Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Approach

Sina Mohseni
Department of Computer Science & Engineering
Texas A&M University
sina.mohseni@tamu.edu

Eric D. Ragan
Department of Computer & Information Science & Engineering
University of Florida
eragan@ufl.edu

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Research in interpretable machine learning proposes different approaches to evaluate model explanations. Our work contributes to evaluating the human interpretability of machine learning explanations. We present an evaluation benchmark for instance explanations from text and image classifiers. We create our benchmark using multi-level human attention heatmaps drawn from multiple human annotators. We review the benchmark, demonstrate its utility for quantitative evaluation of model explanations, and discuss the future work for this research.

Author Keywords
interpretable machine learning; Human subject evaluation; local explanations; Human computer interaction.

Introduction
With the recent and continuing advancements in robust deep neural networks, the prominence of artificial intelligence models is growing for automated decision-making. In such cases, human experts, operators, and decision-makers can also take advantage of advanced machine learning techniques to assist in taking real-world actions. However, these people need to be able to understand and trust machine learning model predictions. Thus, for more effective Human-AI collaboration, advancements are needed in achieving explainability and supporting human under-

standing. This is the primary goal of recent interdisciplinary research thrusts in *explainable artificial intelligence* (XAI). While a multi-faceted topic, the ultimate goal is for people to can understand machine models, and it is therefore essential to involve human feedback and reasoning as a requisite component for evaluating the explainability or understandability of XAI methods and models.

Different approaches have been proposed to evaluate model interpretability and XAI systems at different system design steps [7]. In machine learning research, various computational methods are used to measure fidelity of interpretability techniques to the black-box model [1]. On the other hand, in the field of HCI, human-grounded evaluation studies measure human factors such as user satisfaction, mental model, and trust. Human-AI task performance is also a quantitative measure to the effectiveness of interpretability as an application grounded measure [3] Alongside human-subject studies to evaluate interpretability with human feedback, the human annotation of data (e.g., object segmentation) is also used as a quantitative measure for the quality of explanations. For example, Du et al. [4] used object localization metrics to evaluate saliency maps as a weakly supervised object localization tasks. However, it is not clear that what is the relation between the two human-grounded evaluation methods, being *1) Human review of explanations and feedback* and *2) Human annotation of data as the baseline*, for evaluation purposes.

In this paper, we present a human-grounded evaluation benchmark for evaluating instance explanations from images and text classification model. The benchmark consists of human-annotated samples of images and text documents to approximate the most important regions for human understanding and recognition. Unlike image segmentation datasets, our benchmark provides multi-level heatmap of human attention on image regions and words in documents. Our benchmark allows the quantitative evaluation of instance explanations as a model trustworthiness baseline as well as a baseline for comparison between multiple interpretability techniques. We aim to contribute to the XAI research by studying the relation between the two aforementioned human-grounded evaluation methods being 1) human review and feedback and 2) human annotation baseline. We made this benchmark publicly available online[1] for research purposes.

Human-Grounded Evaluation

We review two main classes of approaches for human-grounded evaluation of interpretability, with the difference depending on whether users have prior knowledge or access to the model explanations itself. In one way, users review existing explanations and provide subjective feedback for those explanations. The other way is to capture users' thoughts and opinions of salient features on input based on the targeted output. Although human explanations could be in any form such as descriptive verbal or salient features explanation, we limit our scope to salient features explanation on image and text data. The following subsections provide further details for the two evaluation types:

1) Evaluating with Explanation Review and Feedback
For the purposes of evaluating existing known explanations, it is possible to collect user feedback about the quality of the explanation given the original input, model explanation, and the targeted output. The user feedback could be in form of subjective rating or correcting the model generated explanation. In this case, quantifying the difference (e.g., IoU) between the model explanation and the user-edited explanation could give a precise measure of quality for the

[1] https://github.com/SinaMohseni/ML-Interpretability-Evaluation-Benchmark

(a)

The concrete simply sucks all the electrons ou
them into the ground.
Another explanation, implausible as it is, is th
needs to be periodically charged (topped-off),
self-discharges and then undergoes irreversib

(b)

Figure 1: Examples of human annotations of salient features for image and text samples. (a) Heat map views from 10 users for drawing contours around the area which explains the object in the image. (b) Heat map view from two expert reviewers for highlighting words related to the document topic (“electronic”).

model explanation. A different advantage of user feedback evaluation is the ability for a clear comparison of explanations from multiple interpretability techniques. For example, users could review several options to choose the best machine-generated explanation and provide justifications for their choices. Another means of capturing user feedback would be letting a user interactively refine model generated explanations for active learning [8]. This method has more flexibility in allowing the rejection of wrong features and adding new features for retraining.

However, multiple external and internal factors could affect users’ subjective ratings in this method. Examples of external factors include users’ prior-knowledge and expectation of model performance. Internal factors include user mental model and trust in the model that could potentially change over time. Also, user feedback might not be re-usable for active feature learning as the retrained models generate different explanations and require new human review.

