Robustness to object rotation in humans and deep neural networks

Haider Al-Tahan (haltaha@uwo.ca)
Graduate Program in Neuroscience, Schulich School of Medicine & Dentistry
Western University

Farzad Shayanfar (farzad.shayanfar@hotmail.com)

Ehsan Tousi (ekahooka@uwo.ca)
Graduate Program in Neuroscience, Schulich School of Medicine & Dentistry
Western University

Marieke Mur (mmur@uwo.ca)
Department of Psychology, Department of Computer Science
Western University

Abstract

Invariant object recognition, a cornerstone of human vision, enables recognizing objects despite variations in rotations, positions, and scales. To emulate human-like generalization across object transformations, computational models must perform well in this aspect. Deep neural networks (DNNs) are popular computational models for human ventral visual stream processing, though their alignment with human performance on visual tasks remains debated. We examine robustness to object rotation in human adults and pretrained feedforward DNNs. We find that object recognition performance is better preserved in humans than in DNNs, although they show a similar pattern of how performance drops as a function of rotational angle. Furthermore, humans and models make different errors, which suggests different processing strategies. Finally, model architecture minimally influences DNN performance, while DNNs trained on richer visual diets and semi-supervised learning goals excel. Our study suggests that visual diet and learning goals may play an important role in the development of invariant object recognition in humans.

Keywords: Deep neural networks; Invariant Object Recognition; Ventral Visual Pathway

Introduction

The ventral visual pathway exhibits a robust ability to identify objects irrespective of their orientation, position, and size (Biederman, 1987; DiCarlo & Cox, 2007; Eger, Ashburner, Haynes, Dolan, & Rees, 2008; Freiwald & Tsao, 2010). This ability has been referred to as invariant object recognition. Any comprehensive model aiming to simulate human object recognition should emulate this capability, demonstrating proficiency in recognizing objects even when they undergo various real-world transformations (DiCarlo, Zoccolan, & Rust, 2012; Peters & Kriegeskorte, 2021; Bowers et al., 2022). Recent work in computer vision has begun to examine robustness of deep neural networks (DNNs) to variations in object orientation, particularly through the use of 3D graphics renderers (Madan et al., 2022; Alcorn et al., 2019; Abbas & Deny, 2022). However, studies comparing performance of these networks to human capabilities are lacking. This leaves open the following questions: Are DNNs as invariant to variations in object orientation as humans are? Which design features make them most invariant?

We addressed these questions by developing a stimulus set that systematically varies object orientation, and that enabled direct comparison between humans and DNNs on object recognition performance. Furthermore, we tested a diverse range of DNNs with different design features to explore what makes some DNNs display more invariant object recognition behaviour than others. Design features of interest included network architecture, learning objective, and visual diet (Smith, Jayaraman, Clerkin, & Yu, 2018; Konkle & Alvarez, 2022; Goyal & Bengio, 2022). This approach provides an opportunity to estimate the unique contribution of each design feature to invariant object recognition.

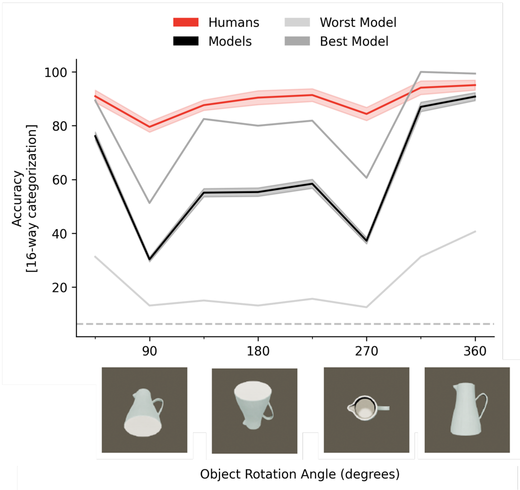

Figure 1: Summary of human and DNN model performance for images of objects that were rotated in depth. The dashed line at the bottom denotes chance performance. Shaded areas indicate standard error of the mean across humans or models.

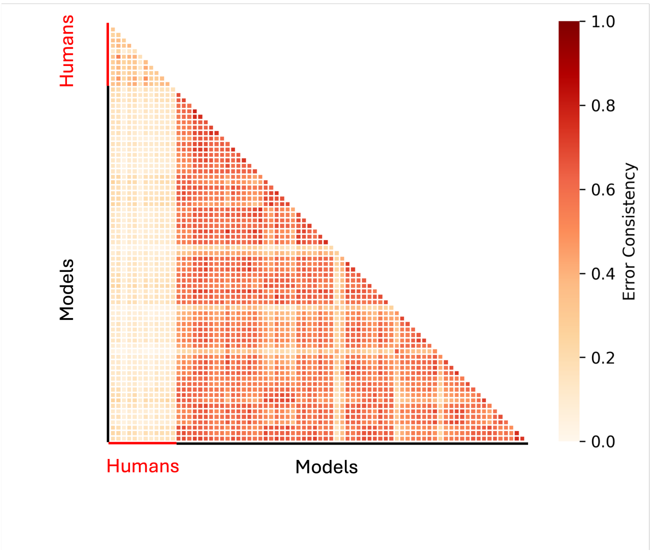

Figure 2: Error consistency for humans and DNN models when categorizing objects after in-depth rotation. Results reveal a gradient from low consistent errors (lighter colors) between human participants and DNN models, to moderately consistent errors among human participants, to highly consistent errors (red) among DNN models.

