Current DNNs are Unable to Integrate Visual Information Across Object Discontinuities

Ben Lonnqvist (ben.lonnqvist@epfl.ch)
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Elsa Scialom (elsa.scialom@epfl.ch)
School of Life Sciences
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Zehra Merchant (zehra.merchant@epfl.ch)
School of Life Sciences
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Michael H. Herzog (michael.herzog@epfl.ch)
School of Life Sciences
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Martin Schrimpf (martin.schrimpf@epfl.ch)
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Abstract

Particular deep neural networks (DNNs) are the current best models of the primate visual ventral stream and the core object recognition behaviors it supports. Despite a rich history of studying visual function in the field of psychophysics however, DNNs have not been thoroughly evaluated via classical psychophysical experiments. To address this gap, we designed a 12-AFC object recognition experiment with object stimuli containing various degrees of contour discontinuities. Humans ($n = 50$ in-laboratory participants) perform well above chance even for images containing very few fragments, with performance scaling logarithmically with the number of fragmented elements, up to near-perfect performance. Leading DNN models on the other hand fail to recognize these fragmented objects, performing at chance throughout. Attempting to remedy this object recognition gap, we fit a linear decoder on model activations to fragmented stimuli, but even with additional supervised trials model representations were unable to support human-level fragmented object recognition performance. Despite this, models as well as humans performed better on directional segment stimuli compared to phosphene-like dot stimuli. Taken together, our results show a striking failure case of current models of the human visual system that is not trivial to rescue – suggesting a critical difference in how models and humans integrate visual information.

Keywords: Psychophysics; Object Recognition; Visual Grouping; Deep Neural Networks; Brain-Score

Introduction

As measured by alignment to neural and naturalistic behavioral data in primates, particular deep neural networks (DNNs) are currently considered the best models of the visual system (Yamins & DiCarlo, 2016; Schrimpf et al., 2020). While the best models currently explain around half the variance across a range of datasets, their similarity to the human visual system under targeted manipulations remains unclear. Specifically, state-of-the-art DNNs have not yet been thoroughly evaluated on their alignment to measurements from the field of psychophysics where experiments aim to uncover the edge conditions of visual processing (Bowers et al., 2023). We here present first results of testing leading DNN models of primate vision on their similarity to humans in grouping experiments inspired by classical psychophysics (Wagemans et al., 2012). We designed an object classification task in which object stimuli are shown with sparse, fragmented visual information — one subset with locally directional edge information (segments), and one without (phosphenes) — which we tested on 50 in-laboratory participants and a selection of leading DNN models of primate core object recognition.

Human Experimental Methodology

Experimental Setup We recruited 50 human participants to perform a 12-AFC fragmented object recognition task in the Laboratory of Psychophysics at EPFL. Participants signed a consent form and were compensated 25CHF/hour for their participation in the study. Participants were sat in a darkened chamber, and stimuli were displayed foveally, spanning 8 degrees of visual angle. Stimuli were presented for 200ms, followed by a 200ms $1/f$ noise mask.

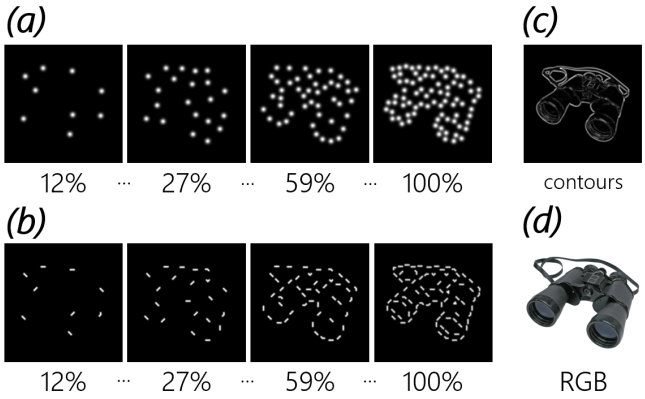

Figure 1: **Example stimuli used in the experiments.** We constructed deteriorated object stimuli, of which a selection is shown here. Percentages denote relative number of elements. ***(a)***: Phosphene (dot-like fragmented) stimuli. ***(b)***: Segment (edge-like fragmented) stimuli. ***(c)***: Contour (edge-filtered full) stimuli. ***(d)***: RGB (full object without background) stimuli.

