An Integrative 3C evaluation framework for Explainable Artificial Intelligence

Completed Research Full Paper

Xiaocong Cui
Georgia State University
xcui4@gsu.edu

Jung min Lee
Georgia State University
jlee469@gsu.edu

J. Po-An Hsieh
Georgia State University
jjhsieh@gsu.edu

Abstract

The emergence of machine learning (ML) based artificial intelligence (AI) bring about fear because of its power and uncontrollability. In response, scientists and engineers are developing explainable AI (XAI) techniques to tackle this concern. However, the literature is short of a systematic approach to assess the various XAI techniques in a balanced and comprehensive manner. To address this gap, we survey the current XAI technique and propose an integrated framework with three evaluation criteria (correlation, completeness, and complexity) to evaluate XAI. Applying this framework, we find the rule extraction method is the most advanced and promising method among current XAI.

Keywords

Explainability, Evaluation Framework, Correlation, Completeness, Complexity, XAI

Introduction

Machine learning based artificial intelligence (AI) has achieved great success in image recognition, text mining, audio analysis and other areas in terms of its high accurate predictability. However, this success is counterbalanced by the unexplainable features which may lead to acceptance, trust and usefulness problems in AI (Hoffman et al., 2018). Motivation to develop explainable AI (XAI) is so strong in both academy and industry that many new techniques are developing rapidly. This raises a new question about how to evaluate these techniques. The question is complicated because it crosses the boundary of computer science, psychology, and social science. This is the motivation for our research.

There are two basic roles in explanation scenario: explainer – information provider, and explainee – information accepter. Most study on XAI mainly focuses on explainee's perspective, which is easy to fall into a persuasive explanation (Herman, 2017). Herman (2017) suggested that we should keep alert when evaluating explanation from explainee's perspective because human evaluation often involves with biases. Therefore, we propose an objective functional explanation framework in this research. We define functional explainability as the ability to fill the information gap between explainer and explainee in an understandable way. It is like a product brochure, in which product specifications, terms, ingredients, mechanisms, and instructions are expressed objectively. The functional explanation should tell explainee how the information gap is filled completely by inference processes and how they are fully connected. What's more, the information from explainer should simple enough so that it is reachable to explainee. We propose three criteria to evaluate functional explanation: level of correlation, level of completeness, and level of complexity. Then we use these criteria to evaluate current XAI techniques.

The rest of the paper is organized as follows. Section 2 is a survey of current XAI techniques. In section 3 we proposed a definition for functional explainability and three evaluation criteria: correlation, completeness, and complexity. Then we evaluate current XAI techniques in section 4. The last part is the discussion and conclusion of this paper.

Survey of current Explainable Artificial Intelligence

The accuracy itself is an evaluation for interpretability (Guidotti et al., 2018, Andrews et al., 1995, Doshi-Velez and Kim, 2017, Freitas, 2014; König and Niklasson, 2004). However, it only provides how accurate the prediction is. The most basic reasons about how and why such prediction was made are not provided. When dysfunctions or unethical solutions happen, accuracy by itself cannot support users (both experts and lay users) to debug or repair the AI (Weller, 2017; Ras et al., 2018). Therefore, there is a pressing need to develop interpretable skills for machine learning techniques. In this section, we survey recent major progress in XAI.

In order to develop an evaluation framework for XAI techniques, this paper conducts a brief survey on XAI technologies that extracted from a couple of most recent, comprehensive and widely cited review papers on XAI technologies (e.g. Guidotti et.al (2018), Adadi and Berrada (2018) and Mueller et.al (2019)). Due to the rapid development of XAI technologies, this survey may not cover all emerging XAI technologies. However, classical and mainstream XAI techniques are considered, especially for techniques that are commonly included in Guidotti et al. (2018), Adadi and Berrada (2018), and Mueller et al. (2019) and so forth.

