Demystify Deep-learning AI for Object Detection using Human Attention Data

Jinhan Zhang
jinhanz@hku.hk
Department of Psychology,
University of Hong Kong

Guoyang Liu
gyliu@sdu.edu.cn
School of Integrated Circuits,
Shandong University

Yunke Chen
cyk1028@connect.hku.hk
Department of Psychology,
University of Hong Kong

Antoni B. Chan
abchan@cityu.edu.hk
Department of Computer Science, City University of
Hong Kong

Janet H. Hsiao
jhhsiao@ust.hk
Division of Social Science, Hong Kong University of
Science & Technology

Abstract

Here we present a new Explainable AI (XAI) method to probe the functional partition in AI models by comparing features attended to at different layers with human attention driven by diverse task demands. We applied this method to explain an object detector Yolo-v5s in multi-category and single-category object detection tasks. We found that the model's neck showed higher similarity to human attention during object detection, indicating a reliance on diagnostic features in the neck, whereas its backbone showed higher similarity to attention during passive viewing, indicating salient local features encoded. With this understanding of its functional partition, using Yolo-v5s as a model for human cognition, our comparative analysis against human attention when providing explanations for object detection revealed that humans attended to a combination of diagnostic and salient features during explaining multi-category general object detection but attended to mainly diagnostic features when explaining single-category human/vehicle detection in driving scenarios.

Keywords: object detection; explainable AI; human attention; eye tracking; deep learning

Introduction

Nowadays, artificial intelligence (AI) has achieved outstanding performance in many computer vision tasks such as image classification and object detection. However, the black-box nature of deep neural networks (DNN) has obscured their internal decision-making mechanisms (Rudin, 2019). This impacts both machine learning scientists' evaluation of models and human users' trust in the system, which drives an increasing demand for model interpretability and explanations that are accessible to humans (Mittelstadt et al., 2019; Hsiao et al., 2021). Many explainable AI (XAI) methods thus have been developed. In computer vision, a prevalent strategy involves generating saliency maps that highlight features attended to by AI, through gradient-based or perturbation-based methods (Adadi & Berrada, 2018). However, these methods typically offer limited information about the model's functioning. For example, the convention of Grad-CAM (Selvaraju et al., 2020), a commonly used gradient-based XAI method, is to backpropagate to the last convolutional layer to provide a coarse localization map of features relevant to the model's output (Molnar, 2020), whereas the functional role of representations developed in different layers are often unclear.

Recently, a new concept of XAI, Artificial Cognition (Ritter et al., 2017; Taylor & Taylor 2021), has been proposed to approach the black box of DNNs by adapting the experimental traditions with which cognitive psychologists have long been addressing the similar black-box challenge in studying the human mind. It utilizes experimental psychology approaches to the understanding of machine behavior or DNNs (Goodfellow et al., 2009; Rajalingham et al., 2018; Richard-Webster et al., 2018). For example, Ritter et al. (2017) conducted experiments analogous to the shape bias tests well-established in human word learning to one-shot word learning models and identified a similar bias. Volokitin et al. (2017) examined different mechanisms of the visual crowding effect in DNNs with different architectures. These attempts showed the potential of experimental psychology methods in interpreting DNNs.

The feasibility of applying cognitive psychology methodologies to XAI arises from both the objectives of XAI and the parallels between human and artificial neural architectures. With the motivation to improve human understanding and trust in machine behavior, Miller (2017) argued for the need to evaluate XAI explanations with human data based on social sciences findings on what people require and expect in explanations. At the same time, the shared hierarchical and distributed processing properties of neuronal and artificial neural networks naturally lead us to invoke references from the knowledge representations in the human brain to comprehend AI (Cichy et al., 2016).

Specifically, findings on the hierarchical and functional structure of the human brain have inspired the study of the architecture of DNNs. The human visual cortex exemplifies the hierarchical structure involving multiple levels of feature abstraction and the functional distributions of human cognition for a complicated task. Low-level visual features, such as orientations and contours, are encoded by the neurons in the primary visual cortex (V1), and neurons in higher-level regions along the hierarchy respond to more complex features such as shape and depth. There is also a functional division into the dorsal and ventral pathways for spatial location and object identification functions, respectively (Prinz, 2012). Then, the inferior temporal area contained specialized regions for different object categories.