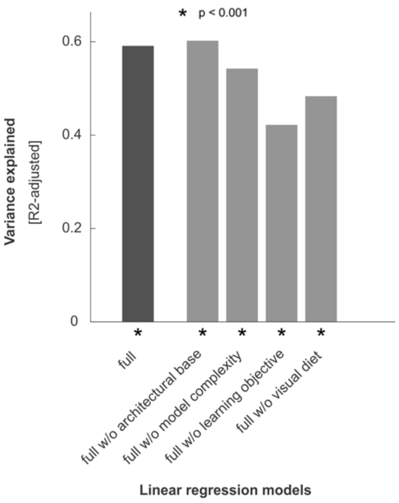

Figure 3: Explained variance in model performance achieved through linear regression, using model architecture (base and complexity), learning objective, and visual diet as predictor variables. Asterisks indicate regression models that explained significantly more variance than a regression model with a constant term only.

Methods

Human experiment

We recruited 17 healthy human adults for the study. Two participants were excluded due to low performance on practice blocks, leaving 15 participants for analysis (mean age: 25 years, five females). The experiment consisted of 13 blocks (128 trials per block). The first three blocks were practice blocks and the remaining were experimental blocks. The images for these blocks were generated using the ThreeDWorld platform (Gan et al., 2021), by applying object rotation in depth on our stimulus set of 176 sourced 3D objects from online repositories. Objects were sourced from 16 categories known to both humans and DNNs (Geirhos et al., 2020, 2021).

On each trial, a grey screen with a central white fixation cross was presented, followed by an image displayed for a duration of 200 ms. The image was followed by a coloured mask for 200 ms, which served to increase task difficulty and reduce effects of recurrent processing on performance. The latter enables a fairer comparison between human participants and DNNs, whose architectures only allow for feedforward information flow (Geirhos et al., 2020). The next phase involved presenting the participants with a response screen with multiple category buttons. This remained available for a maximum period of 1500 ms or until a selection was made by the participant, whichever was shorter. During the response phase, participants were tasked with selecting the category they deemed most congruent with the stimulus. The experiment took approximately 1.5 hours to complete.

Results & Discussion

Humans are more invariant than DNNs to object rotation in depth

In the context of object rotation in depth, our analysis showed that humans and DNNs demonstrate a parallel pattern of performance as the degree of rotation varies. However, humans consistently surpass model accuracy, as observed in Figure 1. Both humans and models face challenges when objects are rotated to their lateral sides, leading to a notable drop in performance. This challenge becomes even more pronounced when objects are observed from unconventional angles, such as the top or bottom.

Humans and DNNs make different errors

Our investigation not only highlights the discrepancy in error rates between humans and DNNs but also underscores the distinct nature of their error patterns (Figure 2). We employed an error consistency measure that accounts for the anticipated consistency between two observers based on their task performance (Geirhos et al., 2021). Our results further indicate that DNNs tend to exhibit relatively high error consistency among themselves, with the exception of some models such as the Masked AutoEncoders (MAE). One potential explanation for these findings is the visual diet experienced during learning. While DNNs often share a common training dataset, leading to a certain level of similarity in their learned representations, humans experience diverse and unique visual diets throughout their developmental trajectory. This individualistic exposure to visual stimuli can contribute to a higher degree of dissimilarity among humans than among DNNs.

Visual diet and learning objectives are important for developing invariant object recognition

In the context of assessing the factors influencing model performance, Figure 3 presents explained variance in model performance achieved through linear regression. We employed model architectural base, model complexity, learning objective, and visual diet as predictor variables to estimate their contributions to the observed variability in model performance. Architectural base reflects whether a DNN is a convolutional or a vision transformer network; model complexity reflects the size of the network; learning objective reflects whether the network was trained using a supervised, semi-supervised, or self-supervised learning objective; and visual diet reflects the number of images in the training set.

We found that architectural distinctions between models may not be the primary driver of the observed variation in their performance. In contrast, model complexity emerges as a moderate determinant, accounting for approximately 5% of the variability in model performance. Furthermore, learning objectives and visual diet surfaced as more influential factors, contributing to roughly 20% and 15% of the variability in model performance, respectively.

References

Abbas, A., & Deny, S. (2022). *Progress and limitations of deep networks to recognize objects in unusual poses.* arXiv. (arXiv:2207.08034 [cs])

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., & Nguyen, A. (2019). Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4840–4849). Long Beach, CA, USA: IEEE.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. doi: 10.1037/0033-295X.94.2.115

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2022). *Deep Problems with Neural Network Models of Human Vision.* PsyArXiv.

DiCarlo, J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341.

DiCarlo, J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434.

Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R. J., & Rees, G. (2008). fMRI activity patterns in human LOC carry information about object exemplars within category. *Journal of Cognitive Neuroscience*, *20*(2), 356–370.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., ... Yamins, D. L. K. (2021). *Threedworld: A platform for interactive multi-modal physical simulation.*

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). *Partial success in closing the gap between human and machine vision.* arXiv.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2020). *Generalisation in humans and deep neural networks.* arXiv. (arXiv:1808.08750 [cs, q-bio, stat])

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *478*(2266), 20210068. (Publisher: Royal Society)

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491. (Number: 1 Publisher: Nature Publishing Group)

Madan, S., Henry, T., Dozier, J., Ho, H., Bhandari, N., Sasaki, T., ... Boix, X. (2022). When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, *4*(2), 146–153. (Number: 2 Publisher: Nature Publishing Group)

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, *5*(9), 1127–1144.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, *22*(4), 325–336.