layer as its input, passes it through several independent, linear convolutional filters, and feeds its output to the next layer after overlaying a nonlinearity. For a particular image, the network assigns probabilities to each object class and picks out the object class with the highest probability. While training a network, the weights of the filters are altered using variants of a method called "gradient descent" (Rumelhart et al., 1986; Bottou, 2010). Networks are often trained to classify images into one of the 1000 categories in a standard dataset from ImageNet (Deng et al., 2009). But testing 1000-class classification is impractical in biological experiments and hence, to compare DNNs' performance with primates', the classification layer is often retrained to do *N*-class classification, *N* being the number of object classes in the primate experiment.

The exact mechanism underlying the functioning of DNNs is still unclear. Neuroscientists' interest in studying emergent properties of another poorly understood complex network, the brain, led to their interest in understanding DNNs, which produced a wave of work comparing the two systems. Some of these studies suggested that DNNs and primate brains represent images in similar ways. For example, Khaligh-Razavi and Kriegeskorte (2014) showed that the pattern of activity elicited in response to natural objects in the population of artificial neurons in the final layer of DNNs is similar to that in the primate inferior temporal (IT) cortex, one of the higher visual areas that encodes object identities. It has also been shown that brain activation patterns in higher visual areas, discerned through fMRI (Güçlü and van Gerven, 2015) and electrode arrays (Yamins et al., 2014; Yamins and DiCarlo, 2016), can be predicted by the activation patterns in the later layers of a DNN. Other studies, however, suggest differences between DNNs and the primate visual system. For example, Pramod and Arun (2016) showed that two images considered very dissimilar by humans were often not considered so by DNNs. Additionally, Katti et al., (2017) showed DNNs to be systematically worse than humans at real-world object detection. Although neural networks match some activation properties of human visual areas, Grill-Spector et al., (2018) showed that their receptive fields are very different. These studies indicate that though DNNs and primates often show similar overall performance, their internal information processing mechanisms might be different. This difference should be reflected as a difference in the pattern of mistakes made by the two systems. For example, even if both primates and DNNs classify 80% of "horse" images as horse, do both the systems get the same images wrong?

A recent study by Rajalingham et al., (2018) published in *The Journal of Neuroscience* addresses this question by investigating whether similar image representations un-

Received Sept. 24, 2018; revised Dec. 20, 2018; accepted Dec. 21, 2018.

This work was supported by the KVPY fellowship program from the Department of Science and Technology (Government of India) to P.G., and by a Junior Research Fellowship from the University Grants Commission to J.D. We thank Dr. SP Arun, members of Vision Lab, IISc (Pramod, Harish, Thomas, Aakash, Zhivago, Georgin, and Jyothsna), and Hansika Chhabra for their valuable inputs and comments.

The authors declare no competing financial interests.

Correspondence should be addressed to Prabaha Gangopadhyay at prabahag@iisc.ac.in.

https://doi.org/10.1523/JNEUROSCI.2458-18.2018

 0270-6474/19/390946-03$15.00/0

derlie object classification in DNNs and primates. To compare the performance of humans, monkeys (*Rhesus macaque*), and DNNs, the authors compared the image discriminability patterns of the three systems. To test object recognition independent of context, an image dataset was generated by overlaying 3D models of 24 different object classes onto randomly chosen natural backgrounds. Humans and monkeys were asked to do a binary discrimination task where they classified a test image by clicking on a choice screen showing canonical views of two objects; the target (same object class as the test image) and a distractor (a different object class). DNNs were retrained to classify the same images as classified by primates, by determining the probability that the test image belonged to each of the 24 possible object classes. The probabilities assigned by DNNs to the target and distractor classes were compared and the one assigned the higher probability was taken as the DNN's choice in the binary discrimination task.

