Evaluating model-to-human alignment on image occlusion

Ehsan Tousi
Neuroscience Graduate Program
Department of Psychology
Western University
London, ON, Canada
ekahooka@uwo.ca

Haider Al-Tahan
Neuroscience Graduate Program
Department of Psychology
Western University
London, ON, Canada
haltaha@uwo.ca

Farzad Shayanfar
Research Volunteer
Western University
London, ON, Canada
farzad.shayanfar@hotmail.com

Marieke Mur
Department of Psychology
Department of Computer Science
Western University
London, ON, Canada
mmur@uwo.ca

Abstract

Object recognition under occlusion is a significant challenge for both artificial and biological visual systems. We compared human performance with that of deep neural networks on a rapid object categorization task using systematically occluded natural images. Vision Language Models (VLMs) not only matched but exceeded human performance on heavily occluded images. In contrast, most Vision Only models showed steep performance drops. Confusion matrix analyses revealed that VLMs make semantically meaningful errors similar to humans, whereas Vision Only models show systematic biases toward specific categories regardless of input. VLMs' integration of linguistic context appears to enable more human-like inference of occluded object parts, suggesting a more object-centric approach compared to traditional pixel-based models. These results highlight the importance of multi-modal integration in developing more human-aligned visual recognition systems that maintain robustness under challenging viewing conditions.

Keywords: object recognition; occlusion; human visual perception; neural networks; vision language models; multi-modal learning

Introduction

A growing body of work in psychology and neuroscience has challenged the idea that deep neural networks (DNNs) are adequate models of human vision (Bowers et al., 2022; Wichmann & Geirhos, 2023). Although DNNs perform well on standard object classification tasks, they often fail to replicate hallmark features of human object perception (Peters & Kriegeskorte, 2021). These discrepancies have raised concerns about the alignment between DNNs and human vision, especially in tasks that probe deeper visual understanding.

Occlusion, in particular, challenges models to go beyond surface-level classification and approximate the richer inferences humans draw from limited visual input. Humans can typically identify objects that are only partially visible – such as a pedestrian stepping behind a car – while the extent to which DNNs succeed under similar conditions remains unclear. Computer vision and human experimental psychology often rely on different methodologies (Hassabis, Kumaran, Summerfield, & Botvinick, 2017). As a result, few studies have directly compared human and model performance using the same stimuli and task demands (but see Tang et al., 2018; Zhu, Tang, and Yuille, 2019).

Here, we evaluate the alignment between DNNs and humans on object recognition under occlusion, using a tightly controlled paradigm with circular aperture masks. While less naturalistic than real-world occlusion, this setup enables identical stimuli across systems and systematic control over visible information. We test a diverse set of models – from vision-only to vision-language architectures – to examine which design features yield more human-like performance. Our approach provides a framework for studying human-model alignment and can be extended to more naturalistic scenarios.

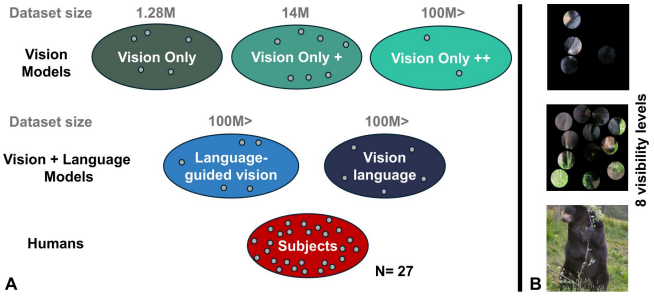

Figure 1: (A) Models (n = 24) evaluated in this study, grouped by learning objective and data scale. Vision Only models (green ovals) were trained on 1.28M, 14M, or 100M+ images using category supervision or self-supervised objectives. Language+Vision models (blue ovals) include Language-Guided Vision models (trained with text supervision, e.g., hashtags) and Vision-Language models (trained on image-text pairs with both modalities encoded). Dots within ovals represent individual models. Human participants (red oval, n= 27) are shown for reference. (B) Example stimuli illustrating the circular occlusion masks used to vary image visibility.

