Bridging AI Developers and End Users: an End-User-Centred Explainable AI Taxonomy and Visual Vocabularies

Weina Jin[*]
Simon Fraser University

Sheelagh Carpendale[†]
Simon Fraser University

Ghassan Hamarneh[‡]
Simon Fraser University

Diane Gromala[§]
Simon Fraser University

Abstract

Researchers in the re-emerging field of explainable/interpretable artificial intelligence (XAI) have not paid enough attention to the end users of AI, who may be lay persons or domain experts such as doctors, drivers, and judges. We took an end-user-centric lens and conducted a literature review of 59 technique papers on XAI algorithms and/or visualizations. We grouped the existing explanatory forms in the literature into the *end-user-friendly XAI taxonomy*. It consists of three forms that explain AI's decisions: **feature attribute**, **instance**, and **decision rules/trees**. We also analyzed the visual representations for each explanatory form, and summarized them as the *XAI visual vocabularies*. Our work is a synergy of XAI algorithm, visualization, and user-centred design. It provides a practical toolkit for AI developers to define the explanation problem from a user-centred perspective, and expand the visualization space of explanations to develop more end-user-friendly XAI systems.

1 Introduction

As artificial intelligence (AI) and deep learning advance significantly and begin to influence society and our everyday life in unprecedented ways, the "black-box" nature or lack of transparency issue of many AI applications becomes a notable problem, especially in domains which involve AI in critical decision-support scenarios, such as medicine, finance, law, military, and autonomous driving. The re-emerging research field of interpretable or eXplainable AI (XAI) aims to tackle the interpretability problem of AI, and to explain AI's decisions to users in terms that users can understand [2]. Although many different XAI algorithms have been proposed in recent years, much attention is on developing XAI approaches for AI experts to visualize, understand, debug or improve the AI models, leaving the major users of AI who are the end users ignored.

The end users are people who do not have prior knowledge in data science, machine learning (ML) or AI. They can be lay persons, or domain experts such as doctors, bankers, judges, drivers. Compared to creating explanations for AI experts, generating explanations for end users is more challenging, since it is unrealistic to ask the end users to interpret the internal parameters and complex computations of the ML models, and they have a diverse range of needs and requirements of using XAI system.

To bridge the gap between the XAI techniques and end uses, we first refer to the theories of explanations [11]: psychological experiments show that people prefer simple, **selected** (but may be biased), and **causal** explanations; explanations are **contrastive** to some other related predictions; similarly, people tend to seek causal reasoning in a **counterfactual** fashion, i.e. what would the prediction be if some features in the input had been different; explanation is **contextual**, **a social process** and usually formed as a conversation [17]. Based on the above insights from end users' perspective, we reviewed 59 XAI technique papers by searching "explainable/interpretable/visualizing AI/ML" and identifying relevant works from these papers' reference. We excluded works that do not have an evaluation of the proposed XAI method. Based on the method availability and popularity, we proposed the **end-user-centred XAI taxonomy** (Figure 1). The taxonomy categorizes the existing XAI techniques regarding their explanatory forms. It also provides the possible visualization (the **visual vocabularies**) for each form. We regard the explanatory process between the XAI system and its end users as dynamic rather than monolithic [11], i.e., communications and interactions happen frequently and back-and-forth, to allow users to interpret AI's decisions from different perspectives and iteratively build AI's theory of mind. The XAI taxonomy and its visual vocabularies provide the needed vocabularies to have such a conversation. We choose visualization as the primary explanation format in XAI, not only because it is the prevalent form of representing explanations in the XAI technique papers, but also it augments human's information processing capability by leveraging the high-bandwidth visual processing channel in our brain.

