User-Centric Enhancements to Explainable AI Algorithms for Image Classification

Severine Soltani (ssoltani@ucsd.edu)
University of California, San Diego, Department of Bioengineering, San Diego, CA 92093 USA

Robert Kaufman (rokaufma@ucsd.edu)
University of California, San Diego, Department of Cognitive Science, San Diego, CA 92093 USA

Michael Pazzani (mpazzani@ucsd.edu)
University of California, San Diego, Halıcıoğlu Data Science Institute, San Diego, CA 92093 USA

Abstract

The introduction of deep learning and CNNs to image recognition problems has led to state-of-the-art classification accuracy. However, CNNs exacerbate the issue of algorithm explainability due to deep learning's black box nature. Numerous explainable AI (XAI) algorithms have been developed that provide developers insight into the operations of deep learning. We aim to make XAI explanations more user-centric by introducing modifications to existing XAI algorithms based on cognitive theory. The goal of this research is to yield intuitive XAI explanations that more closely resemble explanations given by experts in the domain of bird watching. Using an existing base XAI algorithm, we conducted two user studies with expert bird watchers and found that our novel averaged and contrasting XAI algorithms are significantly preferred over the base XAI algorithm for bird identification.

Keywords: Explainable AI; Explanation; Image Classification

Introduction

Significant progress in image classification using deep learning (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015) has created considerable interest in applying these techniques to a broad range of applications including medical diagnosis and classification of animals. Furthermore, a variety of eXplainable Artificial Intelligence (XAI) methods have emerged for explaining image classification to developers or end users (Lapuschkin et al., 2016; Chattopadhyay et al., 2018). These approaches typically highlight regions of the image that are important to the classification decision.

As many researchers turned their attention to XAI, Miller et al. (2017) argued

> While the re-emergence of explainable AI is positive, this paper argues most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users. But explainable AI is more likely to succeed if researchers and practitioners understand, adopt, implement, and improve models from the vast and valuable bodies of research in philosophy, psychology, and cognitive science; and if evaluation of these models is focused more on people than on technology.

In spite of this, most XAI papers, even those with user studies, focus on developer-centric explanations. Langley (2021) draws a distinction between these two purposes for explanations. Here, we explore how cognitively motivated modifications to an existing XAI algorithm can result in increased acceptance to end-users.

Our motivation for exploring principled modifications that might apply to any XAI is twofold. First, we seek to test how changes to existing explainability algorithms based on known human explanation and reasoning practices can improve end-user experiences. Second, we seek to push against the methodological trend where highly dissimilar explainability types are compared and instead focus on the principles behind how a cognitively-motivated change to any existing explanation approach impacts the quality of explanation. This can bring an understanding of how specific properties or techniques may improve visual explanations for image classification more generally.

Deep Learning and XAI

Within the domain of deep learning, convolutional neural networks (CNNs) are specialized for image recognition. CNNs are able to capture complex, non-linear relationships present in images. The basic CNN architecture consists of successive pairs of convolutional and pooling layers: convolutional layers apply a filter to the image to yield feature maps. These feature maps then pass through an activation function, which serves to introduce non-linearity to the network. The output image is then passed through the pooling layer to reduce the dimensionality of the image by downsampling pixels in close proximity to each other. One may serially add pairs of convolutional and pooling layers to their desire before flattening the last output to feed into a fully connected layer, which classifies the image using an activation function. The network may be further trained by using the back-propagation algorithm where the weights of the network may be fine-tuned for greater classification accuracy (Kim, 2017). Convolutional neural network architectures contribute to the obscured, black box nature of deep learning algorithms, however, thus limiting their value.

There have been several general approaches to gaining insight into why deep learning algorithms arrive at a certain classification through XAI:

Increasing Transparency of Opaque Methods Research in this field identifies regions of importance for the classification of images. Some approaches, such as GradCAM

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Figure 1: Heatmaps from GradCAM on the same image from two deep nets. It can be observed that entirely disparate areas of the bird are emphasized as important areas for classification.

(Selvaraju et al., 2017) analyze the activation of hidden units to visualizing the regions of input that are "important" for predictions from these models. Others, such as LRP (Binder et al., 2016) propagate relevance information from the output layer of the neural network to input layer to find the regions most responsible for activating the output.