A discrimination measure called d', which quantifies the separability of two groups (e.g., dog vs not-dog) by comparing "hits" (e.g., dog being called a dog) and "false alarms" (e.g., not-dog being called a dog), was used for all the systems, at four different resolutions. Two object-level discriminations were used: (1) *one object versus rest* (e.g., dog vs not-dog; 24 values), and (2) *between each pair of objects* (e.g., dog vs camel, dog vs bear etc.; ${}^{24}C_2$ values). However, these object-level discrimination measures are computed across all images of an object, taking away the effect of individual images. To capture the latter, two image-level discriminations were also used (where 10 random images were selected from each object class, for a total of 240 images): (1) *one image of an object versus rest of the objects* (i.e., dog vs not-dog for each dog image; 10×24 values), (2) *between each image-object pair* (i.e., dog vs camel, dog vs bear etc. for each dog image; 240×23 values). It should be noted that since discriminability at the object-level is obtained by averaging across images, to capture the variability purely due to individual images, the object-level mean was subtracted from the image-level measures.

To compare the performance of different systems with humans, a metric called "*human-consistency*" was used. It is the correlation between discriminability pattern observed in the system and that observed in humans, normalized by the mean intersubject consistency (the split-half reliability) of the system and humans. Thus, by definition, humans have *human-consistency* of 1. Because a split-half measure requires multiple subjects, for DNNs, "subjects" were generated by varying training-image presentation order and initial conditions of filter weights, keeping the model architecture and training-image set fixed. An artificial network was called similar to primates if its *human-consistency* fell within the "*primate zone*": the range of *human-consistency* of monkeys, to *human-consistency* of humans (the latter being 1).

The *human-consistency* at the object-level resolution for all DNN models tested were found to fall in the *primate-zone*, but when inspected at the image-level, none of the DNNs reached the *primate-zone* (Rajalingham et al., 2018, their Figs. 2, 3). Human data were available for many subjects (1476) but very few monkey/DNN subjects were available. To make the comparison fair, the *human-consistency* at image-level was extrapolated for an infinite number of subjects for monkeys and DNNs, but still DNNs failed to reach the lower limit of the *primate-zone* (Rajalingham et al., 2018, their Fig. 4). Thus, at the coarse resolution of object classes, DNNs and primates show similar discriminability patterns, but the patterns differ at the finer resolution of individual images; DNNs and primates misclassify a different set of images. It should be noted that the image-level discriminability patterns of different DNN architectures were highly correlated (Rajalingham et al., 2018, their Fig. 6) implying similarity in internal processing across them. Thus, none of the architectures used the classification mechanism of primates. It was also shown that introducing simple changes even in the most human-consistent model, like making the input more similar to retinal sampling or changing the classifier or the test-image set, did not make DNNs similar to primates. Nonetheless, analysis based on image attributes revealed that primate and artificial models use somewhat similar attributes (e.g., contrast, size etc.) for object categorization (Rajalingham et al., 2018, their Fig. 7); i.e., their respective performances change in a similar manner with change in image attributes.

These results can be summarized as follows: (1) object-level discrimination behavior of primates and DNN models are statistically indistinguishable, (2) image-based object discrimination by DNN models does not match that of primates, and (3) simple modification of DNN models do not make them classify objects in the same manner as primates, indicating some intrinsic difference in their processing.

The study is a welcome addition to the field of vision science because it provides an essential test of the validity of using DNNs to understand primate visual representation and processing. DNNs, independent of the biology they are often linked with, are an astounding feat in the field of computer vision research, but that does not necessitate their being a good representation of the primate visual cortex: indeed, the study by Rajalingham et al., (2018) suggests that they are not.

Though the important question of whether DNNs and primates differ in image classification has been addressed, what the exact differences are, remains unknown. The authors show that the two systems use similar image attributes for classification, but still there are images that are correctly classified by primates but not by DNNs. Future studies should determine which attributes of the image are most important for classification by primates in such cases. Would increasing sensitivity toward these features improve DNN performance? Also, because background and context have been reported to play an important role in object detection, behaviorally (Davenport and Potter, 2004; Munneke et al., 2013), neutrally (Brandman and Peelen, 2017), and computationally (Katti et al., 2017); it would be interesting to investigate whether discriminability patterns of DNNs become more similar to primates when tested on datasets with congruent objects and backgrounds.

