Probing Human Visual Strategies Using Interpretability Methods for Artificial Neural Networks

Yousif Kashef Alghetaa (yousif95@yorku.ca)
Department of Biology, York University
Toronto, Ontario, M3J1P3, Canada

Simon Kornblith (simon@simonster.com)
Anthropic, USA

Kohitij Kar (k0h1t1j@yorku.ca)
Department of Biology, York University
Toronto, Ontario, M3J1P3, Canada

Abstract

Unraveling human visual strategies during object recognition remains a challenge in vision science. Existing psychophysical methods used to investigate these strategies are limited in accurately interpreting human decisions. Recently, artificial neural network (ANN) models, which show remarkable similarities to human vision, provide a window into human visual strategies. However, inconsistencies among different techniques hinder the use of explainable AI (XAI) methods to interpret ANN decision-making. Here, we first develop and validate a novel surrogate method, *in silico*, using behavioral probes in ANNs with explanation-masked images to address these challenges. Finally, by identifying the XAI method and ANN with the highest human alignment, we provide a working hypothesis and an effective approach to explain human visual strategies during object recognition – a framework relevant to many other behaviors.

Keywords: Object Recognition; XAI; ANN; Explanation masked images

Introduction

Humans can rapidly and accurately identify and categorize objects. Despite the apparent ease with which humans perform this task, the underlying visual strategies employed by the human brain remain largely unknown (Kar & DiCarlo, 2023). Existing psychophysical methods, such as Bubbles (Gosselin & Schyns, 2001) and Classification Images (Eckstein & Ahumada, 2002), have limitations in accurately interpreting human decisions and may not fully capture the complexities of human visual processing (Murray, 2011). ANN models

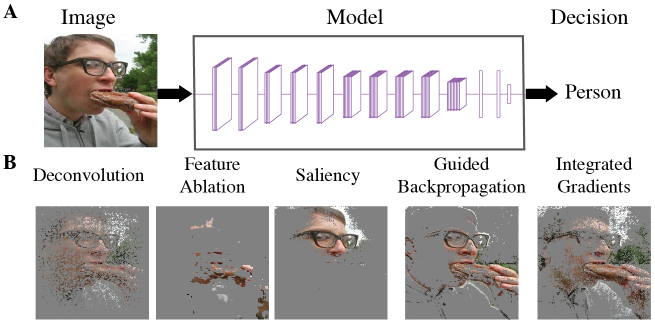

Figure 1: **A.** ANNs perform the same visual object recognition behavior as humans. **B.** Different XAI methods produce significantly different image attribution maps (shown here as a filtered version of the images).

have shown remarkable similarities to human vision (Khaligh-Razavi & Kriegeskorte, 2014; Rajalingham et al., 2018) and could provide insights into human visual strategies (Kar, Kornblith, & Fedorenko, 2022) (**Figure 1A**). However, interpreting ANN decision-making is challenging since the XAI methods designed to address this issue are hindered by inconsistencies (Hooker, Erhan, Kindermans, & Kim, 2019) among techniques (**Figure 1B**) and the need for full model access. To overcome these challenges, we propose a method combining XAI tools in ANNs and human behavioral testing to discover human decision-making strategies.

Results

We developed a behavioral method that bypasses direct explanation comparison between models.

Estimating the true differences in explanations

We define a **Target** model, e.g., ResNet-50 (He, Zhang, Ren, & Sun, 2016), **Figure 2**, and a **Reference** model. While ultimately, we use humans as the Reference model, we tested multiple image-computable, fully differentiable models with varied architecture and learning (e.g., AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), **Figure 2**) to develop and validate our method. An explanation of a model's output is a heat map indicating how input image features contribute to the output (**Figure 2**; mid-panels). We estimated the ground truth rank order in the similarity of explanations between the Target and Reference models by comparing feature attribution maps produced by ten explanations using L2-distance metrics (200 natural images, spanning 10 object categories from the MS COCO dataset (Lin et al., 2014)). We aim to recover this rank order using a human-compatible surrogate method without full access ("looking under the hood") to the Reference model.