One representative case of regional specialization is the fusiform face area (FFA), first identified by Kanwisher et al. (1997) as a specialized module for face perception by comparing brain activations when participants were presented with face stimuli vs. other types of stimuli. It has also been evidenced by extensive behavioral, neuropsychology, and other brain-imaging studies (Puce et al., 1996; see Kanwisher

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

& Yovel, 2009 for a review). Meanwhile, researchers have debated on the actual function of the FFA, such as whether it is specialized for stimuli on which individuals have expertise, by comparing brain activation when participants performed different tasks associated with different levels of expertise (Bilalić et al., 2013; Gauthier et al., 2000; Gauthier, 2017; Righi et al., 2013). These examinations illustrated well how experimental investigations and empirical discussions on the functional partitioning of the brain enhanced our understanding of the information-processing mechanisms underlying the human mind (Kanwisher, 2017).

Analogous to the hierarchical representations for object recognition in the human brain, deep learning AI model's enhanced performance stems from multiple layers of feature abstraction (Bengio et al. 2013). However, in contrast to humans who are general problem solvers capable of performing different tasks to allow the use of experimental approaches to infer the functional organization of the brain., AI models typically are trained to perform a particular task, and thus we are unable to use exactly the same approach as human studies to examine their functional partitioning. However, since humans have the flexibility to perform different functions/tasks, if we know humans' information use when they perform different functions/tasks, we can use this knowledge to probe the functions of the representations developed at different layers of a deep learning AI model through similarity analysis.

Accordingly, here we proposed a new XAI method that uses the similarity of the important features used by AI at different layers to human attention when performing different functions/tasks on the input stimuli to probe the functional role of different layers in the AI model. As a proof of concept, here we focused on Yolo-v5s, a representative one-stage object detection model (Ultralytics 2021). We first obtained saliency maps highlighting features associated with AI's decisions using FullGrad-CAM (Liu et al., 2023) and ODAM (Zhao & Chan, 2023), two gradient-based methods derived from Grad-CAM with higher faithfulness and plausibility for explaining object detection models, from different layers in Yolo-v5s. To probe the functional role of different layers, we collected human attention data when they performed different tasks on the input images, including passive-viewing, object detection, and explanation (i.e., providing explanations of how a particular target is detected). As human attention is shown to be driven by task demands (Hsiao & Chan, 2023; Henderson, 2017), these tasks would result in different attention maps: In passive-viewing, human attention generally follows bottom-up saliency on interesting regions; in object detection, participants need to accumulate sufficient positive information indicating the existence of a target; in explanation, participants may attend to all positive and negative (i.e., discriminative) features relevant to identifying the target. We then examined the similarity of the human attention maps to the saliency maps obtained from different layers of the AI model, with the layers with the highest similarity to each task forming a functional partitioning of the model. In Study 1, we tested this method using images from Microsoft Common Objects in COntext (MSCOCO) dataset (Lin et al., 2014) to understand the functional partitioning of the layers in the AI model. In Study 2, we used a harder dataset, BDD-100K, a popular database for object detection during autonomous driving with occluded and degraded targets (Yu et al., 2020), to examine how the task demand change could help us further understand the functional partitioning of the model**.**

Study 1: General Object Detection

Methods

AI Model In this study, we used a Yolo-v5s model pre-trained on the MSCOCO dataset.

To investigate the functional partition across different stages in object detectors, we selected the last layers from 17 functional modules in Yolo-v5s, including all bottleneck blocks and convolutional blocks. We will present the similarity scores between Yolo-v5s and human attention in different tasks for all 17 layers. However, in this paper, the last four layers in the neck (F14-17 in Figure 1) were excluded from the statistical analysis and inference about layers' functional roles, as they are not directly comparable to human attention. The neck of Yolo-v5s functions to integrate features at different scales and outputs three feature maps to separate prediction heads for detecting small, medium, and large objects in an image, respectively. During the detection of a small object proposed by the small-object detecting head (Head 1), the feature map was output earlier from a middle layer in the neck (F13) and was sent to Head 1 through an immediate connection. This resulted in an empty saliency map at F17 due to the zero gradients with respect to Head 1. Therefore, in multi-object detection, a low similarity between saliency maps from F17 and human attention during detection could be attributed to the layer ignoring the object (because of its scale) instead of differences in the patterns of features attended by AI & humans, assuming that humans have successfully detected all targets.

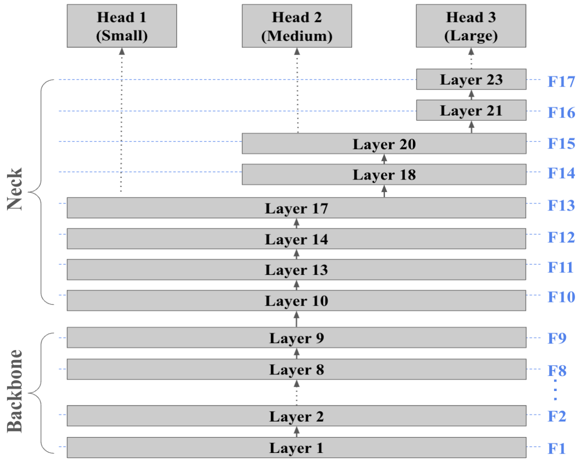

Figure 1: Three pathways in Yolo-v5s to detect small, medium, and large objects. Numbers in black indicate the original order of the functional modules. Module numbers in blue are the functional naming we used in this paper.

