A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning

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

Eric D. Ragan
Department of Visualization
Department of Computer Science & Engineering
Texas A&M University
eragan@tamu.com

The benchmark is available online at https://github.com/SinaMohseni/ML-Interpretability-Evaluation-Benchmark for research purposes.

Abstract
In order for people to be able to trust and take advantage of the results of advanced machine learning and artificial intelligence solutions for real decision making, people need to be able to understand the machine rationale for given output. Research in explain artificial intelligence (XAI) addresses the aim, but there is a need for evaluation of human relevance and understandability of explanations. Our work contributes a novel methodology for evaluating the quality or human interpretability of explanations for machine learning models. We present an evaluation benchmark for instance explanations from text and image classifiers. The explanation meta-data in this benchmark is generated from user annotations of image and text samples. We describe the benchmark and demonstrate its utility by a quantitative evaluation on explanations generated from a recent machine learning algorithm. This research demonstrates how human-grounded evaluation could be used as a measure to qualify local machine-learning explanations.

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

ACM Classification Keywords
H.5.m [Information interfaces and presentation]: Miscellaneous

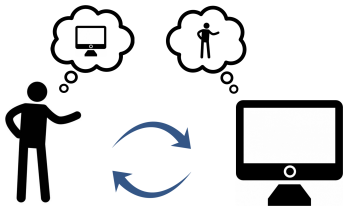


Figure 1: Human and machine interactions have been studied in different levels, but mainly anchored/centered to the human side of conversation. Interpretable machine learning is about self-describing causal machine learning algorithms that explain decision making rules and process to the human.

Introduction

With the recent and continuing advancements in robust deep neural networks, the prominence of machine learning and artificial intelligence models is growing for automated decision-making support and especially in critical areas such as financial analysis, medical management systems, military planning, and autonomous systems. In such cases, human experts, operators, and decision makers can take advantage of new machine learning techniques to assist in taking real-world actions. In order to do so, however, these people need to be able to trust and understand the machine outputs, predictions, and recommendations. Unlike shallow machine learning models that can be interpretable and easier to understand in terms of classification logic, deep learning models are significantly more complex and often considered as black-box models due to their poor transparency and understandability.

Thus, for machine-assisted decision-making using new machine learning technology, advancements are needed in achieving explainability and supporting human understanding. This is the primary goal of recent research thrusts in *explainable artificial intelligence* (XAI). For a system to effectively serve human users, people need to be able to understand the reasoning behind the machine's decisions and actions. Numerous researchers have recently been working to advance explainability through methods such as visualization [9, 3, 10] or model simplification [8, 15]. Different types of interpretability and explainability are possible. For human explainability, for instance, *local explanations* can be used to explain the connection between a single input instance the the resulting machine output, while *global explanations* aim to provide more holistic presentation of how the system works as a whole or for collections of instances. While a multi-faceted topic, the ultimate goal is for people to can understand machine models, and it is therefore important to involve human feedback and reasoning as a requisite component for evaluating the explainability or understandability of XAI methods and models. However, since the majority of research in the area of XAI is lead by experts in machine learning and artificial intelligence, relatively little work has involved human evaluation.

In this paper, we describe an novel evaluation methodology for assessing the relevance and appropriateness of *local* explanations of machine output. We present a human-grounded evaluation benchmark for evaluating instance explanations of images and textual data. The benchmark consists of human-annotated samples of images and text articles to approximate the most important regions for human understanding and classification. By comparing the explanation results from classification models to the benchmark's annotation meta-data, it is possible to evaluate the quality and appropriateness of XAI local explanations. To demonstrate the utility of such a benchmark, we perform a quantitative evaluation of explanations generated from a recent machine learning algorithm. We have also made the benchmark publicly available online for research purposes.

Background

Researchers have argued the importance of interpretable machine learning and how its demand rises from the incompleteness of problem formalization (e.g., [2]). For instance, in many cases, user might lose trust to the system with the doubt that if the machine has taken all necessary factors into account. In this situation, an interpretable model can assist user with generating explanations. Lipton [7] states interpretable machine learning is needed when there is a mismatch between machine objectives and real-world scenarios, which means a transparent machine learning model should share information and decision making details with a user to prevent mismatch objectives problems. The goals of

XAI naturally motivate a merger between human-computer interaction (HCI) and artificial intelligence (AI) disciplines for the creation and evaluation of solutions that are interpretable and explainable for users. It is important that these communities work together to achieve useful and meaningful explanations of machine learning technology.