Model Agnostic (Black Box Methods) Instead of exposing the logic of an underlying model, model agnostic methods make changes to the image and examine the changes to the decisions made by any "black box" algorithm. Black box XAI algorithms include occlusion sensitivity (Zeiler & Fergus, 2014), and Local Interpretable Model-Agnostic Explanations, known as LIME (Ribeiro, Singh, & Guestrin, 2016).

Our observation from applying several XAI algorithms and several architectures revealed some common problems: far too often, the XAI algorithms missed some features that bird guides considered important in distinguishing one bird from another. This may even vary among different runs with different random weights: e.g., sometimes the XAI algorithm will focus on the wing area, other times the head, and sometimes both. Figure 1 illustrates an example of this problem using GradCAM. In addition, the XAI algorithms occasionally highlighted areas of the bird image that bird guides do not consider particularly important. Even so, these highlightings may give insight into how the XAI algorithm operates, which lends itself to a developer-centric focus for the algorithm.

We propose that defects of XAI algorithms can be alleviated through leveraging cognitive theories related to what makes for a satisfactory explanation and applying them to XAI algorithms. By addressing the pitfall that XAI algorithms often produce explanations that are not well-aligned with human experts, we aim to pivot away from the established developer-centric focus of XAI and towards a user-centric focus.

XAI and Cognition

Significant work in the fields of cognitive science, psychology, and philosophy have contributed to theories of explanation (Miller, 2019; Hoffman et al., 2018). Because the studies in this paper focus on the task of classifying birds by species, we start with a short discussion on classification. Murphy (2004) goes into much more details on these issues. We discuss two issues here: contrasting explanations and discriminative classification.

One of our studies and algorithm modifications is motivated by the importance of drawing contrasts. In the philosophy literature, Lipton (1990) differentiates the *fact* (which did occur) from the hypothesized event which did not occur, called the *foil*. Miller (2019, 2021) summarizes the cognitive literature on the utility of contrasts in explanation. In Artificial Intelligence, Pazzani et al. (2018) and Le et al. (2020) have argued for explainable AI algorithms that are sensitive to contrasting categories. Contrasts occur in tasks such as differential diagnosis or distinguishing between easily confused animal species. In these cases, explanations tend to focus on features that differentiate the true fact class from the related-but-false foil class.

The second study focuses on whether it is better to focus on one way to discriminate between categories or to include all features with high cue validity (Rosch & Mervis, 1975). In discrimination learning, a single way of separating one class from others is found (Medin, 1975). Neural nets are examples of discrimination learning. A single network will overlook features with high cue validity if they are not necessary to distinguish an example in one category from examples in other categories. Furthermore, a single network may find an irrelevant feature that is not shared with other networks. We investigate whether by combining the results from an ensemble of networks, we may mitigate these problems.

Several recent studies have focused on end-user evaluation of XAIs as well. Gunning et al. (2021) summarize a variety of studies associated with DARPA's XAI program. Numerous studies showed that users prefer systems that provide decisions with explanations over systems that provide only decisions. However, since many alternative explanation strategies have been compared to the strategy of having no explanation, these studies yield no insight into the nature of explanation. Other studies (Muddamsetty et al., 2021; Petsiuk, Das, & Saenko, 2018) compare different algorithms, but the differences are not intended to show why users prefer one algorithm to another.

An interesting study by Nourani et al. (2019) on images with manually created highlights showed that users could distinguish between highlighting on relevant and irrelevant parts of the image. Relevant regions for identifying dog species included the eyes and ears while the trees in the background are irrelevant. XAI systems that emphasized less relevant areas were judged to be less accurate than they actually were.

Here, we explore two minor variations of an existing XAI algorithm. In one variation, the algorithm is changed to emphasize regions that distinguish one class from another closely-related (contrasting) class. In the other variation, we use an ensemble of several neural nets to classify, and we combine the explanation produced by each network. This allows features with high cue validity to be identified even if there are alternative ways to discriminate categories that do not include them.

The XAI explanations tested in this work were motivated by cognitive theories, such as cue validity and work on contrastive explanations. We hypothesize these explanations will be preferred by end-users of a system due to their propensity to generate more human-meaningful information.

User Studies of XAI for Image Classification

In this paper, we aim to address the shortcomings of XAI algorithms by exploring slight modifications to an existing XAI algorithm and how the modifications impact preference for the novel algorithm. We used with the MATLAB implementations of VGG16 for image classification (Simonyan & Zisserman, 2015) and ImageLIME for XAI (*imageLIME*, 2022). ImageLime with the default superpixel algorithm and did not experiment with alternative superpixel approaches (Schallner et al., 2019). We consider two enhancements to using VGG16 and LIME: averaging explanations over multiple network models and explaining contrasting categories—henceforth referred to as averaged LIME and contrasting LIME.