Human Attention Data Participants. We recruited a sample of 118 participants aged 18 to 37 (*M* = 21.67; *SD* = 3.78; 91 females) with normal or corrected-to-normal vision. 59 of them (42 females) participated in the object detection task and explanation task, and 59 (49 females) participated in the passive viewing task.

Materials & Apparatus. The materials consisted of 160 images from the MSCOCO dataset. In 80 classes, we used two test images containing the object. Two extra images were selected and used in practice trials. For human experiments, we resized and padded the images to ensure the object for explanation was large enough for participants to view. Specifically, if the largest side of that object is smaller than 230 pixels, the images were resized, and we downscaled images with big objects so that objects had similar sizes across all images. We cropped the images so that the object's position in the cropped image is relatively the same as in the original image. All images had a 1270 x 784 resolution, and we kept the original width-height ratio and resized the width or height to fit 1270 or 784. The boundaries are padded with a transparent background. To generate XAI saliency maps, we used the same images under these operations without the extra padding at boundaries.

The programs were built using SR Research Experiment Builder (version 2.3.38) and controlled by a PC computer operating on a Win10 system. The stimuli were displayed at the horizontal center of a 15.6-inch FHD Monitor (1280 x 1024 resolution). We left ¼ of the space above the image and ¾ of the space below the image for the textbox. The stimuli span 58° x 38° of visual angle at a viewing distance of 30 cm. Participants' eye movements were recorded using EyeLink 1000 Plus, installed on a tower mount and set to head-stabilized mode, with a chin rest to keep participants' heads stable. A standard nine-point calibration procedure was performed before the experiment and whenever the drift check error exceeded 1° of visual angle.

Procedure. In the object detection task, participants were instructed to detect common objects for 160 images in 4 blocks, 40 images per block, with randomized block orders and within-block trial orders. They were presented with the full list of the labels before the experiment to ensure that they understood all the labels' meanings. Each trial started with a drift check at the screen center and then a fixation cross at the center of the stimuli presentation for 500 ms. Participants were presented with a class label for 1000 ms and asked to detect all the objects belonging to the class label in the image they were about to view and pressed a key as soon as they finished detecting. To assess detection performance, immediately after the key press, participants were asked to use a mouse click to place a marker at each detected object location on a blank screen. Then, they were asked to click again on the same objects they had clicked previously on the original image to confirm their selection (Figure 2A). Before the formal trials, they completed two practice trials to ensure their understanding of the task. Participants' eye movements when they viewed the stimuli and before pressing the key to indicate the end of detection were used for analysis. We separated the visual search phase from the clicking phase to avoid interference from sensorimotor planning of clicking during the eye movement recording of the visual search task.

In the object explanation task, participants were instructed to provide explanations of 160 images which were blocked and randomized the same way as in the detection task. The images were presented with the largest target object labelled with a blue bounding box. Each trial started with a drift check at the screen center. Then, after a 500-ms fixation cross, participants saw the image's class label for 1000 ms and saw the image (Figure 2B). They were asked to type an explanation in a textbox about why the object in the bounding box should be identified as its labeled class. We prompted them to imagine explaining to someone who has no existing knowledge of the visual object and to provide sufficient information to help a person identify the object or assign correct labels to the object. Before starting the formal trials, participants were instructed to complete two practice trials with the two extra images. Experimenters provided feedback to them and made sure they understood the task before the formal trials. Their eye movements were recorded when they viewed the images. The same set of participants performed the detection task and the explanation task, and the explanation task was conducted last to avoid introducing a familiarity effect to the detection task.

In the passive viewing task (Figure 2C), each trial started with a drift check at the screen center. A fixation cross was then displayed at the center for 500 ms. Participants were asked to view 160 images one at a time, each for 5 s, and rated how much they liked the image on a Likert scale of 1 to 5. The images were blocked and randomized the same way as in the detection task. Their eye movements were recorded when they viewed the images.