Convolutional Neural Network (CNN)

The interpretable CNN aims to explain what patterns and features of images are learned in the network by visualizing the distribution of object parts (Zhang, Wu, and Zhu, 2018). Typical papers that work on the visualization of CNN representations include Zeiler, Taylor, and Fergus (2011) and Zeiler and Fergus (2013 and 2014), which are very effective to invert the feature maps hidden inside of filters to images. Specifically, the main technique is the gradient-based method which is widely used in interpretable CNN. Another method is the up-convolutional net that inverts features learned by CNN to images, although the recovered images are not perfectly original images. Another progress in interpretable CNN is to diagnose CNN representations to explore the semantic meanings expressed from features learned by CNN. Techniques mainly focus on how to estimate the image regions that directly lead to the corresponding output (e.g., Fong and Vedaldi, 2017; Selvaraju et al., 2017). Research in this area also includes vulnerable points estimation (e.g., Su et al., 2017; Koh and Liang, 2017) and biased representations in CNN (Zhang et al., 2018a). CNN disentangling is another powerful technique to reveal hidden semantic meanings from mixed features learned by the network, such as the spatial relationship between two patterns (Zhang et al., 2016; Zhang et al., 2018b).

Local Interpretable Model-agnostic Explanations (LIME)

LIME, proposed by Ribeiro et al. (2016), is a powerful proxy model to explain the prediction of any classification methods (e.g., random forests and neural networks) by learning from local perturbations (Ribeiro et al., 2016; Gilpin et al., 2018). LIME uses an interpretable model to approximate the black-box model (Gilpin et al., 2018). The interpretable model is often a simple model (e.g., linear model) with limited complexity, in which only the most important features are used for explanation. Specifically, the interpretable model learned from perturbations of an input. By resampling from the vicinity of the original instance, a small sample is generated and then used to train the simple model (such as a linear model), which is the approximated interpretable model.

Feature Importance Ranking Measure (FIRM)

The FIRM method is based on the underlying correlation structure of the features. The feature importance is estimated by its total impact on the prediction. For a specific feature, the importance depends on the conditional expected value of correlated features (Zien et al., 2009; Guidotti et al., 2018). So, FIRM takes indirect effect among variables into account. There are three important advantages of FIRM: objective - it is not sensitive to translation and rescaling of the features, universal – it can evaluate all features and generalized to any learning methods, and intelligent – it reveals correlations between features that are not considered in the predictor (Zien et al., 2009). However, the effect of FIRM builds upon the assumption that input features should follow normal distributions, which may impair the generalizability.

Rules extraction

Rule-extraction methods attempt to add explainability to black-box AI models by generating classification models. Rule extraction methods focus on three approaches: decompositional approach, pedagogical approach and eclectic approach (Andrews et al., 1995; Ras et al., 2018). Decomposition approach translates components of a model (hidden layer) into sets of rules, and then combine and prune the rule sets. The pedagogical approach learns the target model by using the target model as an input training data, which is a similar process to build a prediction model. The eclectic approach incorporates both pedagogical and decompositional approaches in a clustering and training way. The generated outputs of those rule extraction approaches provide comprehensible information on how a model processed its input, in forms of sets of rules or decision trees. These forms allow users to trace back to have a glimpse on the decision process of the given outcome. However, there are limitations in rule-extraction method. First, there is a trade-off between accuracy and interpretability (Johansson, 2007). Second, there is no quantitative method to decide the size of the rule set. It relies mainly on the user's experience and intuition (Ras et al., 2018).

Sensitivity analysis (SA)

SA studies how the change of one input variable within a range can influence the output while controls other input variables constant. Therefore, by calculating the change of output, the contributions of the input variable can be assessed (Olden and Jackson, 2002). Garson (1991) is one of the pioneers who tried to crack the black box of a neural network by calculating the relative importance of each input variable. The basic idea is to do the multiplication according to weights connection among different layers (Garson, 1991). However, it ignores the direction of the relationship between input and output variables, and it only works for small neural network structure, while the cost of calculation is too high for complex networks.

The breakthrough is from Olden and Jackson (2002). It provides a randomization approach to make sure that the weights evaluation and input variable contribution can be assessed statistically. Later, other SA methods are proposed. For example, Cortez and Embrechts (2011) proposed a Global SA (GSA) method in which features simultaneously vary in different range. GSA measures the impact of input interactions on output by providing a ranking of input variables based on their importance (Cortez and Embrechts, 2011). Therefore, the method can both explain the model and provide a new way to select variables/features (Cortez and Embrechts, 2011).