We conducted two user studies to determine the preference for standard LIME-generated explanations compared to explanations generated from averaged LIME and contrasting LIME. We solicited the opinion of bird watchers on which annotation methods produced the best heatmaps for highlighting areas of birds that are pertinent to identification. In order to gauge the generalizability of the heatmaps beyond tools for experts, we additionally asked bird watchers if they would recommend the heatmap to others for identification of the bird species. The bird watchers were also asked to classify the bird species to see if the aid of either novel explanation method resulted in increased classification accuracy.

Methods

Heatmap Algorithms

We consider two changes to how LIME is used with VGG16. Although our techniques are general and work with any technique for heatmaps, we selected LIME in our studies because it finds regions with coarse quantum differences in importance as opposed to a continuous gradient of importance. These quantum areas often correspond to regions of the bird such as the bill or an area of the wing.

Averaging Explanations Over Multiple Network Models Ensemble learning (Sagi & Rokach, 2018) has been shown to increase the accuracy of prediction by averaging over multiple models. Here, we focus on using ensembles to improve the explanation. Instead of a single run of VGG16, we run VGG16 several times, compute the relevance of each pixel of each run separately, and then average over several runs. The motivation for combining explanations is that although each run finds an accurate classifier, they may find different local minima, each finding a sufficient way of making the classification. By combining across runs, we may find important but redundant features. Furthermore, irrelevant features that are seldom present have lower weights in the combined model and are often de-emphasized entirely. In our study, we take the average of 11 networks, but we have seen similar results for averaging of 7 through 15 networks. Figure 2 illustrates how the averaged explanation is computed.

Explaining Contrasting Categories The goal behind this approach is to identify regions that distinguish one class from its most similar classes (Pazzani et al., 2018). That is, instead of answering why is this an *X*, they are designed to answer why is this an *X* and not a *Y*. In medical systems, these are analogous to methods that attempt to find regions important for differential diagnosis (e.g., to distinguish bacterial from viral pneumonia). We use a method suggested by (Feghahati et al., 2020) that looks for the regions that are suggestive of belonging to category *X* and features that are suggestive of belonging to category *Y* and finds the difference. This is computed on a pixel-by-pixel basis on the heatmaps. There are several ways to identify the foil category *Y*. One is to identify the contrasting class as the one with the second-highest activation of the neural network. Alternatively, such as in this study, the AI may be given the second most likely class. An example of contrastive explanations is such: an image *A* may have the ground-truth label of "western grebe". Image *B* of a "Clark's grebe" may be the second most likely class label for image *A*. Thus, the difference between the heatmap for "western grebe" and the heatmap for "Clark's grebe" images are computed pixel-wise to yield a contrastive explanation for why image *A* belongs to the "western grebe" class. This method is applicable to any algorithm that generates heatmaps, but we use imageLime in our studies. It should be noted that this method can be used at multiple levels of a hierarchy (e.g., explaining why a bird is a western vs. Clark's grebe or more generally a grebe vs. a heron). Figure 3 depicts an example of how contrasting explanations are generated.

Data Selection

We collected a database of publicly available images from Flickr of 66 bird species for a total of 14,380 images. Due to the number and quality of images, most deep learning algorithms achieve less than 2% error on classification, allowing one to study the explanation of accurate classifiers. Due to the accuracy of the classifier and large number of high-quality examples per class, we did not observe the XAI algorithm focusing on totally irrelevant areas of the image such as water or trees.

Participant Recruitment

Participants for both studies were expert bird watchers who were recruited from mailing lists that report rare bird sightings in Southern California. 28 participants were recruited for the LIME vs. averaged LIME study, and 20 participants were recruited for the LIME vs. contrasting LIME study. One participant from each of the LIME vs. averaged LIME and LIME vs. contrasting LIME studies self-excluded due to lack of familiarity with the bird species included in the studies. One participant was excluded from the LIME vs. averaged LIME study for being under the age of 18, which may point