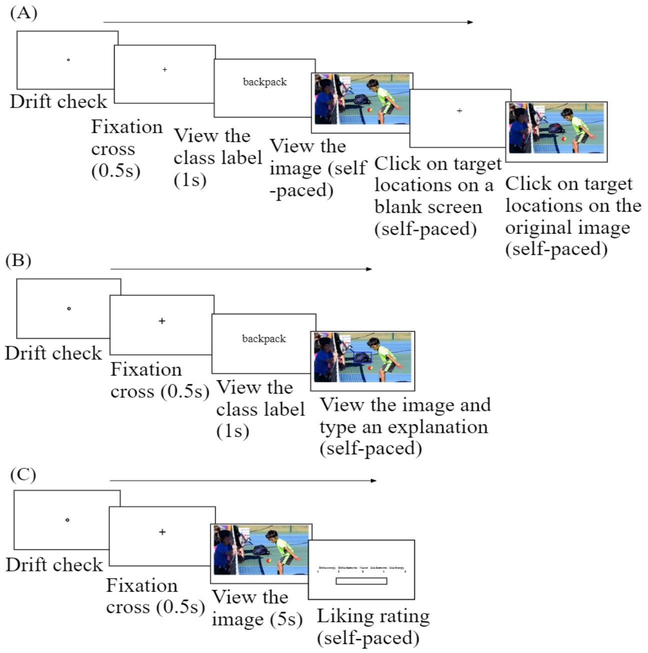
(A)
Drift check
Fixation cross (0.5s)
backpack
View the class label (1s)
View the image (self -paced)
Click on target locations on a blank screen (self-paced)
Click on target locations on the original image (self-paced)
(B)
Drift check
Fixation cross (0.5s)
backpack
View the class label (1s)
View the image and type an explanation (self-paced)
(C)
Drift check
Fixation cross (0.5s)
View the image (5s)
Liking rating (self-paced)

Figure 2: Procedure for collecting human attention data for the (A) Detection, (B) Explanation, and (C) Passive Viewing Tasks on the MSCOCO dataset.

Human Attention Maps & XAI Methods We generated human attention maps by applying a Gaussian smoothing kernel with a 21-pixel SD (equivalent to 1° of visual angle) on each fixation point over all subjects in each task.

To generate AI saliency maps comparable to human attention maps during object detection, we applied FullGrad-CAM, an XAI algorithm that was derived from Grad-CAM and was specially designed for object detection models. Let N_{obj} be the total number of objects detected, then the FullGrad-CAM saliency map for all objects of that class is:

$$S_F = \sum_{m=1}^{N_{obj}} \mu \left(ReLU \left(\sum_{k=1}^{N_{ch}} \frac{\partial y^m}{\partial A^k} \odot A^k \right) \right), \qquad (1)$$

where y_m is the output classification probability of m-th detected object, A^k is the activation map from the k-th layer with N_{ch} channels in total, $ReLU$ is the rectified linear unit activation function, μ is the max-min normalization function that ensures the saliency maps scale between 0 to 1, and $\odot$ represents the Hadamard product. By replacing the global average pooling operation in the conventional Grad-CAM method, FullGrad-CAM preserves the spatial information in the gradient maps, which is particularly informative for object detection.

To compare object detectors' attention with human attention during the explanation task, where a specific object in an image is asked for explanation, ODAM is applied to generate the corresponding instance-specific saliency map of the object. An ODAM saliency map for the m-th object is:

$$S_O{}^{(m)} = ReLU \left(\sum_{k=1}^{N_{ch}} \frac{\partial y^m}{\partial A^k} \odot A^k \right), \qquad (2)$$

When there is only one ground truth target, ODAM's output is equivalent to a FullGrad-CAM saliency map. While in multi-object detection, FullGrad-CAM can be viewed as a normalized saliency map that is averaged over multiple ODAM maps corresponding to all detected objects.

In this study, we generated saliency maps using both XAI methods on all 17 convolutional layers selected. We measured the similarity between human and XAI saliency maps using Pearson correlation coefficient (PCC) to probe the function of each layer across different stages in object detectors by leveraging our understanding of human attention strategies, thus revealing the functional partitioning of the network. In particular, we correlated FullGrad-CAM saliency maps from each layer separately to human attention maps in the detection and passive viewing tasks, and between ODAM saliency maps and human attention maps in the object explanation task. Out of 160 images, we used only those where Yolo-v5s successfully detected objects to ensure a valid comparison to human attention.