Random forest (RF)

RF reflects the thoughts of ensemble learning, which integrates the strengths of multiple learning algorithms to build a better prediction model than any of these algorithms (Friedman et al., 2016, pp 605). The effectiveness of RF is based on the use of randomness. Randomness in RF is implemented in two ways: bagging and randomized node optimization. With the bagging technique, the training dataset for each tree is drawn from the sample using a bootstrap aggregating technique, so trees are independent with each other (Friedman et al., 2016, pp 316-317). Randomized node optimization decides how to make a split decision for each tree - a subset of attributes is randomly selected from total attribute set, from which the best split variable is selected (Criminisi et al., 2012). The variable importance in RF based on out-of-bag (OOB) cases, which are samples used to construct a tree in bagging. For each case in OOB, we can predict the response by using trees in which the case was OOB (Friedman et al., 2016, pp 317-318). Then by randomly change the value of a specific variable while control other variables constant, the change of accuracy can be measured. The mean value of prediction change can be used as variable importance. Another way to measure variable importance is the Gini importance index. For a specific variable that is used to split trees, we can get the Gini importance index for this variable by summing the changes of Gini impurity across all trees.

Wachter et al. (2017)'s method

Wachter et al. (2017)'s counterfactual XAI technique aims to construct XAI in the level of human-understandable counterfactuals. The algorithm is very straightforward, especially when we compare the cost functions of this technique with cost functions of traditional machine learning algorithms.

$$\arg\min_{w} \ell(f_w(x_i), y_i) + \rho(w) \quad (1)$$

$$\arg\min_{x'} \max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x') \quad (2)$$

Where (1) is the cost function of traditional machine learning algorithms, (2) is Wachter et al. (2017)'s method, in which x' is the counterfactual of the original instance x. $f_\omega(x')$ is the predicted value of x'. The goal is to find such x' that close to x as possible and at the same time, to make sure $f_\omega(x')$ is equal to y'.

Wachteret al. (2017)'s method is model-agnostic. The optimizer used in equation (2) can be any of the existing machine learning algorithms. The distance between the original instance and the corresponding counterfactual one is controlled by Manhattan distance, which gives the algorithm several advantages. First, the counterfactual instance is kept close to the original instance, even when the instance varies across the dataset. Second, the Manhattan distance metric is robust to outliers compared to other distance measures, such as mostly used standard deviation metric. Third, the Manhattan metric often utilizes a sparser trained model compared to other methods, which leads to a simpler trained model.

Evaluation criteria for explainability

Functional explainability

There are many different definitions on the explainability of XAI (e.g., Van Lent et al., 2004; Miller, 2018; Guidotti et al., 2018; Adadi and Berrada, 2018). In this research, we define explainability from a functional perspective. Explainability is the ability to fill the information gap between explainer and explainee in an understandable way. In this research, we distinguish two types of explanation: functional explanation and social explanation. The typical scenario of functional explanation is between AI experts, such as AI engineer and AI developer (Ras et al., 2018). Because they share a lot of professional AI knowledge. Therefore, the explanation provided by explainer is selective without considering too much about explainee's knowledge background (Lombrozo, 2012). Social explanation mainly happened between AI experts (as explainer) and lay users (as explainee). It focuses on how to present the explanation to make sure that explainee can invest the lowest cost to build a mental model about explanation (Kulesza et al. 2013; Kulesza et al., 2015). It considers explainee's knowledge structure, belief, interests, expectation, preference and even personality (Miller et al., 2017). The social explanation is used to enhance explainee to appropriate trust and use and avoid distrust and misuse (Miller, 2018).

As two of the most comprehensive review on XAI as far as we know, Hoffman et al. (2018) lists a couple of attributes of explanations, including "*understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness*"; Mueller et al. (2019) lists good explanation features, including appropriate detail, veridicality, usefulness, clarity, completeness, observability, and dimensions of variation. As mentioned before, we only consider the functional part of the explanation. Therefore, only some of them are considered. Specifically, after scrutinized overlapped features between Mueller et al. (2019) and Hoffman et al. (2018), and filtered out social aspects features such as trustworthiness, satisfaction, clarity, and sufficiency of detail, three criteria are considered to evaluate functional explanation: correlation, completeness, and complexity. To fill the information gap, the information provided by explainer should consider correlation and completeness (Miller et al., 2017; Miller, 2018). Correlation shows how the information provided by explainer bridges the information gap by showing the reasoning process. The completeness makes sure that the information gap is fully connected. To make the information understandable by explainee, we should consider complexity, which is another factor that influences the understanding process. The simpler the information from explainer, the easier to understand by the explainee (Kulesza et al., 2015).