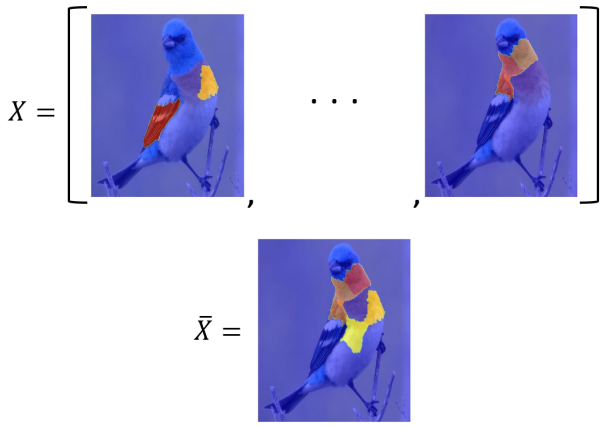

Figure 2: An illustration of how averaged explanations are achieved. $\bar{X}$ contains the output from n individual runs of standard LIME (n is 11 in the case of our studies), and $\bar{X}$ represents the pixel-wise mean of the n images in X. Note that the individual runs in X are likely to contain fewer areas of relevance than the aggregated, averaged final $\bar{X}$ since the output of a single run of LIME is non-deterministic.

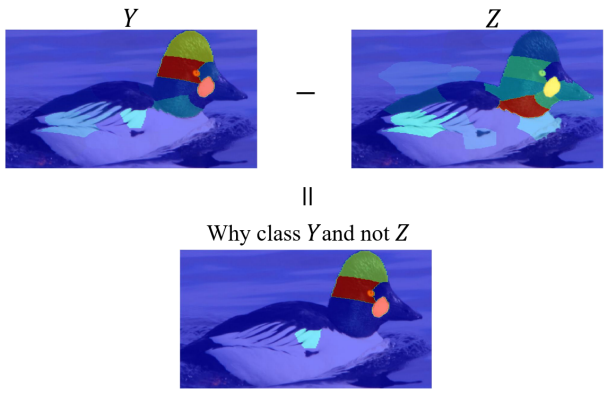

Figure 3: An illustration of how contrasting explanations are achieved. Image Y represents the evidence for the bird being a "common goldeneye" whereas Z represents the evidence for it being a "Barrow's goldeneye." Pixel-wise subtraction of Z from Y yields a contrasting explanation that depicts why this bird is indeed a "common goldeneye" rather than a "Barrow's goldeneye."

to less real-world bird watching experience. Excluding the data from these participants did not alter the significance of the results. In total, 26 and 19 participants were included in analyses for the LIME vs. averaged LIME and LIME vs. contrasting LIME studies, respectively. There was no participant overlap between these studies. The median participant age was 46 and 43 for the LIME vs. averaged LIME and LIME vs. contrasting LIME studies, respectively. Participants in the LIME vs. averaged LIME study had a median of 15 years of bird watching experience while the participants in the LIME vs. contrasting LIME study had a median of 13 years of bird watching experience.

Study Design

Prior to beginning their respective studies, participants from both studies were shown an example of LIME used on an image of a dog to identify the features most important to classifying its breed. This example was intended to familiarize participants with how to interpret the colors on a LIME-generated heatmap.

Each study contained 24 unique bird images. Each of the 24 unique bird images was shown once to each subject with LIME-generated annotations and another time with averaged LIME-generated or contrasting LIME-generated annotations, depending on the study in question. This results in a total of 48 trials evenly split between LIME and averaged or contrasting LIME for each study. Furthermore, the 24 unique bird images were comprised of 10 bird species, and these 10 bird species were chosen such that they were composed of 5 commonly confused pairs: the Barrow's goldeneye and common goldeneye; black-headed grosbeak and blue grosbeak; Clark's Grebe and western grebe; eastern towhee and spotted towhee; indigo bunting and lazuli bunting.

During the study, each trial displayed the heatmap-annotated and unannotated images side-by-side on the left-hand side of the screen; a bird species classification task and questions about annotation preferences were displayed on the right-hand side of the screen. A screen capture of the study interface is shown in Figure 4. Participants were asked to identify the species of the bird shown on the left-hand side by checking 1 of 10 radio buttons, each option being a unique bird species present in the study. Participants were asked to indicate their opinions on the novel highlighting method on the left-hand side by answering two questions: "This highlighting emphasizes the areas of the bird that I think are important for identification" (Question 1), and "I would recommend using this highlighting to help identify this bird" (Question 2). Participants were asked to respond using a 7-point Likert scale ranging from "Strongly Disagree" (a value of 1) to "Strongly Agree" (a value of 7) with "Neutral" (a value of 4) at the midpoint.