Methods

Human experiment. Twenty-seven adult human participants completed the experiment online. We used a rapid forced-choice categorization task where stimuli were presented for 200 ms followed by a visual mask. Stimuli were ImageNet validation images shown at eight visibility levels – ranging from 7% to 100% visibility – created using non-overlapping circular masks randomly placed across each image (Fig. 1B). Categories were 16 ImageNet-compatible, basic-level object categories from the MS COCO database, including "dog" and "car" (Geirhos, Janssen, et al., 2018). Participants performed 896 trials each (˜30 minutes of data).

DNN experiment. The same stimuli were presented to 24 DNNs spanning five model classes varying in learning objective and data scale (Fig. 1A). All models were trained, or fine-tuned on ImageNet-1K, enabling direct comparison to human performance on the same object categorization task (Geirhos, Temme, et al., 2018).

Model-to-human alignment. We computed categorization accuracy and category confusion matrices for both models and humans. To assess alignment, we correlated the diagonals (accuracies per category) and off-diagonals (patterns of confusion between categories) of the matrices between humans and models. Positive correlations indicate that models resemble humans in which categories are challenging (diagonals) or in which categories tend to be confused (off-diagonals). As a reference, we also computed human-to-human alignment using the same metrics.

Results

The occlusion manipulation was effective: both human and model accuracy dropped substantially under occlusion (Fig.