Level of correlation

The role of correlation is to bridge the information gap between explainer and explainee and makes sure that the information is understandable by explainee. In this part, correlation is divided into three levels according to Pearl and Mackenzie' Ladder of Causation: association, intervention, and counterfactual.

Association answers questions like "what is?" or "what does the red spot tell me about a disease?". The answer only requires the correlation rather than causality from data. Intervention not only involves questions about "what does X tell me about Y" but also questions like "what happens to Y if we double X?". This level involves specific causal relationships which cannot be answered just from the data (Pearl, 2018). For example, maybe factor Z also influence Y. Therefore, the right way to answer questions at this level is to study the relationship between X and Y after controlling all other factors that influence Y except X. The highest level is the counterfactual. The typical question in this level is "what if" question. For example, "what if I choose P rather than Q?". It asks for answers for alternatives in a past event. The three levels are hierarchical because when we answer questions at the counterfactual level, we also answered questions at the other two levels. When we answer questions of intervention level, we also answer questions in the association level (Pearl, 2018). The typical model that can make counterfactuals computationally manageable is Structural Causal Model (SCM), which includes three functions: graphical models (e.g., path models), structural equations, and counterfactual and interventional (such as mediation and moderation analysis) logic (Pearl, 2018).

Level of completeness

Kulesza et al. (2013) define completeness as "*the extent to which an explanation describes all of the underlying systems*." There is a consistent conclusion that completeness is an important attribute of explanation (e.g., Hoffman et al., 2018; Kulesza et al., 2015; Martens and Provost, 2014; Doshi-Velez and Kim, 2014). Kulesza et al. (2013) found that completeness helps users to build better mental models about an intelligent agent. Completeness plays a more important role than the soundness of explanation content. In their research, the completeness implementation is based on Lim and Dey (2009), which mainly focuses on the users' interests and demands. Specifically, about the function aspect of an application, users want to know the information about input, output and the conceptual model, which describes how the application process input and generate output. Regarding the non-functional aspect of the application, users interested in control – how to control the application, and certainty – the extent of certainty when an application performs its action. Gilpin et al. (2018) considered accuracy and predictability into completeness. The completeness of an explanation, on the one hand, denotes that the explanation should provide accurate information about the operation of a system; on the other hand, it should allow users to predict system behavior in a different context.

Lim and Dey (2009) indeed provided a comprehensive framework for completeness. However, user's demanding and interests are different from what they should know about the system. "want to do" and "should to do" are different because the former may cause decision-making bias (e.g., confirmatory bias) while the latter is more objective (Bazerman et al., 2013). Therefore, we need to be very careful when using them as evaluation criteria in functional explanation framework, which focuses on explainer's perspective and the corresponding information accepter should be at the same expertise level with the explainer. By integrating existing literature, we proposed the following criteria about completeness.

The first criterion about completeness is on the boundary of inputs and outputs of a model:

(1) Provide complete information on the boundary of inputs: eligibility and range of input variables, sample size, and even how and where to get the sample. Complete information on the boundary of output: the range and accuracy of output variables.

Following criteria about completeness is in regards to the model:

(2) Provide complete information on AI model: illustrate all hyper-parameters and functions/algorithms used to construct the AI model. It should be accurate and complete enough so that other experts can repeat the current model.

(3) Provide complete information on explanation methods: describe information about how to build the explanation methods (e.g., sensitivity analysis, rule extraction, LIME or FIRM) and the function of these explainable methods.

On top of the existing literature, we propose three new criteria that completeness of an explanation should include:

(4) Provide complete information about the limitations and advantages of AI models and explanation methods, and the corresponding benefits and risks of the AI system.

(5) Provide complete information about how to use the AI model and how to interpret results of explanation methods.

(6) The scope covered by explanation and the corresponding evaluation criterion. The global-based explanation is usually more complete than local-based explanation. This is because the former provides information about the behavior of the whole model, while the latter only focuses on the instance and provides information about why the AI model makes a specific decision for this instance (Adadi and Berrada, 2018). Another factor is whether the explanation also provides corresponding quantitative criterion. For example, rule extraction method provides extracted rule as an explanation. However, there is no method to decide the size of the rule set, which is mainly decided by the user's experience and intuition (Ras et al., 2018).

