Do humans and CNN better understand the visual explanations generated by other humans or XAI algorithms?

Romy Müller[a], Marius Thoß[a], Julian Ullrich[a,b], Sascha Weber[a], Carsten Knoll[b], & Steffen Seitz[b]

[a] *Faculty of Psychology, Chair of Engineering Psychology and Applied Cognitive Research, TU Dresden*

[b] *Faculty of Electrical and Computer Engineering, Chair of Fundamentals of Electrical Engineering, TU Dresden*

Emails: romy.mueller@tu-dresden.de, marius.thoss@mailbox.tu-dresden.de, julian.ullrich@hhu.de, sascha.weber@tu-dresden.de, carsten.knoll@tu-dresden.de, steffen.seitz@tu-dresden.de

Visual explanations can shed light into the black box of deep learning. In the context of scene recognition, convolutional neural networks (CNN) use natural images as inputs to select a scene class. Subsequently, methods of explainable artificial intelligence (XAI) can analyze the weights of these networks and highlight areas in the input image that have contributed to the classification decision. These areas can be visualized as attention maps (e.g., heatmaps or binary masks overlaid on the original image). Such visual explanations are useful to cross-check the CNN, for instance by revealing whether it has actually looked at the dog to classify it as a husky, or at the snow in the background. But how understandable are these explanations? Are they more or less understandable than human-generated explanations? Does this depend on the type of image? And does it depend on whether you are a human or a CNN?

The literature allows for divergent predictions. On the one hand, it has been reported that humans can better classify images based on the areas selected by CNN than by humans (Zhang et al., 2019). On the other hand, it has been reported that CNN can better classify images based on the areas selected by humans than by CNN (Rong et al., 2021). Does this mean that humans can better understand CNN, while CNN can better understand humans? Certainly not, as the results of these two studies are hard to integrate due to a number of methodological differences between them. In the study by Zhang et al. (2019), humans saw images of complex natural scenes, split into 50 pre-defined segments, and their task was to manually order these segments according to their relevance for classification. The authors found that humans mainly selected the category-defining object itself, so that their attention maps lacked the rich information provided by the scene context. In consequence, these visual explanations were less understandable for other humans than those of CNN, which did consider the context. In the study by Rong et al. (2021), participants performed fine-grained classification of close-up images showing one of two highly similar bird species, and their eye movements were tracked passively. The authors found that human eye movements were more discriminative than CNN in targeting the feature that differentiated the birds. Thus, the two studies used different methods to elicit human attention maps on different images, which makes it impossible to draw firm conclusions. In the present contribution, we therefore varied three factors in one and the same study: the agent and procedure used for generating the attention map, the agent who had to understand the attention map, and the type of image that needed to be classified.

The task for both humans and CNN was to classify image snippets that represented attention maps on sixty scenes from six classes of the Places365 dataset (Zhou et al., 2017). These classes differed in whether the classification relied on singular salient objects (i.e., *objects: lighthouse, windmill*), arrangements of different objects (i.e., *indoor scenes: office, dining room*), or large sceneries spanning most of the image but not depending on particular objects (i.e., *landscapes: desert, wheat field*). The attention maps were generated in one of four ways: Human attention maps were elicited by (1) tracking human *eye movements* during scene classification of the original, full images, or (2) asking humans to *manually select* the image areas they consider most relevant for classification by drawing a

polygon around them. The maps represented an average of the data elicited from 25 participants. CNN attention maps were generated by visualizing the results of the two XAI methods (3) *Grad-CAM* (Selvaraju et al., 2017) and (4) *XRAI* (Kapishnikov et al., 2019). For each attention map, the most relevant 5% of the image were selected and turned into a binary mask, so that everything outside the most relevant area was hidden.

For the human classification task, a cue word was presented before each image, and participants had to decide as quickly and correctly as possible whether it matched the subsequently presented attention map. Their response times and error rates were analyzed with a 4 (*attention map type: human eye movements, human manual selections, Grad-CAM, XRAI*) x 3 (*image type: objects, indoor scenes, landscapes*) repeated measures ANOVA. For CNN classification, we use a ResNet-152 and analyzed top-5 accuracy as well as certainty for the actual class, comparing these metrics between attention map and image types.

The human results indicated that in general, humans were faster and more accurate when using human-generated attention maps than when using XAI maps, while the type of human map did not matter. However, these results strongly depended on image type. For objects, performance was much worse for Grad-CAM than all other types of maps. For indoor scenes and landscapes, it was XRAI that led to inferior performance, while Grad-CAM maps could be used as quickly and accurately as human maps. Thus, whether human attention is more informative than XAI attention depends of the XAI method and on the images this XAI has to deal with. However, it also depends on who you ask: CNN performance results indicated that the most informative type of attention map was human manual selections, followed by XRAI, while performance with human eye movements and Grad-CAM was much worse. This was true for objects, while for indoor scenes and landscapes, the snippets turned out to be too small to yield interpretable results (i.e., accuracy was at chance). Therefore, in follow-up analyses, we also tested how the CNN results changed with increasing size of the attention maps.

Our findings have theoretical as well as practical implications. They contribute to our understanding of the differences between human and artificial information processing, but also inform the evaluation of XAI methods. In the past, this evaluation often neglected how understandable these maps were for humans, and whether this goes hand in hand with their fidelity to the CNN model. Our results suggest that this may not always be the case.

Keywords: image classification, Convolutional Neural Networks (CNN), explainable artificial intelligence (XAI), attention/saliency maps, eye movements, similarity between humans and CNN

References

Kapishnikov, A., Bolukbasi, T., Viégas, F., & Terry, M. (2019). XRAI: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4948-4957).

Rong, Y., Xu, W., Akata, Z., & Kasneci, E. (2021). Human attention in fine-grained classification. *arXiv preprint arXiv:2111.01628*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision* (pp. 618-626), Venice, Italy: IEEE.

Zhang, Z., Singh, J., Gadiraju, U., & Anand, A. (2019). Dissonance between human and machine understanding. In *Proceedings of the ACM on Human Computer Interaction* (pp. 1-23), New York, NY: Association for Computing Machinery.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Mchine Itelligence, 40*(6), 1452-1464.