Deep Networks as Models of Human and Animal Categorization

Bradley C. Love (b.love@ucl.ac.uk)
Olivia Guest (o.guest@ucl.ac.uk)
Department of Experimental Psychology, University College London
London, UK

Piotr Slomka (piotr.slomka@cshs.org)
Department of Imaging Cedars-Sinai Medical Center
David Geffen School of Medicine, University of California Los Angeles
Los Angeles, California, USA

Victor M. Navarro (victor-navarro@uiowa.edu)
Edward Wasserman (ed-wasserman@uiowa.edu)
Department of Psychological & Brain Sciences, The University of Iowa
Iowa City, Iowa, USA

Keywords: Categorization; Neural Network; Deep Learning; Convolutional Neural Network; Animal Learning; Object Recognition; Neural Computation.

Introduction

Convolutional neural networks (CNNs) trained as classifiers learn by associating visual inputs (e.g., photographs of objects) with appropriate output labels (e.g., “crow”, “dog”, “car”). These complex models, which contain millions of weights, are the state-of-the art in machine vision, rivaling humans in object recognition tasks (LeCun, Bengio, & Hinton, 2015; Krizhevsky, Sutskever, & Hinton, 2012). What these networks learn displays some commonalities with human learning (Kubilius, Bracci, & de Beeck, 2016; Lake, Zaremba, Fergus, & Gureckis, 2015). Furthermore, the layers in these networks have been related to neural activity along the ventral stream (Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016)

The similarity spaces created by these models at various network layers allow us to draw parallels with the brain’s neural coding schemes (Guest & Love, 2017). At earlier layers, networks display similarity spaces that reflect the high-level categories found in the input space, e.g., lions and tigers are more similar to one another than to mopeds. At the more advanced layers, similarity structure tends to break down such that representations of different object categories become orthogonal.

Can these networks also shed light on how non-human animals categorize? CNNs can be used to determine at what level of representation (i.e., what network layer) animals are coding similarities between images. For example, are animals learning regularities at a very low level, close to the pixels in the image, or are they seizing upon more abstract shape features? In this contribution, we address this question by examining data from pigeons trained to categorize images of cardiograms as normal or abnormal.

Pigeons are excellent at classifying visual stimuli (Bhatt, Wasserman, Reynolds, & Knauss, 1988). For example, pigeons trained to discriminate between medical images of normal and cancerous breast tissue generalized to novel stimuli and attained human-level accuracy (Levenson, Krupinski, Navarro, & Wasserman, 2015). Importantly, knowledge transfer was only true in certain circumstances. Pigeons only generalized within image magnification levels — they were not scale-invariant. Also, generalization was significantly compromised, although still above chance, when tested on grayscale images (perhaps to be expected given the loss of hue and brightness cues). However, the pigeons’ performance improved with additional training on greyscale images.

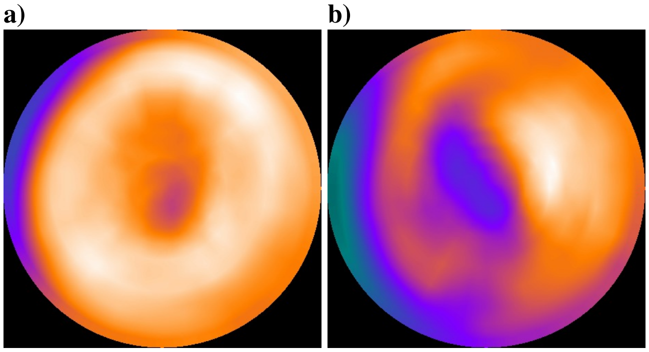

Figure 1: Two examples of the stimuli that the pigeons and network are asked to classify: **a**) a normal cardiogram without any perfusion damage; and **b**) an abnormal cardiogram with total perfusion damage 20 (of a maximum of 51).

Can CNNs explain such patterns of performance? At the most advanced layers of these networks, representations should be somewhat invariant to changes in size, luminance, translation, etc. However, at lower layers the network will be more sensitive to such changes and will not generalize as broadly. Which network layer best captures how pigeons categorize?