Level of complexity

Complexity is another commonly accepted criterion to evaluate explanation. Thagard (1989) proposed Theory for Explanatory Coherence, in which he suggested that there are seven principles about how explanations fit with prior belief. He emphasized that all things being equal, people prefer simpler explanations that cited fewer causes but more general. This viewpoint is verified by the experiments from Read and Marcus-Newhall (1993), in which participants like simple explanations with fewer causes. Carroll and Rosson (1987) proposed a minimalist explanation model which argued that people favor short explanations rather than a long one. Rosson et al. (1990) verified this model and found that a simpler explanation helps programmers to understand faster than traditional explanation. Recently, Kulesza et al. (2015) advocate that the explanation should not be overwhelmed. Miller (2018) proposed that for providing causality and generate trust, simplicity is a necessary condition for an intelligent agent.

By integrating existing research, the complexity depends on the number of cases and the length of explanations. When applying this criterion in the XAI context, it is easy to find that global-based explanation provides more information in a simple way, while local-based explanation provides simple information in a complex way.

An Integrative Evaluation Framework for XAI

Level of correlation

Wachter et al. (2017)'s method. According to Pearl and Mackenzie's ladder of causation, counterfactual is the highest level in causal reasoning (Pearl and Mackenzie, 2018). Therefore, the counterfactual explanations from Wachter et al. (2017)'s method should be more advanced than current other XAI techniques. However, that's not the case. It indeed offers several advantages. It provides support for decision making, including understand or challenge solutions from black-box AI and correct future behavior (Wachter et al., 2017). It is easy to compute and at the same time, easy to convey the explanation to lay audience. However, it circumvents the most urgent need for revealing internal mechanism in an explainable way, just as its title illustrated – "*counterfactual explanations without opening the black box*." Wachter et al. (2017)'s method is about instance-level counterfactual. It is not comparable to counterfactuals generated from causal models such as structural causal model (SCM), which is more powerful (Wachter et al., 2017). Wachter et al. (2017)'s method can answer questions like "what if I choose P rather than Q?" by providing explanations about "when choosing Q, the output is Y1; If choosing P, the output is Y2". It seems to answer counterfactual questions. However, it does not answer questions "why Q gets Y1 while P gets Y2?" The black-box is still opaque.

Convolutional Neural Network (CNN). Interpretable CNN provides an intuitive way to help people to understand the learning process of the network. It can answer "what" questions by directly show the input, visualized features, and corresponding output with accuracy. However, because of the lack of statistical indices for these visualizations, the interpretation of these visualizations may vary from different perspectives (Hadji and Wildes, 2018). The vulnerable points estimation and biased representation techniques of interpretable CNN can answer some degree of counterfactual questions by showing output for noisy perturbation or biased representation of input images. Similar to Wachter et al. (2017)'s method, it is instance-based counterfactual. But it explains more than Wachter et al. (2017)'s method by providing information about "why I chose P rather than Q."

Local Interpretable Model-Agnostic Explanations (LIME). Similar to Wachter et al. (2017)'s method, LIME provides a local explanation for a single instance. It has the potential to answer counterfactual questions "what if I choose P rather than Q?" by providing explanation for both P and Q. However, unlike Wachter et al. (2017), it has the potential to answer questions about "why I chose P rather than Q" because of the simple model learned from perturbations of an input is an explainable model.

Rule-extraction methods. The rules extracted from rule-extraction methods allow users to trace back decision process of a given outcome. Therefore, answers for counterfactual questions are in the rule set. Compared to interpretable CNN, the explanation provides by rule-extraction methods not only answer questions from instance level but also explain the relationship from some aspects of input instances and corresponding outputs. This is decided by the size of the rule set. Therefore, the answers for counterfactual questions can be instance level (e.g., CNN and Wachter et al. (2017)'s method) from a large size rule set, or local level from a small size rule set, although this is with the price of low accuracy.

Sensitivity analysis (SA). As mentioned before, the rationale of sensitivity analysis is to measure the relationship between the change of the input variable and change of output, rather than to measure the direct relationship between them. Therefore, it focuses on the intervention level of causation. However, it cannot answer intervention questions such as "what will happen if we double input variable X?" The explanations provided by simple linear regression analysis cannot be revealed in SA.