Results

Here we examined the human-AI-attention similarity and its variation across layers in Yolo-v5s (Figure 4). Figure 3 illustrates examples AI saliency maps and human attention maps during different tasks. For each task, a paired-samples t-test was conducted to compare the human-AI similarity score difference between the neck and the backbone, where the score is defined as the average PCC value over all layers within the neck and within the backbone, for 144 images where Yolo-v5s made successful detections. Note that the last four layers in the neck (F14-F17) are excluded from the analysis. Our analysis indicates that attention maps from the neck of Yolo-v5s (*M* = .362, *SD* = .115) exhibited higher similarity to human attention maps during *object detection,* compared to the backbone (*M* = .298, *SD* = .082), *t*(143) = 7.12, *p* < .001. When comparing AI and human attention during *passive viewing*, we found that the neck (*M* = .192, *SD* = .161) exhibited lower similarity to human attention than the backbone (*M* = .208, *SD* = .126) with marginal significance, *t*(143) = -1.94, *p* = .00539. Lastly, when comparing AI and human attention during *object explanation*, the neck (*M* = .292, *SD* = .138) showed higher similarity than the backbone (*M* = .276, *SD* = .096) with marginal significance, *t*(143) = 1.97, *p* = .0504.

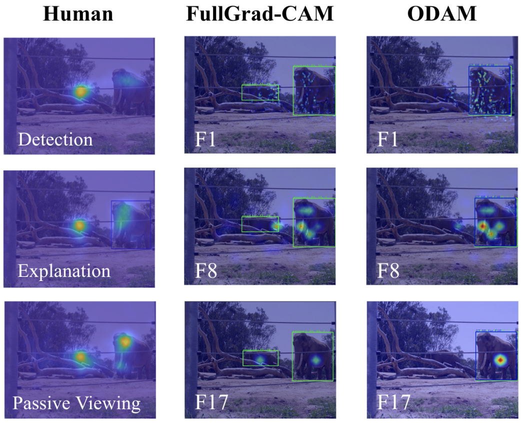

Figure 3: Example human attention maps and XAI saliency maps.

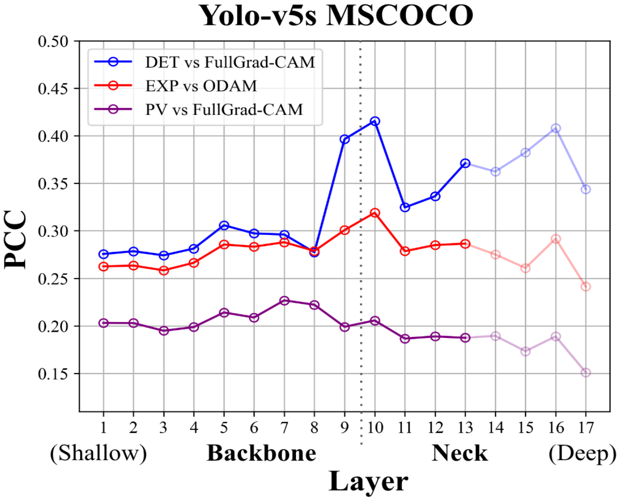

Figure 4: The similarity between human attention maps and XAI saliency maps generated from backbone layers (F1-F9) and neck layers (F10-F17) in Yolo-v5s during different tasks for MSCOCO object detection, as measured by Pearson correlation coefficient (PCC). The similarity scores of F14-17, excluded from statistical analysis, are displayed transparently.

Study 2: Object Detection in Driving Scenarios

Methods

AI Model and Dataset We selected BDD-100K, a popular database for object detection during autonomous driving (Yu et al., 2020). The dataset includes 10 target categories, from which we chose 'car', 'truck', and 'bus' as the target categories for vehicle detection, and 'person' and 'rider' for human detection. We trained Yolo-v5s, from scratch on a self-curated training set with these five target labels. We followed the same criteria of layer selection in Study 1.

Human Attention Data Participants. We recruited 79 participants aged 18 to 37 (*M* = 24.03; *SD* = 4.60; 62 females) with normal or corrected-to-normal vision. 60 of them (48 females) participated in the vehicle task and 60 (48 females) participated in the human task. To identify human experts for driving-scene objects, they all had driving licenses.

Materials & Apparatus. According to the suggestion of existing research that young adults have a location memory limit of 3 to 5 items (Cowan, 2010), we selected stimuli from the dataset that contains 1 to 4 targets. We selected 160 images for vehicle detection and 160 for human detection.

In the detection task, stimuli (1280 x 720 pixels) were displayed at the center of a 15.6-inch monitor (1280 x 1024 pixels) which spans 34.2 ° x 20.8° of visual angle at a viewing distance of 55 cm. In the explanation task, they were presented 30 pixels to the top of the screen with space for the textbox below the image. Participants' eye movements were recorded using the same equipment and setup as Study 1.