Explanation Strategies

Interpretable models such as tree-based models [8] and rule lists [15] have been proposed as examples that can be directly explained or summarized using relatively simple or common visualization methods. For more complex black-box models such as deep neural networks (DNN), other methods have been explore to generate local explanations for each individual instance as well as for global explanations of the entire model. Local explanations in the form of salience maps is a popular way to generate explanations in DNNs. This approach presents features with the greatest contribution to the classification. For example, Simonyan et al. [13], used output gradient to generate a mask of which pixels is the model relying on for classification task. In other work, Ribeiro et al. [11] presented a model-agnostic algorithm that generates local explanations for any classifier in different data domains. As another example, Ross et al. [12] proposed an iterative approach using an input gradient that can improve its explanation by constraining explanation by a loss function.

Data visualization is also a basic tool to show the relationship between data points and clusters. Methods like MDS and t-SNE [9] generate a 2D mapping of high-dimensional data to visualize spatial relation of data clusters. Visual analytics tools such as ActiVIS [3] and Squares [10] take advantage of a 2D mapping of data points along with feature-cluster and instance-cluster views to help users with performance analysis and in understanding classification logic.

Evaluating Explanations

In considering evaluation approaches for XAI, Doshi-Velez [2] proposed three categories: *application-grounded*, *human-grounded* and *functionality-grounded* evaluations. These categories vary in evaluation cost and inclusiveness. In this taxonomy, functionality-grounded evaluation uses formal definitions of interpretability as a proxy for qualifying explanations and no human research is involved. *Application-grounded* evaluation is done with expert users reviewing the model and explanations in real tasks. Analytics tools like *ActiVis* [3] with participatory design procedure and case studies show satisfactory results from the expert users in machine learning field. Krause et al. [4] also proposed a visual analytics tool for the medical domain to debug binary classifiers with instance-level explanations. They worked tightly with the medical team and hospital management to optimize processing times in the emergency department of the hospital.

In contrast, *human-grounded* evaluations are generally performed with non-expert users and simplified tasks. To date, there are few research studies involving human subjects to assess XAI. Ribeiro et al. [11] presented an experiment to study whether users can identify the best classifier using their explanations. In their study, participants reviewed explanations generated for two image classifiers. They also performed a small study where the researchers intentionally trained a classifier incorrectly with biased data to study whether participants could to identify the connection between the incorrect features the resulting erroneous classifications. Also studying interpretability for people, Lakkaruj et al. [5] conducted research with interpretable decision sets, which are groups of independent if-then rules. They evaluated interpretability through a user study where participants looked at the decision-set rules and answered a set of questions to measure their understanding of the model.

(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 2: Examples of users *feedforward explanations* for image and text data. (a) Heat map views from 10 users for drawing a contour around the area which explains the object in the image. (b) Heat map view from two expert reviewers for highlighting words related to the "electronic" topic.

The authors reported that both accuracy and average time spent in understanding the decisions was improved with their interpretable decision sets comparing to a baseline with bayesian decision lists.

Human Evaluation Approaches

We discuss two main classes of approaches for human evaluation for interpetability, with the difference depending on whether users have prior knowledge or access to sample explanations. In one way, users review existing explanations and provide specific feedback for those explanations. The other option is to capture users' thoughts and opinions of the most relevant features based on the input and output without review of example explanations.

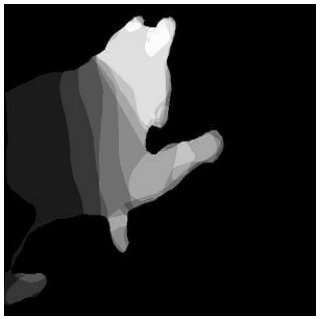
The explanations could be in any form such as verbal or local explanations and on any data such as image, text, or tabular data. The following subsections provide further description of each of human-grounded local explanations evaluation types for machine learning .

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 and the resulting output. For example, users could review several options and choose the best machine-generated explanation for a straight-forward comparison.