Feature Importance Ranking Measure (FIRM). Both FIRM and SA provide a relationship between input and output. However, they have different mechanisms. The feature importance in FIRM is estimated by its total impact on the prediction, while in SA, the importance is estimated by how the change of input influence the change of output. FIRM can also answer counterfactual questions while SA cannot. However, the biggest problem is that FIRM needs a very strong assumption which is difficult to satisfy: the distribution of input data should be a normal distribution.

Random forest (RF). The idea of variable importance in the random forest is very similar to variable importance in sensitivity analysis – measure the changes of output by randomly changing the value of a specific variable while controlling other variables constant. Therefore, random forest also focuses on the intervention level of causation. However, random forest also detects variable interactions, which is powerful than sensitivity analysis.

From mentioned above, Wachter et al. (2017)'s method can only answer the shallow level of counterfactuals causation, while CNN and LIME can provide a deeper explanation for counterfactuals questions. Rule-extraction methods are better than LIME and CNN in terms of counterfactuals. SA and random forest focus on intervention level of causation, but not complete. Random forest is more powerful than SA because it provides information about variable interaction. FIRM is the most powerful method to reveal causation in the data. However, the assumption limited generalizability too much (table 1). In conclusion, all these methods have drawbacks when revealing counterfactuals. They do not truly uncover causality-based explanation as described in Shmueli and Kopius (2011).

Methods	Correlation		Completeness	Complexity
	Intervention	Counterfactual		
CNN		**	**	***
LIME		**	**	***
FIRM		***	****	*

Rule extraction		***	***	**
SA	**		****	*
RF	***		****	*
Wachter et al. (2017)'s method		*	*	***

Table 1 Evaluation results of current XAI techniques

(because all methods meet the association level of causation, we don't show association in this table)

Level of completeness

Because we focus on explanation methods, therefore, the last three of six completeness criteria mentioned before are considered. Because the fifth completeness criterion is from social explanation perspective, we only consider the fourth and sixth completeness criteria. By providing global explanations, the confidence of each decision and explanation, and assumptions to use the method, SA, Random Forest and FIRM have the highest level of completeness. Rule extraction can provide both local and global explanation by adjusting the size of the rule set. However, the method to adjust the rule set is unsolved. Therefore, rule extraction has a relatively lower level of completeness. The lowest level of methods includes CNN and Wachter et al. (2017)'s method. Compared with Wachter et al. (2017)'s method, CNN provides a more detailed local explanation. Therefore, CNN has a higher level of completeness than Wachter et al. (2017)'s method (table 1).

Level of complexity

As mentioned before, a local-based explanation is more complex than a global-based explanation. Therefore, Wachter et al. (2017)'s method, LIME, and CNN have highest level complexity because they mainly provide an instance-based explanation. The next level is the rule extraction method because it can provide a local explanation with high accuracy, while provides a global explanation with low accuracy. FIRM, SA and Random Forest methods have the lowest level of complexity by providing a global level of explanation (table 1).

Discussion and conclusion

XAI is urgently needed in the current AI filed. As a response, a lot of new algorithms are developed recently. However, what is the standard criteria is an unsolved problem. Research about explanation from psychology, sociology, and cognition are borrowed to define and evaluate current XAI techniques. However, most of them are from the explainee's perspective. It is dangerous to do this because it is easy to fall into a persuasion explanation. In this paper, we proposed a new functional explanation framework, which is different from most existing explanation framework in XAI. Specifically, there are three factors to evaluate a functional explanation: correlation, completeness, and complex. The correlation makes sure that the information from explainer can fill the information gap between explainer and explainee. Pearl and Mackenzie' ladder of causation is used to measure the level of correlation. The information from explainer should complete so that it can describe the underlying systems. We proposed six criteria for a complete explanation. For explainee, the explanation should be as simple as possible, although this factor is less important than correlation and completeness. We then use these three factors to evaluate current XAI techniques. The findings suggested that in terms of correlation, rule-extraction method can provide the highest level of causation on counterfactuals level. Regrading completeness and complexity, FIRM, Random Forest and SA are best because they can provide global explanation.

As the most important evaluation criterion, the level of correlation shows that XAI has great advancement compared with black-box AI. Although XAI is far less than perfect, it is indeed on the right track. It is time to consider the combination and superposition effect of these different ideas and methods. Another direction to think about XAI is to develop more complicated causal chain like high-order factor analysis and complicated path analysis in SCM.