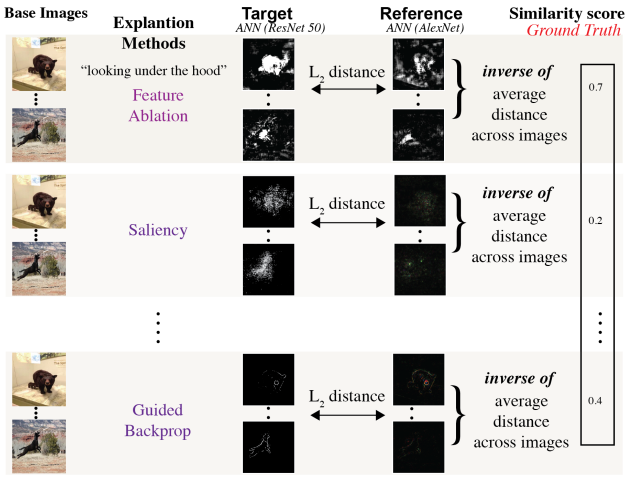

Figure 2: Estimating similarity between ResNet-50 (Target), and AlexNet (Reference) explanations using L2 distance across XAI outputs.Feature attribution maps are generated for each image using 10 XAI methods. The inverse of the mean L2 distance between the maps is used as similarity scores.

Generation of EMI

We generated filtered versions of the original images (explanation masked images, EMI) by retaining the top percentiles of informative pixels based on the feature attribution map of each explanation for the Target model((Hooker et al., 2019)). The EMIs are created in a two-step process (**Figure 3A, B**): (*1*) explanation generation using methods like Saliency (Simonyan,

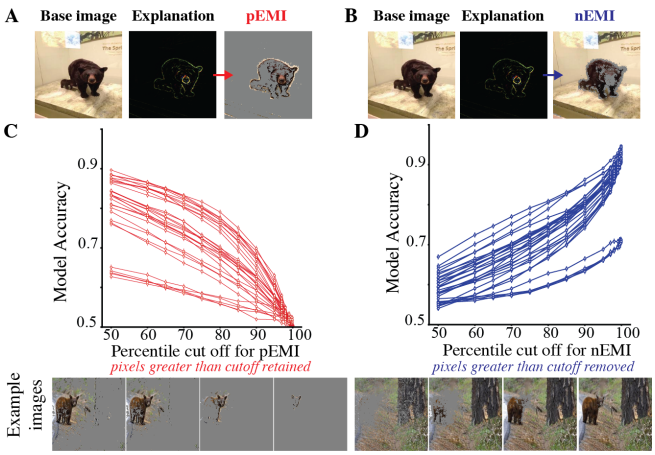

Figure 3: **Estimating EMI and validating it with model accuracy tests. A.** Generation of positive (by retaining the pixels greater than a cut-off) and negative (by removing the pixels greater than a cut-off) EMI. **B.** Reference model accuracies (on EMI from multiple Target images) reflecting the decrease (in red) and increase (in blue) in performance consistent with the expected changes with EMI cut-off levels.

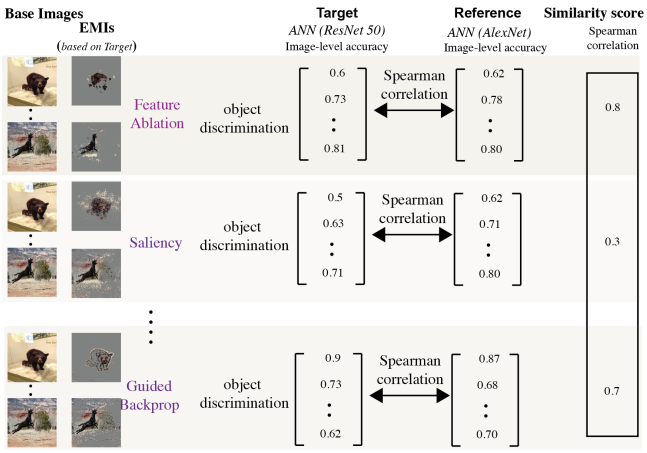

Figure 4: **Behavioral tests on EMI.** EMIs generated from explanation methods (for Target model) are presented to the Target and Reference models. The image-level accuracy pattern is correlated between the models to get a similarity score.