Here we consider data from an a yet unpublished study by Wasserman and colleagues in which pigeons are trained to classify cardiograms as normal or abnormal, see Figure 1. Pigeons can correctly determine whether a cardiogram is abnormal or normal in much the same way as a skilled human, and can correctly classify unseen cardiogram images.

To parallel the pigeons, we also show the same stimuli to a CNN, namely Inception-v3 GoogLeNet (Krizhevsky et al., 2012). In line with the pigeons, the network can also determine whether a stimulus is normal or abnormal. Also like the pigeons, Inception-v3 GoogLeNet is very sensitive to changes in color, having serious problems generalizing when trained on color images and tested on grayscale without additional training. Importantly, even though the model can differentiate between the two classes at the output layer it can also do so at much lower layers. The output layer is trained to represent very high-level conceptual categories (1000 mutually exclusive classes, e.g., sunglasses, moped, jellyfish, etc.). Although these output classes do not contain options for normal and abnormal cardiograms, the network provides a distributed answer across these categories thus solving the classification task. In other words, the output shows a similarity structure matching the normal/abnormal distinction in the inputs.

As mentioned, at lower layers including the input layer, the network can also differentiate the two types of stimuli into normal and abnormal. This means that basic stimulus properties, which are what the network and the pigeons are extracting and learning, are sufficient to separate the two classes of cardiograms shown in Figure 1. This is important because it implies that more complex and abstract features, or even representations of basic shapes, are not required for the type of learning problem the pigeons are solving. In addition, this predicts that generalization will be poor in both the animal and computational models we have considered. We consider the broader implications of these results for how humans and non-human animals categorize.

Six Relevant Papers by BCL

Gigure, G. & Love, B.C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 110 (19), 7613-7618.*

Guest, O., Love, B.C. (2017). What the Success of Brain Imaging Implies about the Neural Code. eLife,6:e21397.

Love, B.C. (2015). The Algorithmic Level is the Bridge Between Computation and Brain. *Topics in Cognitive Science, 7, 230-242.*

Mack, M.L., Love, B.C., & Preston, A.R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. Proceedings of the National Academy of Sciences (PNAS), 113(46), 1320313208.

Mack, M.L., Preston, A.R. & Love, B.C. (2013). Decoding the Brain's Algorithm for Categorization from its Neural Implementation. *Current Biology*, 23, 2023-2027.

Turner, B.M., Forstmann, B.U., Love, B.C., Palmeri, T.J. & Van Maanen, L. (2016). Approaches to Analysis in Model-based Cognitive Neuroscience. *Journal of Mathematical Psychology.* http://dx.doi.org/10.1016/j.jmp.2016.01.001.

Acknowledgments

This work was supported by NIH (Grant 1P01HD080679) to BCL and EW. We thank all the members of the P01 team, led by Vladimir Sloutsky. Additional support from this work comes from the Leverhulme Trust (Grant RPG-2014-075) and a Wellcome Trust Investigator Award (Grant WT106931MA) to BCL, as well as The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Bhatt, R., Wasserman, E., Reynolds, W., & Knauss, K. (1988). Conceptual behavior in pigeons: Categorization of both familiar and novel examples from four classes of natural and artificial stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(3), 219.

Guest, O., & Love, B. C. (2017, jan). What the success of brain imaging implies about the neural code. *eLife*, *6*, e21397. doi: 10.7554/eLife.21397

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput Biol*, *10*(11), 1–29. doi: 10.1371/journal.pcbi.1003915

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput Biol*, *12*(4), e1004896.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Proceedings of the annual meeting of the cognitive science society* (p. 1-6).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi: 10.1038/nature14539

Levenson, R. M., Krupinski, E. A., Navarro, V. M., & Wasserman, E. A. (2015, 11). Pigeons (columba livia) as trainable observers of pathology and radiology breast cancer images. *PLOS ONE*, *10*(11), 1-21. doi: 10.1371/journal.pone.0141357

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365. doi: 10.1038/nn.4244