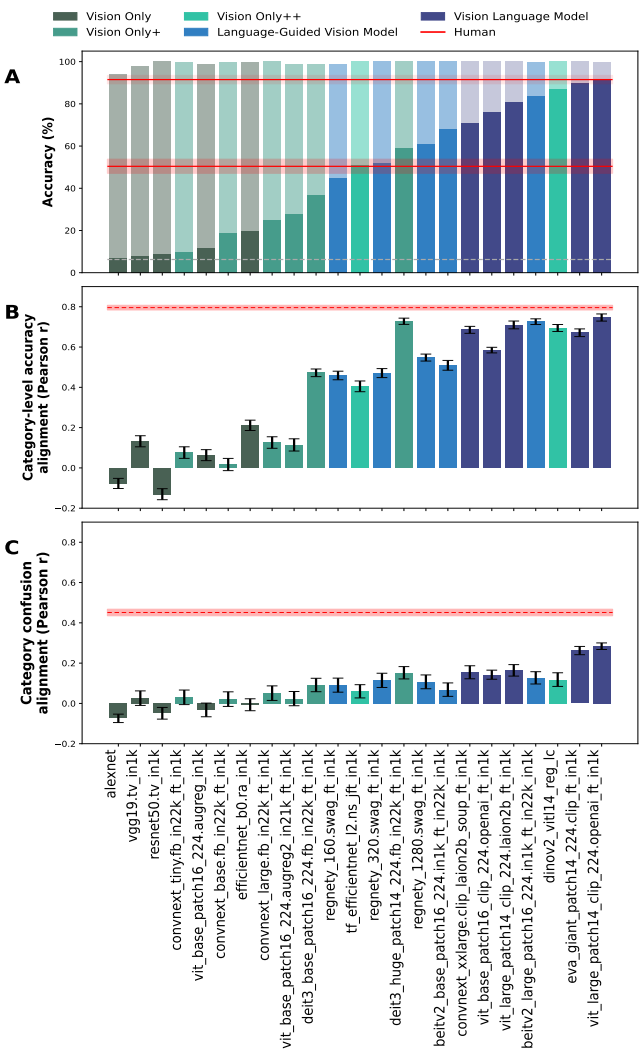

Figure 2: Model-to-human alignment on object recognition under occlusion. (A) Categorization accuracy for fully visible (lighter bars) and occluded (darker bars) images (each bar is the mean across the remaining seven visibility levels). Horizontal red lines indicate average human performance; shaded areas show 95% confidence intervals. Bars show model performance, with bar colors indicating model class as detailed in Fig. 1. The dotted gray line marks chance-level performance. (B) Category-level accuracy alignment for occluded images (averaged across all seven levels). Bars show the average Pearson correlation between each model's category-wise accuracy (i.e., the diagonal of its confusion matrix) and the corresponding human accuracies. Error bars represent the standard error of the mean. Red dashed line and shaded areas indicate human-to-human alignment (average inter-subject correlation and 95% confidence interval). (C) Category confusion alignment for occluded images (averaged across all seven levels). Same conventions as in panel B, but using the off-diagonal entries of the confusion matrix.

2A). While performance on fully visible images was uniformly high, accuracy varied widely under occlusion, revealing differences in model robustness. Accuracy increased with image visibility in both humans and models (not shown).

Several models outperformed humans on the occluded object categorization task (Fig. 2A). This was most evident for Vision+Language models, which maintained robust performance even under severe occlusion, but also included a Vision Only model trained on 100M+ images. In contrast, traditional Vision Only models trained on less data showed steep performance declines with increasing occlusion. These findings highlight meaningful variation in robustness across model classes under identical task conditions.

Models with higher performance generally also showed greater similarity to humans in terms of which categories were most difficult (Fig. 2B). However, none reached the level of human-to-human alignment – all fell below the 95% confidence interval – suggesting that even the best models fail to fully match human patterns of category difficulty.

When examining category confusions, a clearer gap emerged. Vision-Language models (VLMs) performed best, with two models reaching about half the human-to-human alignment, while others showed weak correspondence to human confusions (Fig. 2C). VLMs made semantically meaningful errors – e.g., confusing dogs with bears or trucks with cars – similar to humans, while Vision Only models often exhibited systematic biases toward specific categories such as "clock" and "bottle", whose visual features resemble the circular masks. These results suggest that models at best only partially replicate human category confusions.

Conclusions

Our findings suggest that Vision Language models exhibit more human-like behavior in object recognition under occlusion than traditional Vision Only models. Unlike models that rely primarily on pixel-level features, Vision Language models can integrate linguistic context that potentially helps infer occluded object parts through learned associations. These results highlight the potential of language-guided learning to support more robust, object-centered representations – particularly under challenging, information-limited conditions.

Several factors may contribute to why some models outperformed humans. Time-constrained viewing likely limited human use of context or top-down strategies, while models trained on large-scale datasets may more readily exploit contextual cues. The strong performance of language-guided models suggests that integrating linguistic structure into visual learning could support more abstract or amodally complete representations. Future work should explore how language shapes the internal organization of visual models – and whether this brings us closer to capturing the flexibility of human visual inference.

Acknowledgements

This work was funded by a Natural Sciences and Engineering Research Council Discovery Grant (RGPIN-2019-06741).

References

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2022, April). *Deep Problems with Neural Network Models of Human Vision.* PsyArXiv. doi: 10.31234/osf.io/5zf4s

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2018, December). *Comparing deep neural networks against humans: object recognition when the signal gets weaker.* arXiv. (arXiv:1706.06969 [cs]) doi: 10.48550/arXiv.1706.06969

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, *31*.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017, July). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. doi: 10.1016/j.neuron.2017.06.011

Peters, B., & Kriegeskorte, N. (2021, September). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, *5*(9), 1127–1144. (Number: 9 Publisher: Nature Publishing Group) doi: 10.1038/s41562-021-01194-6

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., ... Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840.

Wichmann, F. A., & Geirhos, R. (2023, September). Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? *Annual Review of Vision Science*, *9*(Volume 9, 2023), 501–524. (Publisher: Annual Reviews) doi: 10.1146/annurev-vision-120522-031739

Zhu, H., Tang, P., & Yuille, A. L. (2019). Robustness of object recognition under extreme occlusion in humans and computational models. *CoRR*, *abs/1905.04598*.