Procedure. In the detection task, the procedure was identical to that in Study 1 except that participants were not presented with the target label in each trial (Figure 5A). Before the experiment, they were instructed to detect all cars, buses, and trucks in the vehicle detection task and pedestrians or riders in the human detection task. The explanation task procedure was identical to that in Study 1 except that participants did not see the object label in each trial (Figure 5B). Additionally, in each image as the target for explanation, we selected the target object influenced by the most difficult conditions: occlusion (with part of the object occluded by other objects or image boundary) and degradation (with object features influenced by uneven lightening, shadow, reflection, motion blur, or night vision). These conditions were determined according to the majority choice of three raters with good inter-rater reliability (Occlusion: α = .881; Degradation: α = .877; Cronbach, 1951). Among targets influenced by the same number of difficult conditions, we chose the one with the largest bounding box.

Human Attention Maps & XAI Methods The procedure of human and AI attention maps generation was identical to Study 1, except that due to the difference in image sizes, here we applied a Gaussian smoothing kernel with a 30-pixel SD (equivalent to 1° of visual angle on BDD images) when generating human attention maps.

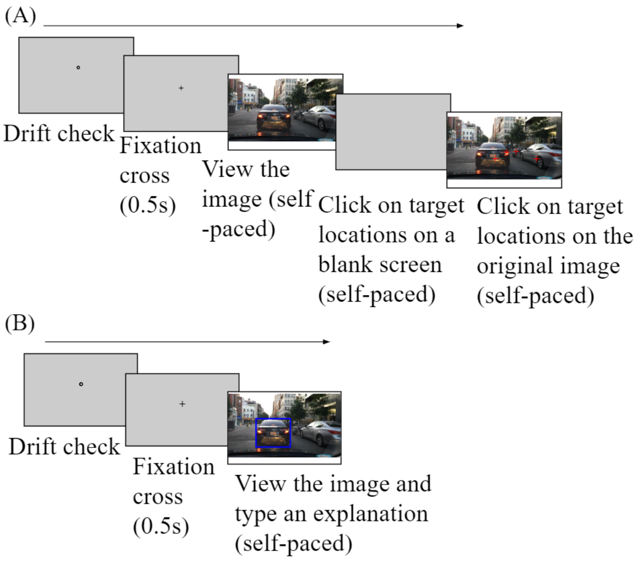
(A)
Drift check
Fixation cross (0.5s)
View the image (self-paced)
Click on target locations on a blank screen (self-paced)
Click on target locations on the original image (self-paced)
(B)
Drift check
Fixation cross (0.5s)
View the image and type an explanation (self-paced)

Figure 5: Procedure for collecting human attention data for the (A) Detection and (B) Explanation Task using the BDD dataset.

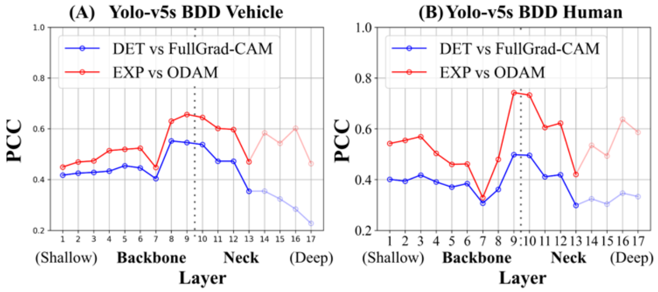

Figure 6: The similarity between human attention maps and XAI saliency maps generated from backbone layers (F1-F9) and neck layers (F10-F17) in Yolo-v5s for (A) vehicle detection and (B) human detection, as measured by Pearson correlation coefficient (PCC). The similarity scores of F14-17, excluded from statistical analysis, are displayed transparently.

Results

Yolo-v5s made successful detections on 86 images for vehicle detection and 68 images for human detection. Comparing between AI and human attention during *vehicle explanation,* the neck of Yolo-v5s (*M* = .579, *SD* = .153) exhibited higher similarity to human attention than the backbone (*M* = .521, *SD* = .116), *t*(85) = 4.33, *p* < .001. Similarly, for *human explanation*, the neck of Yolo-v5s (*M* = .595, *SD* = .084) also exhibited higher similarity to human attention than the backbone (*M* = .516, *SD* = .072), *t*(67) = 6.86, *p* < .001. Comparing AI and human attention during *vehicle detection*, there was no significant difference between the neck (*M* = .459, *SD* = .147) and the backbone (*M* = .456, *SD* = .135), *t*(85) = .299, *p* = .766. Similarly, for *human detection*, the difference between the neck (*M* = .406, *SD* = .136) and the backbone (*M* = .392, *SD* = .114) was not significant, *t*(67) = 1.63, *p* = .108.

Discussion

Here we introduced a new XAI method to investigate the functional partition in an AI model (Yolo-v5s) by comparing it with human cognition. Inspired by how cognitive neuroscientists investigate the functional architecture of the brain by asking participants to perform different cognitive tasks, and the observation that human attention is task-specific, here we utilized human attention in different tasks and examined its similarity to the representations from different layers of Yolo-v5s to comprehend their functional roles.