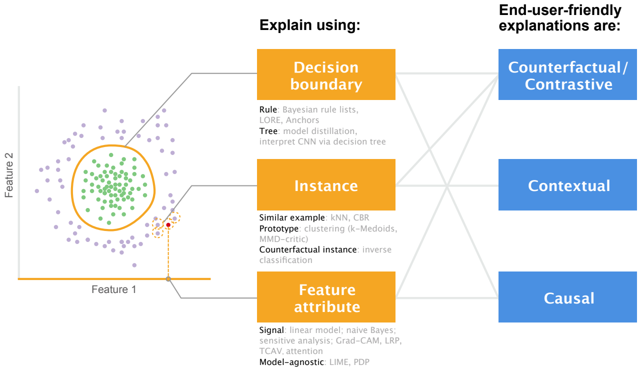
Explain using:
End-user-friendly explanations are:
Feature 2
Feature 1
Decision boundary
Rule: Bayesian rule lists, LORE, Anchors
Tree: model distillation, interpret CNN via decision tree
Instance
Similar example: kNN, CBR
Prototype: clustering (k-Medoids, MMD-critic)
Counterfactual instance: inverse classification
Feature attribute
Signal: linear model; naive Bayes; sensitive analysis; Grad-CAM, LRP, TCAV, attention
Model-agnostic: LIME, PDP
Counterfactual/ Contrastive
Contextual
Causal

Figure 1: The plot shows a toy example of a machine learning classification task. For an unknown data (red dot), the end-user-centred XAI taxonomy shows explanations can be created at three levels. Each has its subcategories and exemplars listed below, and is mapped to the characteristics of explanations.

2 The End-User-Centred XAI Taxonomy

The XAI taxonomy and its visual vocabularies consist of three primary explanation forms: explain using **1) feature attribute**, **2) instance**, and **3) decision-tree/rules**. Serendipitously, it corresponds to the granularity of the learned representations of ML models at the feature level, instance level, and decision boundary level.

2.1 Explaining using *feature attribute*

Mapping to explanation theories Feature attribute is the most common form of explanation. The feature attribute relates to **causal** explanation since people tends to explain the causality of prominent features to the prediction. The feature attribute explanations are handy to show **contrastive** explanations: users can choose a reference prediction (usually the most suspected other than the predicted one), and

[*] e-mail: weinaj@sfu.ca

[†] e-mail: sheelagh@sfu.ca

[‡] e-mail: hamarneh@sfu.ca

[§] e-mail: gromala@sfu.ca

interactively check how the explanatory features differ between the predicted and the referenced one [10]. Some perturbation/sampling-based XAI algorithms can be applied to generate **counterfactual** explanations. Users can check what would change to the prediction if some of the feature values are changed [13].

Exemplars We divide the XAI techniques into two sub-categories: 1) **Signal**: the XAI algorithm needs to access the internal state of the original model, such as the model parameters (LRP [1]), gradients (sensitivity analysis [16], Grad-CAM [15]), activation (CAM [21], attention [20]), and the learned representations (TCAV [7]). 2) **Model-agnostic**: the XAI algorithms only need to access the input-output pairs without accessing the model's internal parameters, such as LIME [13] and partial dependent plot (PDP) [3].

Visual vocabulary The visual representation of feature attribute largely depend on the feature data types. For example, for image data, the **saliency map** is a common form to visualize the fine-grained feature importance score at the pixel level, by overlaying a color map on the input image. Other popular methods include using masks, segmentation maps, or bounding boxes on image data. **Scatter or line plot** can show the effect of individual feature on the overall prediction (feature shape). **Bar plot** or **box plot** are typical choices to visualize the multiple feature attributes in tabular or text data. The variations of bar plot include **waterfall plot, treemap, wrapped bars, packed bars, piled bars, Zvinca plots, tornado plot**. The variations of box plot include **violin plot and beeswarm plot** that show more detailed data distributions. **2D or 3D heatmap** is used to visualize the feature interactions and their predictions. More complicated feature-feature interactions can be visualized with **matrix heatmap**, **node-link network**, or **contingency wheel**.

