Building More Explainable Artificial Intelligence with Argumentation

Zhiwei Zeng[1], Chunyan Miao[1], Cyril Leung[2], Chin Jing Jih[3]
[1]Nanyang Technological University, [3]Tan Tock Seng Hospital, Singapore
[2]The University of British Columbia, Vancouver, Canada
i160001@e.ntu.edu.sg

Abstract

Currently, much of machine learning is opaque, just like a "black box". However, in order for humans to understand, trust and effectively manage the emerging AI systems, an AI needs to be able to explain its decisions and conclusions. In this paper, I propose an argumentation-based approach to explainable AI, which has the potential to generate more comprehensive explanations than existing approaches.

Introduction

The advent of big data era has brought great success to machine learning. Together, the abundance of data and the development of various machine learning techniques has led to an explosion of new AI models and applications. However, much of machine learning still remains opaque as a "black box". The effectiveness of AI systems, especially in critical applications such as disease diagnosis, stock trading and autonomous vehicles, will be limited by the machines' inability to explain their decisions and conclusions to humans. Thus, it is important to build more explainable AI, so that humans can understand, trust and effectively manage the emerging AI systems (Gunning 2016).

Integrating the taxonomies proposed in (Gunning 2016) and (Biran and Cotton 2017), existing research in explainable AI (XAI) can be categorized into three broad approaches: (1) *explanations based on features*, (2) *model approximation* and (3) *interpretable models*.

For *explanations based on features*, usually a non-interpretable complex model and its predictions are given. This approach focuses on generating justifications for the predictions by extracting and identifying the features that have significant effects on the prediction results. Martens et al. (2008) explain the results of an SVM classifier by extracting rules that can produce similar results to the SVM based on a small subset of features. Landecker et al. (2013) interpret the classification results of hierarchical networks by studying the degree of importance of different components to the classification results. Hendricks et al. (2016) generate explanations for image classification results of a CNN using a LSTM, based on both prominent image features and class discriminative features.

The second approach, *model approximation*, involves model-agnostic methods that infer an explainable model from any black-box models. Robnik-Šikonja and Kononenko (2008) decompose a model's prediction to the level of individual features by comparing the model results when a feature value is present and absent. More recently, Ribeiro, Singh, and Guestrin (2016) explain a prediction instance by constructing interpretable model locally around it, which is only an accurate approximation of the global model in the vicinity of that instance.

The third approach, *interpretable model*, aims to construct models that are inherently structured and interpretable, such as rule lists and decision trees. Si and Zhu (2013) use And-Or-Trees to represent the possible component structures of objects in images, which can be used as compositional models for explaining results of object detection. Lake, Salakhutdinov, and Tenenbaum (2015) learn a generative model of character images, and explain a character recognition result using the generation process of that character.

Despite much work having been done in this field, I have identified several gaps that need to be bridged. Firstly, existing XAI models can answer the question "why this decision or conclusion", while they cannot answer "**Why not**". Secondly, most of them offer explanations as a form of evidences rather than reasons. Good explanations need to reveal the underlying reasoning process and are best presented in human-interpretable terms. To fill these gaps, I decide to take an argumentation-based approach to XAI, which has the potential to generate more comprehensive explanations.

Progress to Date

Argumentation is the study of how reasonable decisions or conclusions can be reached by constructing for and against arguments and evaluating these arguments accordingly. Argumentation-based approach to decision making is expected to be more akin with the way humans deliberate, evaluate alternatives and make decisions. This endows argumentation-based approach unique benefits, including transparent decision making process and the ability to offer understandable reasons underlying the decisions made. Argumentative explanations have also attracted increasing research interests in recent years. Fan and Toni (2014) study argumentative explanations for acceptable decisions or conclusions while Fan and Toni (2015) focus on generating ex-