We first tested our method on Yolo-v5s for general object detection on the MSCOCO dataset to examine its functional partitioning. We correlated saliency maps from each layer with human attention during object detection, object explanation, and passive viewing tasks, and compared the overall similarity scores between the backbone and the neck of Yolo-v5s under these three task conditions.

Interestingly, Yolo-v5s exhibited an opposite trend in human-AI similarity score when compared with human's attention during *detection* and *passive viewing*. Specifically, the neck had a higher similarity score than the backbone in the comparison with human attention during *detection*, and a lower similarity score than the backbone for *passive viewing* (although this difference was not significant in the t-test). During object detection, humans attend to diagnostic features for identifying the target objects from background distractors (e.g., Qi et al., 2023, although in object categorization). In the model, a higher similarity to human object-detection attention maps in the neck than in the backbone suggested that diagnostic features for identifying the target objects were better represented in the neck than in the backbone. In contrast, during passive viewing, there was no specific task demand and human attention was guided mainly by bottom-up saliency (Elazary & Itti, 2008). Therefore, in the model, a higher similarity to human passive-viewing attention maps in the backbone than the neck suggests that the backbone encoded more bottom-up saliency information.

Our findings above also contributed to an extended understanding of human's attention strategies. The similarity of Yolo-v5s's neck to human attention during *explanation* fell between the similarity scores for human attention during the other two tasks, and there was no significant difference in similarity between the neck and backbone. This result suggested that humans may employ a combination of diagnostic features and salient features when explaining object detection. This finding is congruent with previous studies comparing human and deep neural networks in image classification (Qi et al. 2023), which revealed that humans employed a focused fixation strategy on diagnostic features, attending to only sufficient information for making the decision, in contrast to an explorative fixation strategy on more relevant and contextual features when explaining how the object was classified to a given category.

We have also tested our methods on Yolo-v5s under more realistic and difficult conditions: vehicle and human detection/explanation in driving scenarios with occlusions and degradations using stimuli from BDD. The comparison with human attention during *object detection* showed a similar pattern to Study 1 with MSCOCO, where the neck exhibited higher similarity than the backbone (although this difference was not significant in the t-test). Interestingly, the comparison with human attention during *vehicle/human explanation* showed a similarity score pattern different from our findings in the first study: the neck exhibited higher similarity to human attention than the backbone. This result suggests that humans may use a different strategy to explain vehicle and human detection with images from BDD as compared with that used in explaining the detection of general object categories with images from MSCOCO. For explaining difficult human or vehicle detection, participants attended to more diagnostic features than salient features, in contrast to detecting objects of multiple general categories where both salient and diagnostic features were attended to. This explanation strategy change may be related to task demands. There were multiple categories in MSCOCO that differed in many low-level features. Thus, participants may use these differences in low-level salient features to explain how to identify an instance in one category against other categories. In contrast, detecting humans or vehicles with BDD involved detecting only one category. Therefore, the explanations may focus on the minimum amount of diagnostic/positive features that give them the confidence to identify it as a hit.

The method described in this paper has enhanced our comprehension of the mechanism of Yolo-v5s with a human-centered explanation. While preliminarily we interpreted the functions of layers based on the predefined backbone/neck split, our method can potentially be used to discover a more detailed human-attention-guided functional partitioning in more complicated AI systems, which may differ from the typical architectural partitioning specified by AI developers.

In conclusion, we introduced a new XAI method that probes the functional partition in an AI system by comparing its attended features at different layers with human attention driven by different task demands. Using object detection model Yolo-v5s as an example, we showed that in multi-category general object detection with MSCOCO, human attention for object detection had higher similarity to features in Yolo-v5s's neck than the backbone, suggesting that the neck encoded diagnostic features for detection. In contrast, human attention during passive viewing had the opposite trend, suggesting that the backbone encoded salient local features that humans typically attended to during passive viewing. By comparing the human-AI-similarity score of the explanation tasks, we found that humans attended to a combination of diagnostic and salient features during explaining multi-category general object detection but attended to mainly diagnostic features when explaining human/vehicle detection in driving scenarios. Thus, in addition to providing human-centered explanations on the functional partition of AI systems, our method demonstrated potential applications in understanding human cognition using the functional partition in AI as a model.