Results

For each of the two studies, we compared the median preference ratings for LIME trials to averaged or contrasting LIME

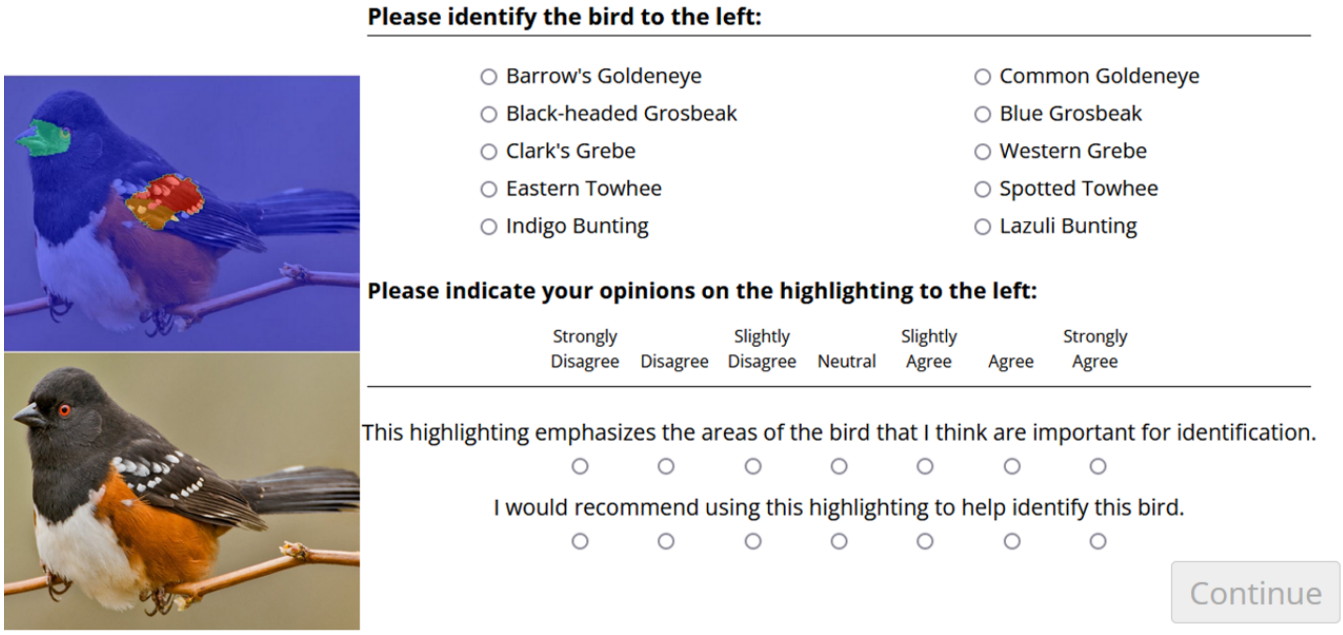

Figure 4: An example screen capture from the LIME vs. contrasting LIME study. Images were placed side-by-side for ease of comparison.

trials. Additionally, we compared overall bird species classification accuracy for LIME trials to averaged or contrasting LIME trials. Because accuracy may be a dubious measure of whether one is qualified to advise on useful annotations for bird identification, we did not use classification accuracy to exclude individual trials or participants. An experienced bird watcher may be aware that the shape of a white patch on the face of a common goldeneye or Barrow's goldeneye is a reliable method of distinguishing the two species. While a bird watcher may fail to accurately assign the round patch to the common goldeneye and the crescent patch to the Barrow's goldeneye (i.e. failure of the classification task), they remain knowledgeable on which areas of a bird are crucial for classification. Additionally, excluding trials with incorrect classifications did not result in a significant difference in the distribution of median ratings for any question; this is true for both studies (uncorrected p values of .7 or greater). This may be due to the consistently high classification accuracy across all participants.

Thus, all participants and their data were used for the subsequent analyses. All reported p values are Bonferroni-corrected for 3 pairwise Wilcoxon signed-rank tests in each study. Broadly, subjects exhibit a significant preference for averaged or contrasting LIME over standard LIME. The p values and median ratings for the two questions regarding preferences can be found in Table 1 and Table 2, respectively.