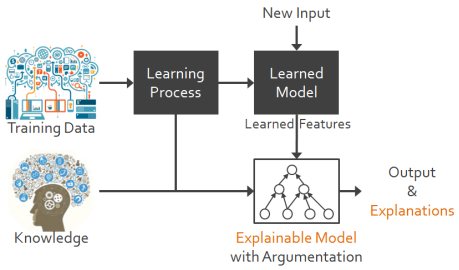

Figure 1: The Conceptual Framework of Proposed Approach

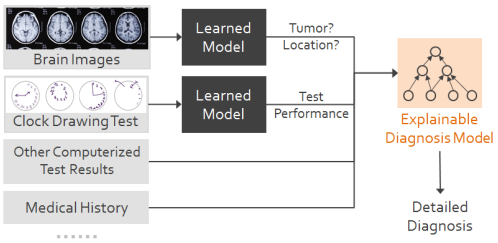

Figure 2: Application of the Proposed Approach in Early Detection of Dementia

planations for non-acceptable ones.

Inspired by these works, I am currently working towards a new argumentation-based approach to XAI. The conceptual framework of my proposed approach is illustrated in Fig 1. Instead of using only a black box model, an explainable model is created to perform reasoning based on the latent patterns learned and to generate explanations for the output. It would fall into the third approach reviewed above —*interpretable model*. As a concrete example, Fig 2 shows the application of the proposed model in early detection of dementia. Machine learning techniques for computer vision, such as CNN, can be used to extract visual features from brain images and clock drawing tests. The learned features, such as size and location of tumours and drawing test performance, together with other computerized test (Zeng et al. 2017) results and medical history, will be used as the inputs to an explainable diagnosis model. The model will not only be able to present a positive or negative diagnosis, but also detailed reasons leading to the diagnosis.

As the first step to realize the proposed approach, I have came up with an explainable model for making context-based decisions. In (Zeng et al. 2018), I presented a graphical representation for modelling decision problems involving varying contexts, Decision Graph with Context (DGC), and a reasoning mechanism for making context-based decisions which relies on the Assumption-based Argumentation formalism. Based on these constructs, I formalized two types of explanations with their computation, *argument explanation* and *context explanation*, identifying the reasons for decisions made from an argument-view and a context-view respectively.

Future Work

In the future, I would like to further explore the properties of different argumentative explanations. This may include but is not limited to the selection of the best explanation from a set of available explanations and the evaluation of the explanations generated. Then, I hope to apply the proposed approach in applications such as the early detection of dementia and building an explainable model for computational persuasion. In order to do so, I plan to identify suitable machine learning techniques based on the application contexts and find a viable way to integrate them with the explainable decision making model I developed.

Acknowledgement

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative, and the Singapore MOH under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017).

References

Biran, O., and Cotton, C. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, 8–13.

Fan, X., and Toni, F. 2014. On computing explanations in abstract argumentation. In *Proc. of ECAI 2014*, 1005–1006. IOS Press.

Fan, X., and Toni, F. 2015. On explanations for non-acceptable arguments. In *Proc. of TAFA 2015*, 112–127. Springer.

Gunning, D. 2016. Explainable artificial intelligence (xai). *Technical report, DARPA/I2O*.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *European Conf. on Computer Vision*, 3–19. Springer.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.

Landecker, W.; Thomure, M. D.; Bettencourt, L. M.; Mitchell, M.; Kenyon, G. T.; and Brumby, S. P. 2013. Interpreting individual classifications of hierarchical networks. In *2013 IEEE Symp. on Computational Intelligence and Data Mining*, 32–38. IEEE.

Martens, D.; Huysmans, J.; Setiono, R.; Vanthienen, J.; and Baesens, B. 2008. Rule extraction from support vector machines: an overview of issues and application in credit scoring. *Rule extraction from support vector machines* 33–63.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Robnik-Šikonja, M., and Kononenko, I. 2008. Explaining classifications for individual instances. *IEEE Trans. Knowl. Data Eng.* 20(5):589–600.

Si, Z., and Zhu, S.-C. 2013. Learning and-or templates for object recognition and detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(9):2189–2205.

Zeng, Z.; Miao, C.; Leung, C.; and Shen, Z. 2017. Computerizing trail making test for long-term cognitive self-assessment. *International Journal of Crowd Science* 1(1):83–99.

Zeng, Z.; Fan, X.; Miao, C.; Wu, Q.; and Cyril, L. 2018. Context-based and explainable decision making with argumentation. In *AAMAS-18 (submitted)*. IFAAMAS.