Target article: "Deep Problems with Neural Network Models of Human Vision"

Why psychologists should embrace rather than abandon DNNs.

Galit Yovel[1,2] & Naphtali Abudarham[1]
[1]School of Psychological Sciences
[2]Sagol School of Neuroscience
Tel Aviv University, Tel Aviv, Israel, 69987

Corresponding author:
Galit Yovel
School of Psychological Sciences
Sagol School of Neuroscience
Tel Aviv University, Tel Aviv, Israel, 69987
gality@tauex.tau.ac.il

Word count:
ABSTRACT: 59
MAIN TEXT: 936
REFERENCES: 417
ENTIRE TEXT (TOTAL + ADDRESSES etc.): 1759

Target article: "Deep Problems with Neural Network Models of Human Vision"

Abstract

Deep neural networks (DNNs) are powerful computational models, which generate complex, high-level representations that were missing in previous models of human cognition. By studying these high-level representations, psychologists can now gain new insights into the nature and origin of human high-level vision, which was not possible with traditional handcrafted models. Abandoning DNNs would be a huge oversight for Psychological Sciences.

Target article: "Deep Problems with Neural Network Models of Human Vision"

Computational modeling has long been used by psychologists to test hypotheses about human cognition and behavior. Prior to the recent rise of deep neural networks (DNNs), most computational models were handcrafted by scientists who determined their parameters and features. In vision sciences, these models were used to test hypotheses about the mechanisms that enable human object recognition. However, these handcrafted models used simple, engineered-designed features (e.g., Gabors), which produced low-level representations that did not account for human-level, view-invariant object recognition (Biederman & Kalocsai, 1997; Turk & Pentland, 1991). The main advantage of DNNs over these traditional models is not only that they reach human-level performance in object recognition, but that they do so through hierarchical processing of the visual input that generates high-level, view-invariant visual features. These high-level features are the *"missing link"* between the low-level and output representations of the handcrafted models of object recognition. They therefore offer psychologists an unprecedented opportunity to test hypotheses about the origin and nature of these high-level representations, which were not available for exploration so far.

In this issue of BBS, Bowers and colleagues propose that psychologists should abandon DNNs as models of human vision, because they do not produce some of the perceptual effects that are found in humans. However, many of the listed perceptual effects that DNNs fail to produce are also not produced by the traditional handcrafted computational vision models, which have been prevalently used to model human vision. Furthermore, although current DNNs are primarily developed for engineering purposes (i.e., best performance), there are myriad of ways in which they can and should be modified to better resemble the human mind. For example, current DNNs that are often used to model human face and object recognition (Khaligh-Razavi et al., 2016; O'Toole & Castillo, 2021; Yamins & DiCarlo, 2016) are trained on static images (Cao et al., 2018; Jia Deng et al., 2009), whereas human face and object recognition are performed on continuous streaming of dynamic, multi-modal information. One way that was recently suggested to close this gap is to train DNNs on movies that are generated by head-mounted cameras attached to infants' forehead (Fausey et al., 2016), to better model the development of human visual system (Smith & Slone, 2017). Training DNNs initially on blurred images also provided insights about the potential advantage of the initial low acuity of infants' vision (Vogelsang et al., 2018). Such and many other modifications (e.g., Parisi et al., 2019) in the way DNNs are built and trained may generate perceptual effects that are more human-like. Yet, even current DNNs can advance our understanding of the mechanisms that enable the generation of the complex high-level representations that are required for face and object recognition (Abudarham et al., 2021; Hill et al., 2019) but are still undefined in current neural and cognitive models. This significant computational achievement should not be dismissed.

Bowers and colleagues further claim that DNNs should be used to test hypotheses rather than to solely make predictions. We fully agree and further propose that psychologists are best suited to apply this approach by utilizing the same procedures they have used for decades to test hypotheses about the hidden representations of the human mind. Since the early days of psychological sciences, psychologists have developed a range of elegant experimental and stimulus manipulations to study human vision. The same procedures can now be used to explore the nature of DNNs' high-level hidden representations as potential models of the human mind (Ma & Peters, 2020). For example, the *face inversion effect is* a robust, extensively studied, and well-established effect in human vision, which refers to the disproportionally large drop in performance that humans show for upside-down compared to upright faces (Cashon & Holt, 2015; Farah et al., 1995; Yin, 1969). Because the low-level features extracted by traditional, handcrafted algorithms are similar for upright and inverted faces, these traditional models do not reproduce this effect. Interestingly, a human-like face inversion effect that is larger than an object inversion effect is found in DNNs (Dobs et al., 2022; Jacob et al., 2021; Tian et al., 2022; Yovel et al., 2022). Thus, we can now use the same stimulus and task manipulations that were used to study this effect in numerous human studies, to test hypotheses about the mechanism that may underlie this perceptual effect. Moreover, by manipulating DNNs' training diet, we can examine what type of experience is needed to generate this human-like perceptual effect, which is impossible to test in humans where we have no control over their perceptual experience. Such an approach was recently used to address a long-lasting debate in cognitive sciences about the domain-specific vs. the expertise hypothesis in face recognition (Kanwisher et al., 2023; Yovel et al., 2022).