REFERENCES

Adadi, A., and Berrada, M. 2018. "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*. (6), pp. 52138–52160.

Bazerman, M. H., Moore D. A, Wiley J. and Sons. 2013. *Judgment in Managerial Decision Making*, 8th Edition, New York, ISBN: 978-1-118-06570-9.

Carroll, J., & Rosson, M. 1987. *Paradox of the active user*. In *J. M. Carroll (Ed.), Interfacing Thought: Cognitive Aspects of Human-Computer Interaction,* pp. 80–111.

Cortez, P. and Embrechts, M. J. 2013. "Using sensitivity analysis and visualization techniques to open black box data mining models," *Info. Sci.* (225), pp. 1–17.

Doshi-Velez, F. and Kim, B. 2017. "Towards A Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608.

Fong, R. C., Vedaldi, A, 2017. "Interpretable explanations of black boxes by meaningful perturbation," *IEEE Int Conf on Computer Vision*, pp.3429-3437. https://doi.org/10.1109/ICCV.2017.371

Freitas, A. A. 2014. "Comprehensible classification models: A position paper," *ACM SIGKDD Explor*. Newslett. (15 :1), pp. 1–10.

Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., and Giannotti, F. 2018. "A survey of methods for explaining black box models," *arXiv preprint arXiv:1802.01933*.

Gilpin, L.H, Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. 2018. "Explaining explanations: An approach to evaluating interpretability of machine learning," *arXiv:1806.00069*.

Hadji, I. and Wildes, R.P. 2018. "What do we understand about convolutional networks?" *arXiv preprint arXiv:1803.08834*.

Herman, B, 2017. "The promise and peril of human evaluation for model interpretability". *arXiv preprint arXiv:1711.07414*.

Hilton, D. J. 1990. "Conversational processes and causal explanation," *Psychological Bulletin*, 107(1):65–81.

Hilton, D. J. 1996. "Mental models and causal explanation: Judgements of probable cause and explanatory relevance," *Thinking & Reasoning*, 2(4):273–308.

Hoffman, R.R., Miller, T. Klein, G., & Litman, J. 2018. "Metrics for Explainable AI: Challenges and Prospects," A report on the DARPA Explainable AI Program, DARPA, Washington, DC.

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., and Clancey, W. J. 2018. "Explaining Explanation, Part 4: A Deep Dive on Deep Nets," *IEEE Intelligent Systems*, (33:3), 87-95.

Josephson, J. R., Josephson, S. G. 1996. Abductive inference: Computation, philosophy, technology, *Cambridge University Press*.

Koh, P. and Liang, P, 2017. Understanding black-box predictions via influence functions. In *Proc 34th Int Conf on Machine Learning*, pp.1885-1894.

Kulesza, T., Stumpf, S., Burnett, M. M., and Yang, S, 2013. "Too much, too little, or just right? Ways explanations impact end users' mental models," In *proceedings of the 2013 IEEE Symposium on Visual Languages and Human-Centric Computing*, pp. 3–10.

Kulesza, T., Burnett, M., Wong, W-K. and Stumpf, S. 2015. Principles of Explanatory Debugging to personalize interactive machine learning. *In: O. Brdiczka & P Chau (Eds.), Proceedings of the 20th International Conference on Intelligent User Interfaces*. (pp. 126-137). New York, USA: ACM. ISBN 9781450333061.

Lim, B., & Dey, A. 2009. "Assessing demand for intelligibility in context-aware applications". *In Proc. Ubicomp,* pp. 195–204.

Lipton, P. 1990. "Contrastive explanation," *Royal Institute of Philosophy Supplement* (27), pp. 247-266.

Lombrozo, T. 2012. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pp. 260–276.

Martens, D. and Provost, F. 2014. Explaining data-driven document classifications. *MIS Quarterly* (38:1). pp. 73-99.

Miller, T. 2018. "Explanation in Artificial Intelligence: Insights from the Social Sciences," *ArXiv:1706.07269* [Cs]. Retrieved from http://arxiv.org/abs/1706.07269

Miller, T., T., Howe P., and Sonenberg, L. 2017. "Explainable AI: Beware of inmates running the asylum," *in Proceedings of the IJCAI Workshop on Workshop on Explainable Artificial Intelligence*.

Mueller, S.T., Hoffman R.R., Clancey, W, Emrey, A, and Klein, G. 2019. "Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI," *DARPA XAI Program*. arXiv:1902.01876.

