Give Users What They Want: Labeled Arrows

Severine Soltani (ssoltani@ucsd.edu)
University of California, San Diego, Department of Bioengineering

Michael Pazzani (mpazzani@ucsd.edu)
University of California, San Diego, Halicioglu Data Science Institute

Introduction

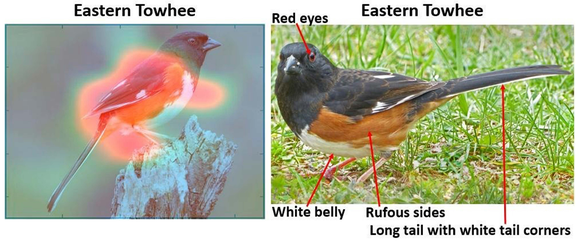
Users want insight into how an algorithm makes its classification and rely on explainable artificial intelligence (XAI) methods to do so. XAI methods using saliency maps or heatmaps highlight the importance of a region in a classification, such as the popular GradCAM[1].

We argue that the explanations provided by heatmaps are incongruent with the explanations that users expect and experts give. We propose that users may prefer and benefit from explanations that are closer to what experts produce, such as arrows with labels, rather than heatmaps.

Study 1 Methods: Experts

21 expert bird watchers were asked the utility of 5 types of image explanations for bird classification:

- Bounding boxes
- Hand-drawn arrows pointing at relevant features
- Hand-drawn arrows with labels
- Grad-CAM generated heatmaps
- Verbal descriptions



An example of heatmap and labeled arrow annotations.

Study 2 Methods: Novices

336 novices learned to classify similar-looking birds with feedback until proficiency (9 out of 10 correct in a running window).

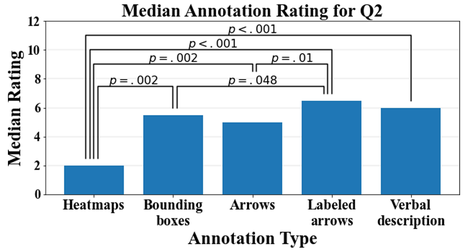
The experimental groups were shown bird images with annotations that emphasize areas of the bird important to its classification; the control group did not receive any annotations.

Explanation types (all compared to not having any explanations) evaluated include:

- Best-case scenario hand-drawn heatmaps
- Bounding boxes
- Verbal descriptions
- Labeled arrows

Study 1 Results: Experts

Pairwise Mann-Whitney U test on the median ratings for all combinations of explanation types for "I would recommend using this explanation to help identify this bird" (Q2):

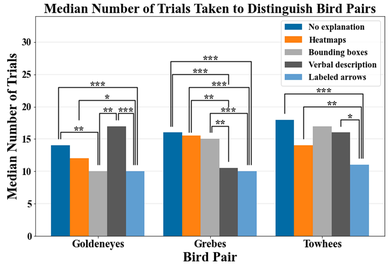


The Bonferroni-corrected *p* values are statistically significant as are the difference in median scores between the heatmaps and all other explanation types.

Labeled arrows are strongly preferred to heatmaps.

Study 2 Results: Novices

Pairwise Mann-Whitney U test for all combinations of explanation types within each bird pair:



The Bonferroni-corrected *p* values are statistically significant for the median number of trials until classification proficiency. Significance levels of $p < .05$, $p < .01$, and $p < .001$ are denoted by *, **, and ***, respectively.

Labeled arrows as feedback results in the lowest (or tied for lowest) number of trials until classification proficiency.

References

[1] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 618–626, Oct. 2017. ISSN: 2380-7504.

Acknowledgments

This work was supported with funding from the DARPA Explainable AI Program under a contract from NRL and from NSF grant 2026809. This work was also supported by a grant from the National Institutes of Health, USA (NIH grant T32GM008806).