Vedaldi, & Zisserman, 2013) or Occlusion (Zeiler & Fergus, 2014) to rank pixels, and (*2*) percentile cutoff calculation and separation to generate two types of EMIs - positive EMI (pEMI) with top 'x' percentile pixels and negative EMI (nEMI) with lower '100 - x' percentile pixels. We experiment with various cutoffs to evaluate the impact of significant versus non-significant features. We hypothesized that the way EMIs drive the behavior of two systems, computed as image-level behavioral accuracies similar to (Rajalingham et al., 2018), might be symptomatic of how similar the underlying explanations (visual strategies) used to generate the EMIs are (**Figure 4**).

Validation of the proposed surrogate method

We measured the behavioral accuracies (previously explained in (Rajalingham et al., 2018; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019)) of all ANNs on the EMIs generated by 6 models. The Spearman correlation between the image-level accuracies (Target-vs-Reference)for the EMIs across each XAI method (**Figure 4**) provided the rank order of their behavior similarity. A strong correlation between these rankings and the ground truth (**Figure 2**) would validate our approach. Indeed, we observed significant positive correlations (Spearman R $\sim$0.7, all models, **Figure 5A**). We observed that nEMIs produced lower correlations compared to pEMIs. Therefore, we only used the pEMI for the human behavioral study.

Approximating human explanations

Next, we tested human subjects (n=300; pooled) and measured their object discrimination performances (methods identical to (Kar et al., 2019)) for the EMIs generated from different ANN models and XAI methods (**Figure 5B**). We observed that VGG-16 under the saliency-method (noise tunneling smooth gradient) (Smilkov, Thorat, Kim, Viégas, & Wattenberg, 2017) method yields the highest alignment (Spearman R=0.45) with human behavioral patterns. Therefore, this remains our working hypothesis for the human image attribution map during object recognition among the tested alternatives – which can be further probed with newer, more brain-aligned models.

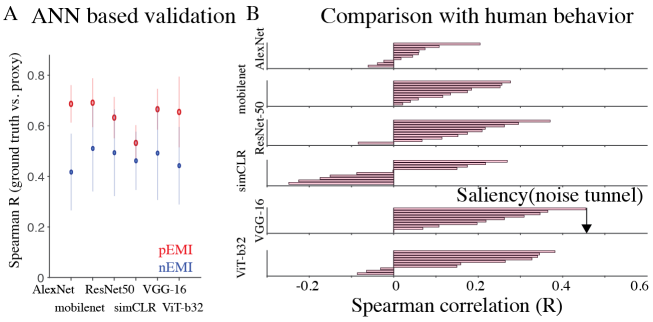

Figure 5: **A.** The surrogate method produces explanation rank-orders highly correlated ($\sim$0.7) with ground truth validating the approach. pEMIs yield higher correlations than nEMIs **B.** Distribution of alignment of human behavior and model predictions on EMIs from each XAI tool. VGG-16 with noise tunnel smooth gradient yields the best match ($\sim$0.45).

Conclusion

Our study introduces an accessible approach to unraveling human visual strategies. It addresses the limitations of existing psychophysical methods and the challenges of interpreting ANN decisions. This innovative method has the potential to bridge the gap between artificial and biological vision (Fel, Rodriguez Rodriguez, Linsley, & Serre, 2022), further advancing our understanding of human visual processing.

Acknowledgments

KK has been supported by funds from the Canada Foundation for Innovation (CFI), the Canada Research Chair Program, the Simons Foundation Autism Research Initiative (SFARI, 967073), the Canada First Research Excellence Funds (VISTA Program), and a Google Research Award.

References

Eckstein, M. P., & Ahumada, A. J. (2002). Classification images: A tool to analyze visual strategies. *Journal of vision*, *2*(1), i–i.

Fel, T., Rodriguez Rodriguez, I. F., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, *35*, 9432–9446.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, *41*(17), 2261–2271.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, *32*.

Kar, K., & DiCarlo, J. J. (2023). The quest for an integrated set of neural mechanisms underlying object recognition in primates. *arXiv preprint arXiv:2312.05956*.

Kar, K., Kornblith, S., & Fedorenko, E. (2022). Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nature Machine Intelligence*, *4*(12), 1065–1067.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, *22*(6), 974–983.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Murray, R. F. (2011). Classification images: A review. *Journal of vision*, *11*(5), 2–2.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part i 13* (pp. 818–833).