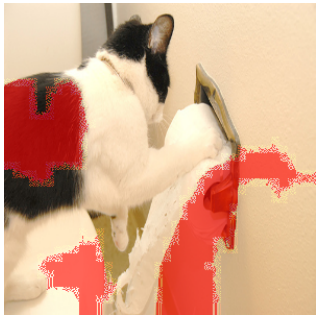
User decisions are made with knowledge about the input, the explanations, and the output label. We would expect users to generally pick explanations that most closely match their logic and background knowledge. One advantage of this method is the ability for a clear comparison of multiple interpretable machine learning algorithms. Another means of capturing user feedback would be letting a user interactively refine machine generated explanations. This method has more flexibility in allowing rejection of wrong features and adding new features to the explanations. Quantifying the difference between an initial given explanation and a user-edited explanation could give a clear measure of quality for the initial machine-generated explanations. The disadvantage of this method is that human review is always a comparison relative to an existing explanation, which means (1) some form of explanation must already exist, (2) the evaluation is specific to the particular explanations reviewed, and (2) reviewing the existing explanation might bias a user's perception.

Evaluating with Input Review and Feedforward
Another option for human evaluation is to collect feedback about the features that would best contribute towards an explanation for a given output—and users could provide such information without seeing example explanations. For example, explanations could be obtained by presenting the user with the input and output label and then asking to find the relevant features corresponding to the label. 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. User choice is made with knowledge about the input along with the output label. 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 user opinions. For example, Figure 2 examples of text and image heatmaps generated by this approach for our benchmark.

This can be thought of as a *feedforward* approach, as the information from reviewers would be independent of any particular explanation. Consequently, this approach can result in a reusable benchmark that can apply to various explanations.



(a)



(b)

Figure 3: (a) User generated weighted-masks for an example image of a cat. We use this weighted-mask to evaluate machine generated explanations accuracy. (b) Machine generated explanation by the LIME algorithm for the same image. Irrelevant red-highlighted regions in this image cause low explanation precision comparing to human generated explanations.

Evaluation Benchmark

Because our goal was a benchmark that could be used for evaluation of known input and classifications, we captured explanations in a feedforward approach where the users were asked to annotate relevant regions in images and words in text articles that are most related to 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* [1] and *20 Newsgroup* [6] data sets. The initial version of this benchmark is available online [1] for research purposes.

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 a review-board approved user 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 recognizing the object, or the portion that, if removed, you could not recognize the object. None of the participants were experts in any of the image categories. Each participant annotated all images in a random ordering.

All user annotations are accumulated to create a weighted explanation mask (see Figure 3a) over the image. Figure 2a shows a heatmap views of user annotated explanations over two 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 user annotations with exact contour shapes to reduce the impact of user imprecision or hand jitter.

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

Annotated Text Examples

All text articles were collected from two categories (medical or *sci.med*, and electronic or *sci.elect*) from the *20 Newsgroup* data set. For each category, expert reviewers highlighted the most important words relevant to the given topic name (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. Figure 2(b) shows a single tone heat map view of user annotated explanations over a partial sample text article.

Use Case

To demonstrate the utility of our benchmark, we present a use case in evaluating local explanations from the well-known LIME explainer [11]. Similar to the previously presented research on LIME [11], we used the pre-trained model from Google’s *Inception v3* [14] for image classification.

Next, we compared the machine-generated explanation with our evaluation benchmark. The comparison is done pixel-wise for each image sample We compared our weighted-masks (see Figure 3a) to the LIME results for all 100 images in our benchmark set. We calculated true positive, false positive and false negative pixels with bit-wise operations, and precision and recall for the set were calculated as 0.39 and 0.58, respectively. The low precision is indicative of extraneous irrelevant regions of the images being highlighted in explanations by the LIME algorithm. Figure 3b shows an example of image explanations from the LIME algorithm where two of the red highlighted patches show regions that do not correspond to the cat in the image. Using this evaluation method, we would hope to see algorithms produce local explanations with closer alignment to user

annotations.

Acknowledgements

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

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition. CVPR.* IEEE, 248–255.

[2] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. (2017).

[3] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2018. A cti V is: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 88–97.

[4] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. 2017. A Workflow for Visual Diagnostics of Binary Classifiers using Instance-Level Explanations. *arXiv preprint arXiv:1705.01968* (2017).

[5] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1675–1684.

[6] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*. 331–339.

[7] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[8] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 150–158.

[9] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[10] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2017. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics* 23, 1 (2017), 61–70.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.

[12] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. *arXiv preprint arXiv:1703.03717* (2017).

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[15] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*. 1013–1022.