The preference for annotations output by averaged LIME was significantly higher than the preference for annotations output by LIME for both questions in the study ($p < .001$). The median rating of Question 1 for the LIME images was 4.0 ("Neutral") while the median rating for the averaged LIME images was 5.5 ("Slightly Agree"/ "Agree"). The median rating of Question 2 for the LIME images was 3.0 ("Slightly Disagree") while the median rating for the averaged LIME images was 5.0 ("Slightly Agree"). Bird species classification accuracy, with a mean of 90.1% and 91.3% for LIME and averaged LIME, respectively, did not differ between these two explanation methods ($p > .99$).

Likewise, the preference for annotations output by contrasting LIME was significantly higher than the preference for annotations output by LIME for both Question 1 ($p = .014$) and Question 2 ($p = .004$). The median rating of Question 1 for the LIME images was 4.5 ("Neutral"/ "Slightly Agree") while the median rating for the contrasting LIME images was 5.0 ("Slightly Agree"). The median rating of Question 2 for the LIME images was 4.0 ("Neutral") while the median rating for the contrasting LIME images was 5.0 ("Agree"). Bird species classification accuracy was again strikingly similar between LIME and contrasting LIME trials with respective mean classification accuracies of 84.9% and 86.0% ($p > .99$).

Discussion

The results of our two user studies showed that expert bird watchers prefer averaged LIME and contrasting LIME explanations significantly more than standard LIME explanations using two measures. Specifically, they report that the modified explanation techniques emphasized the areas they believed to be important for the identification of the bird in question and are more likely to recommend the novel highlighting technique to others as a tool for bird identification.

Table 1: Bonferroni-corrected p values[a] for median ratings in LIME vs. averaged LIME and LIME vs. contrasting LIME studies.

	LIME vs. Averaged LIME (N = 26)	LIME vs. Contrasting LIME (N = 19)
Question 1[b]	$< .001$	.014
Question 2[c]	$< .001$	.004

[a] 3 pairwise comparisons were made for each study using the Wilcoxon signed-rank test.
[b] "This highlighting emphasizes the areas of the bird that I think are important for identification."
[c] "I would recommend using this highlighting to help identify this bird."

Table 2: Median responses on 7-point Likert scale (1 = "Strongly Disagree" and 7 = "Strongly Agree") for LIME vs. averaged LIME and LIME vs. contrasting LIME studies.

	LIME vs. Averaged LIME (N = 26)		LIME vs. Contrasting LIME (N = 19)	
	Lime	Averaged LIME	LIME	Contrasting LIME
Question 1[a]	4.0	5.5	4.5	5.0
Question 2[b]	3.0	5.0	4.0	5.0

[a] "This highlighting emphasizes the areas of the bird that I think are important for identification."
[b] "I would recommend using this highlighting to help identify this bird."

We believe that averaged LIME was preferred over standard LIME for two reasons. At the most basic level, averaging across several runs of LIME means that features that may be present in a single run but truly irrelevant to classification will end up having a lower weight overall. The result is that attention is drawn away from these less relevant features and focused more on the features whose relevance is prominent across more runs. Further, as the algorithm is averaging across more runs, alternative features that are found in perhaps half of the runs but not the other half will be caught. These techniques may be particularly helpful if there are more than one distinguishing feature of a bird in question. Though this redundancy may result in the highlighting of more features than is truly necessary for a given classification, the end result is a cleaner explanation. Highlighting all of the relevant attributes helps users clearly attend to features which are important for consideration in category membership.

We note that these can also be related to Grice's conversational maxims of *quantity*—sharing the right amount of information, no more and no less—and *relation*—keeping communications relevant—as they establish the precedent that pixel-based visual explanations such as those produced using LIME should communicate sufficient information while eliminating anything unnecessary (Grice, 1975). In our case, the cleaner image prominently shows relevant features and has given less weight to features that are less important to classification.

In a similar manner, we believe that contrasting LIME was preferred over standard LIME in that it directs attention to particular features of the bird which may be singularly important or unique to the identification of that bird compared to the other birds in question. For example, by subtracting the features which belong to a "western grebe" (the fact) from those of a "Clark's grebe" (the foil), we are left with the most distinguishing features of a western grebe. By drawing a user's attention to these important features, it streamlines the process of differentiating between the two and allows the user to more easily draw their attention to what is relevant in that determination. The resulting elimination of foils is more efficient and thus less cognitively demanding, and should theoretically lead to increased understanding of features differentiating category membership.