Stimulus Manipulations Stimuli were created from background-removed images of common objects from the BOSS dataset (Brodeur, Guérard, & Bouras, 2014) using a fragment renderer (Rotermund, Scialom, Repnow, Herzog, & Ernst, 2024). Stimuli were split into two distinct groups: *phosphenes* (Figure 1a) and *segments* (Figure 1b). Of the 50 participants, a random half were presented phosphene stimuli, and the other half were presented segment stimuli. The dataset consisted of 9 different levels of numbers of elements, ranging logarithmically from 12% to 100% (maximum number of elements without overlap), resulting in a total of 432 stimuli (4 objects per each of the 12 categories, across 9 different numbers of elements). In each of these groups, participants were shown all stimuli in ascending order of numbers of elements. In addition, all participants were presented the same objects with contours, and in full RGB (Figure 1c,d).

Model Experimental Methodology

Zero-Shot Models To directly test DNNs' capability to perform the fragmented object recognition task in a human-comparable manner, we mapped a total of **14 models'** responses from ImageNet categories (Russakovsky et al., 2015) to our 12 object categories zero-shot using a WordNet synset mapping (Geirhos et al., 2021). We tested several models on the task, including AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), ResNet-50 (He, Zhang, Ren, & Sun, 2016), Regnet (Radosavovic, Kosaraju, Girshick, He, & Dollár, 2020), Effi-

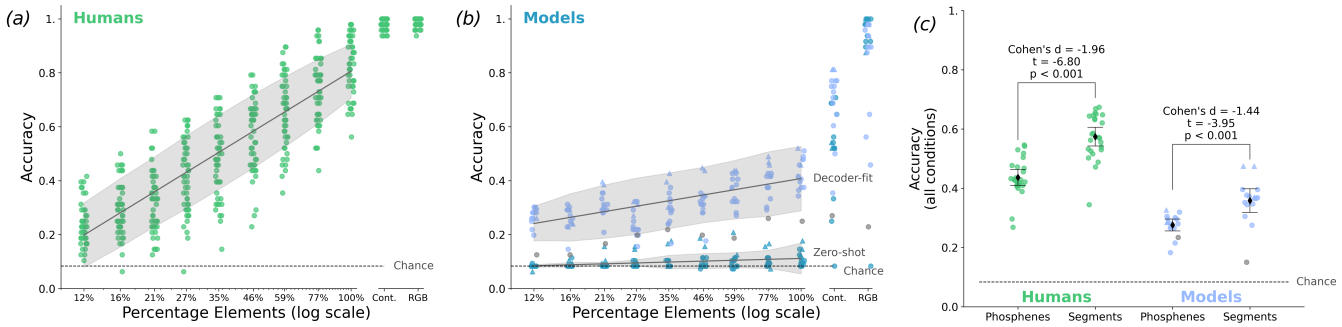

Figure 2: **Humans are able to recognize objects with discontinued visual information, while DNNs fail.** Individual dots represent individual **participants'** or **models'** average performance; error bars are 95% confidence intervals. Chance performance is 8.33%. ***(a)***: Human categorization performance. The x-axis shows different conditions (Figure 1), combined across phosphenes and segments. Human performance scales logarithmically (regression fit: $accuracy = 0.29 * \log(x) - 0.51$, $R^2 = 0.73$, $p < 0.001$). ***(b)***: Model categorization performances. **Dark blue**: Zero-shot models with responses mapped from ImageNet labels *without* additional fitting (regression fit: $accuracy = 0.01 * \log(x) + 0.05$, $R^2 = 0.11$, $p < 0.001$). **Light blue**: Linear decoder-fit models with an additional 120 supervised trials within-condition. Regression fit: $accuracy = 0.08 * \log(x) + 0.05$ ($R^2 = 0.46$, $p < 0.001$). **Gray**: Baseline `pixels` model. **▲Triangles** denote networks trained to exhibit a shape-bias, while **●circles** denote all other networks. ***(c)***: Human and model mean performance split by phosphene and segment stimuli.

cientNet (Tan & Le, 2019), SimCLR (Chen, Kornblith, Norouzi, & Hinton, 2020), two Barlow variants (Zbontar, Jing, Misra, LeCun, & Deny, 2021), 3 CorNets (Kubilius et al., 2019), and 3 shape-biased ResNets (Geirhos et al., 2018).