Another important question the study raises is whether models assuming simple, time-invariant neural activations, like DNNs, sufficiently capture the dynamics in the ventral visual pathway. It is known that neurons in IT display encoding properties essential to object recognition, like view invariance (Ratan Murty and Arun, 2015), and local shape information (Sripati and Olson, 2009), at later stages of firing, which is not evident in the mean response. Including recurrent connections in a layer might incorporate such temporal properties. Independently, multiple studies have reported complex nonlinearities in the receptive fields of visual neurons (Schwartz et al., 2012; Touryan and Mazer, 2015; Gharat and Baker, 2017), which might not be captured by the single ReLU-type (Nair and Hinton, 2010) nonlinearity used in DNNs. Introducing these experimentally observed properties in DNNs will hopefully help us make a comprehensive, concise, yet accurate model of both object feature process-

ing and classification in the ventral visual pathway.

References

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp 177–186. Heidelberg: Physica-Verlag.

Brandman T, Peelen MV (2017) Interaction between scene and object processing revealed by human fMRI and MEG decoding. J Neurosci 37:7700–7710. CrossRef Medline

Davenport JL, Potter MC (2004) Scene consistency in object and background perception. Psychol Sci 15:559–564. CrossRef Medline

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255. Piscataway, NJ: IEEE Computer Society.

Gharat A, Baker CL Jr (2017) Nonlinear Y-like receptive fields in the early visual cortex: an intermediate stage for building cue-invariant receptive fields from subcortical Y cells. J Neurosci 37:998–1013. CrossRef Medline

Grill-Spector K, Weiner KS, Gomez J, Stigliani A, Natu VS (2018) The functional neuroanatomy of face perception: from brain measurements to deep neural networks. Interface Focus 8:20180013. CrossRef Medline

Güçlü U, van Gerven MA (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci 35:10005–10014. CrossRef Medline

He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, v2016-January (2016), pp 770–778. Los Alamitos, CA: IEEE Computer Society.

Katti H, Peelen MV, Arun SP (2017) How do targets, nontargets, and scene context influence real-world object detection? Atten Percept Psychophys 79:2021–2036. CrossRef Medline

Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Comput Biol 10:e1003915. CrossRef Medline

Munneke J, Brentari V, Peelen MV (2013) The influence of scene context on object recognition is independent of attentional focus. Front Psychol 4:552. CrossRef Medline

Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: ICML'10 Proceedings of the 27th International Conference on International Conference on Machine Learning, pp 807–814. Haifa, Israel: Omnipress.

Pramod RT, Arun SP (2016) Do computational models differ systematically from human object perception? 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1601–1609. Los Vegas, NV: IEEE Computer Society.

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ (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. J Neurosci 38:7255–7269. CrossRef Medline

Ratan Murty NA, Arun SP (2015) Dynamics of 3D view invariance in monkey inferotemporal cortex. J Neurophysiol 113:2180–2194. CrossRef Medline

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. CrossRef

Schwartz GW, Okawa H, Dunn FA, Morgan JL, Kerschensteiner D, Wong RO, Rieke F (2012) The spatial structure of a nonlinear receptive field. Nat Neurosci 15:1572–1580. CrossRef Medline

Sripati AP, Olson CR (2009) Representing the forest before the trees: a global advantage effect in monkey inferotemporal cortex. J Neurosci 29:7788–7796. CrossRef Medline

Touryan J, Mazer JA (2015) Linear and nonlinear properties of feature selectivity in V4 neurons. Front Syst Neurosci 9:82. CrossRef Medline

Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19, 356–365. CrossRef Medline

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