2.2 Explaining using *instances*

Mapping to explanation theories People use examples to learn and explain. If the input data itself is structural and interpretable (like image or text), the instance-based explanations are intuitive for a human to interpret. The instance-based explanation also carries more **contextual** information where the typical or untypical/**counterfactual** features reside.

Exemplars The instance-based explanation includes showing instances which are either similar or juxtaposed to the query instance. Nearest neighbour (e.g. kNN) and case-based reasoning ([8]) methods are used to find **similar instances**. k-Mediods, MMD-critic [6], generating representative prototype [9, 16] are proposed to obtain the **prototypical or typical examples** of the prediction.

2.3 Explaining using *decision-tree/rules*

Mapping to explanation theories The form of decision trees or rule lists/sets mimic the **causal** chain of reasoning, thus can easily fit in the users' reasoning process. Since it explicitly gives the decision boundary, it is convenient to obtain the **counterfactual or contrastive** cases from the decision boundary.

Exemplars There are different techniques to learn a **decision tree** from the original black-box models, e.g., model distillation [4], disentangle model's representations [19]. **Rule** learning algorithms include: Bayesian Rule Lists [18], LORE [5], Anchors [14], etc.

Visual vocabulary Different visualization techniques are utilized to visualize the hierarchical structure of trees and rules. For decision tree, the most common representation is to use **node-link tree**, other visual representation formats include **treemap, cladogram, hyperbolic tree, dendrogram, flow chart**. For the representation of rule lists or sets, **IF-THEN text** is the most common format. Other representing formats include **matrix** [12] or **table**.

3 Discussion and Conclusion

In this work, we took an end-user-centred lens to review the XAI technique literature, and proposed the end-user-centred XAI taxonomy and its visual vocabularies. The essence of using them is to combine the proper vocabularies to generate a comprehensive explanation for various audiences and usage scenarios. For example, prototypical explanations can be enhanced with highlighted feature attributes; the nodes in a decision tree can be represented using prototypical examples. Our work can empower AI developers to create more concrete, low-fidelity prototypes to probe end user's specific requirements and get detailed feedback. It can also inform AI developers of the possible end-user-friendly explanatory forms, and guide or regularize the XAI algorithm/visualization development. In our future work, we will evaluate the XAI taxonomy and its visual vocabularies with end users.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, jul 2015.

[2] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. feb 2017.

[3] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, oct 2001.

[4] N. Frosst and G. Hinton. Distilling a Neural Network Into a Soft Decision Tree. Technical report, 2017.

[5] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. Local Rule-Based Explanations of Black Box Decision Systems. may 2018.

[6] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! Criticism for Interpretability, 2016.

[7] B. Kim, W. M., J. Gilmer, C. C., W. J., , F. Viegas, and R. Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) . *ICML*, 2018.

[8] J. L. Kolodner. An Introduction to Case-Based Reasoning. *Artificial Intelligence Review*, 6:3–34, 1992.

[9] O. Li, H. Liu, C. Chen, and C. Rudin. Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. Technical report, 2017.

[10] S. M. Lundberg, P. G. Allen, and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017.

[11] T. Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. jun 2017.

[12] Y. Ming, H. Qu, and E. Bertini. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):342–352, jan 2019.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD*, pages 1135–1144, New York, New York, USA, 2016. ACM Press.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-Precision Model-Agnostic Explanations. *Association for the Advancement of Artificial Intelligence*, 2018.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. oct 2016.

[16] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Technical report, 2013.

[17] D. S. Weld and G. Bansal. The Challenge of Crafting Intelligible Intelligence. mar 2018.

[18] H. Yang, C. Rudin, and M. Seltzer. Scalable Bayesian Rule Lists. In *ICML*, 2017.

[19] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu. Interpreting CNNs via Decision Trees. jan 2018.

[20] Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F. K. Khalil, S. I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, and L. Yang. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019.

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. Technical report, 2015.