Pearl, J. 2018. "Theoretical impediments to machine learning with seven sparks from the causal revolution," *arXiv preprint arXiv:1801.04016*.

Pearl, J. and Mackenzie, D. 2018. The Book of Why: The New Science of Cause and Effect, Hachette UK.

Ras, G, Gerven, M. van, and Haselager, P. 2018. "Explanation methods in deep learning: Users, values, concerns and challenges". Available: https://arxiv.org/abs/1803.07517

Read, S. J., Marcus-Newhall, A. 1993. "Explanatory coherence in social explanations: A parallel distributed processing account," *Journal of Personality and Social Psychology* (65: 3), pp. 429-447.

Ribeiro, M.T., Singh, S., and Guestrin, C. 2016. "Why Should I Trust You?: Explaining the Predictions of Any Classifier," In *SIGKDD*, 2016. 3, 6.

Rosson, M., Carrol, J., & Bellamy, R. 1990. "Smalltalk scaffolding: a case study of minimalist instruction," In *Proc. CHI*, pp. 423–430.

Selvaraju, R.R., Cogswell, M., Das, A., et al. 2017. "Grad-CAM: visual explanations from deep networks via gradient based localization," *IEEE Int Conf on Computer Vision*, pp.618-626. https://doi.org/10.1109/ICCV.2017.74.

Shmueli, G. and Koppius, O.R. 2011. "Predictive analytics in information systems research," *MIS Quarterly* (35: 3), pp. 553-572.

Silva, A, Cortez, P., Santos, M. F., Gomes, L., and Neves, J. 2006. "Mortality assessment in intensive care units via adverse events using artificial neural networks," *Artificial Intelligence in Medicine*, (36:3), pp. 223-234.

Su, J, Vargas, D. V., Kouichi, S, 2017. "One pixel attack for fooling deep neural networks," http://arxiv.org/abs/1710.08864.

Slugoski, B. R, Lalljee, M, Lamb, R, and Ginsburg, G.P. 1993. "Attribution in conversational context: Effect of mutual knowledge on explanation-giving," *European Journal of Social Psychology*, (23:3), pp. 219–238.

Thagard, P, 1989. "Explanatory coherence," *Behavioral and Brain Sciences* (12: 03), pp. 435-467.

Van Lent, M., Fisher, W., Mancuso, M. 2004. "An explainable artificial intelligence system for small-unit tactical behavior," In: PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press. pp. 900–907.

Wachter, S, Mittelstadt, B, and Russell, C. 2017. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," arXiv preprint arXiv:1711.00399.

Weller, A. 2017. "Challenges for transparency," *Workshop on Human Interpretability in Machine Learning – ICML 2017*.

Zeiler, M.D. and Fergus, R, 2014. "Visualizing and understanding convolutional networks," *European Conf on Computer Vision*, pp.818-833.

Zeiler, M.D. and Fergus, R. 2013. "Stochastic pooling for regularization of deep convolutional neural networks," In *International Conference on Learning Representations 2013*.

Zeiler, M.D., Taylor, G.W. and Fergus, R. 2011. "Adaptive deconvolutional networks for mid and high level feature learning," In *ICCV*, pp. 2018 - 2025.

Zien, A., Kramer, N., Sonnenburg, S., and Ratsch, G. 2009. "The feature importance ranking measure," In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 694–709.

Zhang, Q., Cao, R, Wu, Y. N, et al., 2016. "Growing interpretable part graphs on convnets via multi-shot learning," In *Proc 30th AAAI Conf on Artificial Intelligence*, pp.2898-2906.

Zhang, Q., Cao, R, Shi, F, et al., 2018b. "Interpreting CNN knowledge via an explanatory graph," In *Proc 32nd AAAI Conf on Artificial Intelligence*, pp.2124-2132.

Zhang, Q., Wang, W, Zhu, S., 2018a. "Examining CNN representations with respect to dataset bias," In *Proc 32nd AAAI Conf on Artificial Intelligence*, in press.

Zhang, Q., Wu, Y., Zhu, S., 2018. "Interpretable convolutional neural networks," *Proc IEEE Conf on Computer Vision and Pattern Recognition*, in press.

Zhang, Q. and Zhu, S., 2018. "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology & Electronic Engineering* (19:1), pp. 27-39.