Decoder-fit Models For each of our 12 object categories, we first selected 10 ImageNet images from the corresponding ImageNet categories. We then removed backgrounds from these images (Gatis, 2023) and generated 120 novel fragmented images for all percentage levels (Figure 1; 10 images per object category). We fit linear decoders on the penultimate layer activations of a total of **16 models**, including all of the models that performed zero-shot categorization using ImageNet mapping as well as a `pixels` baseline model, and a ResNet-18 trained on ImageNet-21k (Ridnik, Ben-Baruch, Noy, & Zelnik-Manor, 2021).

Results

Human Performance Improves Logarithmically in the Number of Fragments **Human subjects'** performance was near-ceiling in both the RGB and contour conditions. Their performance scaled logarithmically with the number of added elements (Figure 2a), starting from an average of 24% correct at `12%` elements shown, up to 83% correct at `100%` elements shown.

Models Fail Catastrophically **Model performance** in the RGB condition was also high (average 90% correct), validating the effectiveness of the ImageNet mapping procedure; but substantially lower with contour stimuli (51% correct), and near-chance in all fragmented conditions (Figure 2b, **dark blue**). Model performance does not scale with added elements in the same way that human performance scales (two-sided t-test on the difference of regression slopes between **humans** and **models**, $t = 30.59$, $p < 0.001$).

Model Representations are Insufficient to Support Recognition Across Discontinuities To understand whether model representations are capable of supporting object recognition in fragmented stimuli at all, we fit **decoders** to fragmented stimuli. Figure 2b shows that model (**light blue**) performance substantially improved, but did not reach human levels – neither in absolute performance (average performance across conditions was 59% for **humans** vs 36% for **decoder-fit models**), nor in scaling ($t = 18.61$, $p < 0.001$). Even models explicitly trained to exhibit a **▲shape bias** (Geirhos et al., 2018) do not meaningfully outperform **●other models** (Figure 2b).

Both Models and Humans Prefer Directional Fragments We tested whether **humans** and **decoder-fit models** are better at recognizing segment stimuli (Figure 1b) over phosphene stimuli (Figure 1a). Indeed, both humans and models are substantially more accurate on segment stimuli (Figure 2c), with large effect sizes of $d = 1.96$ and $d = 1.44$ respectively.

Conclusion

We demonstrate a **striking DNN model failure** to integrate visual information across object discontinuities in a well-controlled fragmented object psychophysical experiment that **human participants** excelled on (Figure 2a,b). While **supervised decoder training** on additional stimuli enables models to perform above chance, model performance still lags far behind humans. Like humans however, decoder-fit models exhibit a preference to stimuli with directional edge information (Figure 2c). Taken together, our results suggest a clear difference in the way humans and models integrate visual information that is not trivial to rescue.

Acknowledgments

BL and MHH were supported by a grant from the Swiss National Science Foundation (N.320030_176153; "Basics of visual processing: from elements to figures"). ES was supported by ERA-NET NEURON (Ref Nr: NEURON-051).

References

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2023, January). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385. doi: 10.1017/S0140525X22002813

Brodeur, M. B., Guérard, K., & Bouras, M. (2014, September). Bank of Standardized Stimuli (BOSS) Phase II: 930 New Normative Photos. *PLOS ONE*, *9*(9), 1–10. (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0106953

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Gatis, D. (2023). *rembg: A tool to remove image background using pre-trained deep learning models.* `https://github.com/danielgatis/rembg`. GitHub. (Accessed: 2023-04-15)

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. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 23885–23899). Curran Associates, Inc.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018, September). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness..

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). Las Vegas, NV, USA: IEEE. doi: 10.1109/CVPR.2016.90

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10428–10436).

Ridnik, T., Ben-Baruch, E., Noy, A., & Zelnik-Manor, L. (2021, June). ImageNet-21K Pretraining for the Masses..

Rotermund, D., Scialom, E., Repnow, M., Herzog, M., & Ernst, U. (2024, April). *davrot/percept_simulator_2023: V1.0.0 (neuroprosthesis).* Zenodo. doi: 10.5281/zenodo.10978899

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020, November). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, *108*(3), 413–423. doi: 10.1016/j.neuron.2020.07.040

Tan, M., & Le, Q. (2019, 09–15 Jun). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 6105–6114). PMLR.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012, November). A Century of Gestalt Psychology in Visual Perception I. Perceptual Grouping and Figure-Ground Organization. *Psychological bulletin*, *138*(6), 1172–1217. doi: 10.1037/a0029333

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

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 12310–12320). PMLR. (ISSN: 2640-3498)