Acknowledgments

This study was supported by Research Grant Council of Hong Kong (TRS #T45-401/22N-4) and Strategic Research Grant from City University of Hong Kong (Prof. No. 7005995). We thank Yumeng Yang and Linjing Wang for their help in data collection and experiment design.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bilalić, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many Faces of Expertise: Fusiform Face Area in Chess Experts and Novices. *Journal of Neuroscience*, 31(28), 10206–10214. https://doi.org/10.1523/JNEUROSCI.5727-10.2011

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1), Article 1. https://doi.org/10.1038/srep27755

Cowan, N. (2010). The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Curr Dir Psychol Sci*, 19(1), 51-57. https://doi.org/10.1177/0963721409359277

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. Journal of Vision, 8(3), 3. https://doi.org/10.1167/8.3.3

Gauthier, I. (2017). *The Quest for the FFA led to the Expertise Account of its Specialization*. https://doi.org/10.48550/ARXIV.1702.07038

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2), Article 2. https://doi.org/10.1038/72140

Goodfellow, I., Lee, H., Le, Q., Saxe, A., & Ng, A. (2009). Measuring Invariances in Deep Networks. *Advances in Neural Information Processing Systems*, 22. https://proceedings.neurips.cc/paper_files/paper/2009/hash/428fca9bc1921c25c5121f9da7815cde-Abstract.html

Henderson, J. M. (2017). Gaze Control as Prediction. Trends in Cognitive Sciences, 21(1), 15–23. https://doi.org/10.1016/j.tics.2016.11.003

Hsiao, J. H., Ngai, H. H. T., Qiu, L., Yang, Y., & Cao, C. C. (2021). *Roadmap of Designing Cognitive Metrics for Explainable Artificial Intelligence (XAI)* (arXiv:2108.01737). arXiv. http://arxiv.org/abs/2108.01737

Hsiao, J. H., & Chan, A. B. (2023). Visual attention to own- vs. other-race faces: Perspectives from learning mechanisms and task demands. British Journal of Psychology, 114(S1), 17-20

Kanwisher, N. (2017). The Quest for the FFA and Where It Led. *The Journal of Neuroscience*, 37(5), 1056–1061. https://doi.org/10.1523/JNEUROSCI.1706-16.2016

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kanwisher, N., & Yovel, G. (2009). Face Perception. In G. G. Berntson & J. T. Cacioppo (Eds.), *Handbook of Neuroscience for the Behavioral Sciences* (1st ed.). Wiley. https://doi.org/10.1002/9780470478509.neubb002043

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), Article 7553. https://doi.org/10.1038/nature14539

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). Springer International Publishing.

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. H. (2023). *Human Attention-Guided Explainable Artificial Intelligence for Computer Vision Models*. arXiv. https://doi.org/10.48550/arXiv.2305.03601

Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences* (arXiv:1706.07269). arXiv. https://doi.org/10.48550/arXiv.1706.07269

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 279–288. https://doi.org/10.1145/3287560.3287574

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. https://christophm.github.io/interpretable-ml-book/

Prinz, J. (2012). *The Conscious Brain*. OUP USA.

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential Sensitivity of Human Visual Cortex to Faces, Letterstrings, and Textures: A Functional Magnetic Resonance Imaging Study. *Journal of Neuroscience*, 16(16), 5205–5215. https://doi.org/10.1523/JNEUROSCI.16-16-05205.1996

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2023). Explanation strategies for image classification in humans vs. current explainable AI. *Proceedings of the Annual Meeting of the Cognitive Science Society.* https://arxiv.org/abs/2304.04448

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. https://doi.org/10.1523/JNEUROSCI.0388-18.2018

RichardWebster, B., Anthony, S. E., & Scheirer, W. J. (2019). PsyPhy: A Psychophysics Driven Evaluation Framework for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9), 2280–2286. https://doi.org/10.1109/TPAMI.2018.2849989

Righi, G., Tarr, M. J., & Kingon, A. (2013). Category-selective recruitment of the fusiform gyrus with chess expertise. In *Expertise and Skill Acquisition* (pp. 261-280). Psychology Press.

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). *Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study* (arXiv:1706.08606). arXiv. http://arxiv.org/abs/1706.08606

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28(2), 454–475. https://doi.org/10.3758/s13423-020-01825-5

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision.* https://doi.org/10.1007/s11263-019-01228-7

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do Deep Neural Networks Suffer from Crowding? *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/c61f571dbd2fb949d3fe5ae1608dd48b-Abstract.html

Yang, A., Liu, G., Chen, Y., Qi, R., Zhang, J., & Hsiao, J. (2023). Humans vs. AI in Detecting Vehicles and Humans in Driving Scenarios. *Proceedings of the Annual Meeting of the Cognitive Science Society*

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.48550/arXiv.1805.04687

Zhao, C., & Chan, A. B. (2023, April 13). ODAM: Gradient-based instance-specific visual explanations for object detection. *The Eleventh International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2304.06354f