It is worth noting that although we used explainable AI methods in our study, we asked subjects questions that did not use the word "explanation." Instead, we asked how well the "highlighting emphasizes the areas of the bird that I think are important for identification" and whether they "would recommend using this highlighting to help identify this bird." We do this because current XAI methods identify regions of interest, but they do not indicate why they are interesting. We believe a better explanation would not only identify but also describe the regions of interest with terms such as "spotted wing" or "sharp talons." Many books on bird identification both point out distinctive regions with arrows or circles, and also label why they are distinctive.

Conclusion

Our study focused on automatically manipulating visual explanations to have averaged or contrasting properties and seeing whether experts would recommend using this highlighting to help others. Our results reveal that averaging across runs of LIME and utilizing contrasting explanations makes a significant difference in experts' willingness to use and recommend the explanation as opposed to the standard presentation of LIME. At present, the results of our two user studies have implications for both practical applications of XAI and on future methodological directions for XAI explanation research. First, we contend that averaging and contrasting explanation techniques are viable improvements for XAI explanations that use heatmaps. Further, our results lend credence to the notion that methods should be more fine-grained, focus on theory-driven techniques, and incisive in their ability to identify which factors contribute to increased preference for XAI explanations.

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 the NIH Graduate Training Program in Bioinformatics (T32GM008806). We thank Dorrit Billman and Pat Langley for their constructive feedback on our drafts.

References

Binder, A., Montavon, G., Bach, S., Müller, K.-R., & Samek, W. (2016). Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *arXiv:1604.00825 [cs]*. (arXiv: 1604.00825)

Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847. (arXiv: 1710.11063) doi: 10.1109/WACV.2018.00097

Feghahati, A., Shelton, C. R., Pazzani, M. J., & Tang, K. (2020). CDeepEx: Contrastive Deep Explanations. *ECAI 2020*, 1143–1151. (Publisher: IOS Press) doi: 10.3233/FAIA200212

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2*(4), e61. doi: 10.1002/ail2.61

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining explanation, part 4: a deep dive on deep nets. *IEEE Intelligent Systems*, *33*(3), 87–95.

imageLIME. (2022). Retrieved 2022-01-15, from `https://www.mathworks.com/help/deeplearning/ref/imagelime.html`

Kim, P. (2017). Convolutional Neural Network. In P. Kim (Ed.), *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence* (pp. 121–147). Berkeley, CA: Apress. doi: 10.1007/978-1-4842-2845-6

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc.

Langley, P. (2021). Explanation in cognitive systems. *IEEE Intelligent Systems*, *9*, 3–12.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., & Samek, W. (2016). The LRP Toolbox for Artificial Neural Networks. *Journal of Machine Learning Research*, *17*(114), 1–5.

Le, T., Wang, S., & Lee, D. (2020). GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 238–248). New York, NY, USA: Association for Computing Machinery.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi: 10.1038/nature14539

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247–266.

Medin, D. L. (1975). A Theory of Context in Discrimination Learning. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 9, pp. 263–314). Academic Press. doi: 10.1016/S0079-7421(08)60273-X

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Miller, T. (2021). Contrastive explanation: A structural-model approach. *Knowledge Engineering Review*, *36*, E14. doi: https://doi.org/10.1017/S0269888921000102

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum. In *Ijcai 2017 workshop on explainable artificial intelligence (xai)*.

Muddamsetty, S. M., Jahromi, M. N. S., Ciontos, A. E., Fenoy, L. M., & Moeslund, T. B. (2021). Introducing and assessing the explainable AI (XAI)method: SIDU. *arXiv:2101.10710 [cs]*. (arXiv: 2101.10710)

Murphy, G. (2004). *The big book of concepts*. MIT press.

Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 97–105.

Pazzani, M. J., Feghahati, A., Shelton, C. R., & Seitz, A. R. (2018). Explaining contrasting categories. In *Workshop on explainable smart systems (exss), acm intelligent user interface conference.*

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. *arXiv:1806.07421 [cs]*. (arXiv: 1806.07421)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*. (arXiv: 1602.04938)

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*.

Schallner, L., Rabold, J., Scholz, O., & Schmid, U. (2019). Effect of superpixel aggregation on explanations in lime–a case study with biological data. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 147–158).

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 *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). (ISSN: 2380-7504) doi: 10.1109/ICCV.2017.74

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. (arXiv: 1409.1556)

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 818–833). Cham: Springer International Publishing. doi: 10.1007/978-3-319-10590-1_53