2) Evaluating with Human Annotation

Another approach for human-grounded evaluation of explanations is to collect human feedback by annotating the salient features that would best contribute towards explanations for the given output. For example, if the data is a text article about a “computer science” topic, the user would find and annotate words and phrases related to the topic. However, user choice is made with knowledge about the input along with the output label. In this case, increasing the number of users results in capturing a wide spectrum of user explanations on each input. In this method, explanations are weighted features from multiple users’ annotations. Figure 1 presents examples of text and image heatmaps generated by multiple users annotations. Additionally, recent related research propose using segmentation masks (pixel-level human label) to improve model representation learning and hence prediction performance. For instance, Li et al. [6] present GAIN, a method to use human annotation of objects to improve the training process in weakly supervised object localization task. As the user feedback would be independent of any particular explanation, the human annotation in this method could be re-usable for evaluation of explanation as well as training for the same sample.

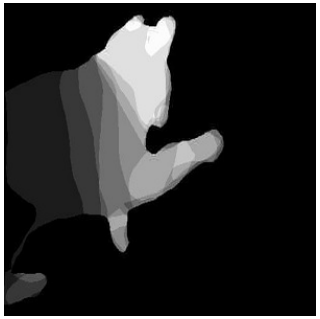
On the other hand, the disadvantage of this method is the cost of multi-level annotation of input samples while maintaining the annotation quality. However, this is yet unclear that the objectivity of human-annotation evaluation methods excels subjective human review and feedback.

Evaluation Benchmark

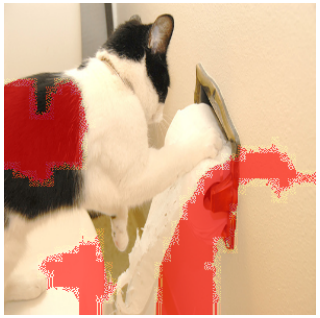
In order to create a human-grounded benchmark for model explanations evaluation, we captured human annotation of salient features where participants were asked to select relevant regions in images and phrases in text documents that are representing the topic or subject. The preliminary deployment of this benchmark consists of a subset of 100 sample images and text articles from the well-known *ImageNet* [2] and *20 Newsgroup* [5] data sets.

Annotated Image Examples

All image samples were collected from the *ImageNet* data set from 20 general categories (example categories include animals, plants, humans, indoor objects, and outdoor objects). Our preliminary benchmark includes 5 images per category for a total of 100 images. In an IRB approved participation study, 10 participants viewed images on a tablet and used a stylus to annotate key regions of the image. We asked them to draw a contour in the image around the area most important to recognize the object or the portion that, if removed, you could not recognize the object. None of the



(a)



(b)

Figure 2: (a) User generated weighted-mask for an example from ImageNet. We use this weighted-mask to evaluate model explanations accuracy. (b) Model explanation using LIME [8] algorithm for the same image. Irrelevant red-highlighted regions in this image cause low explanation score in comparison to the human-grounded annotation.

participants were experts in any of the image categories. Each participant annotated all images in a random ordering.

All participants' annotations are accumulated to create a weighted explanation mask over the image. Figure 1-a shows a heatmap view of participants' annotated explanations over a sample image, where "hot" colors (red) shows more commonly highlighted regions, and "cooler" colors (blue) show areas that were highlighted less frequently. We also masked all participants' annotations with exact contour shapes to reduce the impact of participants' imprecision or hand jitter.

Annotated Text Examples

All text documents were collected from the *20 Newsgroup* data set in medical (*sci.med*) and electronic (*sci.elect*) categories. For each category, two expert reviewers highlighted the most important words relevant to the given topic (i.e., medical or electronic). Reviewers were instructed to highlight words which, if removed, you could not recognize the main topic of the article. Two electrical engineers and two physicians volunteered as experts to annotate 100 documents from each topic. Each expert annotated the documents individually and in random order. Figure 1-b shows a single tone heat map view of user annotated explanations over a partial sample text article.

Conclusion and Future Work

We presented our model explanation evaluation benchmark and raised a question on the relation between the two reviewed human-grounded evaluation approaches (i.e., human annotation vs. human review and feedback). In our future work, we plan to run extensive human-subject studies to capture users' feedback on model explanations and compare users' subjective feedback with the quantitative model explanation scores from our benchmark. Also, we are interested in examining that to what extent our proposed multi-layer heatmap annotation can better represent human explanation in comparison to a segmentation mask.

Acknowledgements

This research is based on work supported by the DARPA XAI program under Grant #N66001-17-2-4031 and NSF award #1900767.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *NIPS*.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE.

[3] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[4] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. 2018. Towards explanation of dnn-based prediction with guided feature inversion. In *KDD*.

[5] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *ICML*. 331–339.

[6] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *CVPR*.

[7] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2019. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv preprint arXiv:1811.11839* (2019).

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*. ACM.