It was psychologists, not engineers, who first designed these neural networks to model human intelligence (McClelland et al., 1995; Rosenblatt, 1958; Rumelhart et al., 1986). It took more than 60 years since the psychologist, Frank Rosenblatt published his report about the *Perceptron*, for technology to reach its present state where these hierarchically structured algorithms can be used to study the complexity of human vision. Abandoning DNNs would be a huge oversight for cognitive scientists, who can contribute considerably to the development of more human-like DNNs. It is therefore pertinent that psychologists join the AI research community and study these models in collaboration with engineers and computer scientists. This is a unique time in the history of cognitive sciences, where scientists from these different disciplines have shared interests in the same type of computational models that can advance our understanding of human cognition. This opportunity should not be missed by psychological sciences.

COMPETING INTERESTS STATEMENT: We have no competing interests.
FUNDING STATEMENT: This paper is funded by an ISF 971/21 and Joint NSFC-ISF 2383/18 to GY.

References

Abudarham, N., Grosbard, I., & Yovel, G. (2021). Face Recognition Depends on Specialized Mechanisms Tuned to View-Invariant Facial Features: Insights from Deep Neural Networks Optimized for Face or Object Recognition. *Cognitive Science*, *45*(9). https://doi.org/10.1111/cogs.13031

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1358), 1203–1219. https://doi.org/10.1098/rstb.1997.0103

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 67–74. https://doi.org/10.1109/FG.2018.00020

Cashon, C. H., & Holt, N. A. (2015). Developmental Origins of the Face Inversion Effect. In *Advances in Child Development and Behavior* (1st ed., Vol. 48). Elsevier Inc. https://doi.org/10.1016/bs.acdb.2014.11.008

Dobs, K., Martinez, J., Yuhan, K., & Kanwisher, N. (2022). Using deep convolutional neural networks to test why human face recognition works the way it does. *BioRxiv*, 1–26.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What Causes the Face Inversion Effect? *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 628–634. https://doi.org/10.1037/0096-1523.21.3.628

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107. https://doi.org/10.1016/j.cognition.2016.03.005

Hill, M. Q., Parde, C. J., Castillo, C. D., Colón, Y. I., Ranjan, R., Chen, J.-C., Blanz, V., & O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, *1*(11), 522–529. https://doi.org/10.1038/s42256-019-0111-7

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*(1), 1–14. https://doi.org/10.1038/s41467-021-22078-3

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/cvprw.2009.5206848

Kanwisher, N., Gupta, P., & Dobs, K. (2023). CNNs Reveal the Computational Implausibility of the Expertise Hypothesis. *IScience*, *26*(2), 105976. https://doi.org/10.1016/j.isci.2023.105976

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2016). Mixing deep neural network features to explain brain representations. *Journal of Vision*. https://doi.org/10.1167/16.12.369

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: towards using deep nets as models for human behavior*. 1–39.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

O'Toole, A. J., & Castillo, C. D. (2021). Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning. *Annual Review of Vision Science*, *7*, 543–570. https://doi.org/10.1146/annurev-vision-093019-111701

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ArXiv Preprint*.

Rosenblatt, F. (1958). The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, *65*(6), 386–408.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology*, *8*(DEC), 1–10. https://doi.org/10.3389/fpsyg.2017.02124

Tian, F., Xie, H., Song, Y., Hu, S., & Liu, J. (2022). The Face Inversion Effect in Deep Convolutional Neural Networks. *Frontiers in Computational Neuroscience*, *16*(May), 1–8. https://doi.org/10.3389/fncom.2022.854218

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(44), 11333–11338. https://doi.org/10.1073/pnas.1800901115

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. In *Nature Neuroscience*. https://doi.org/10.1038/nn.4244

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141.

Yovel, G., Grosbard, I., & Abudarham, N. (2022). Computational models of perceptual expertise reveal a domain-specific inversion effect for objects of expertise. *PsyXiv*.