The XAI Paradox: systems that perform well for the wrong reasons

Cor Steging, Lambert Schomaker, and Bart Verheij

Department of Artificial Intelligence, Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen

Many of the most commonly employed machine learning techniques inherently lack a possibility to provide an explanation of the way in which they reason. The 'black box' character of such systems can lead to distrust by users because the systems cannot explain why particular decisions are made [3].

The response to this lack of transparency is the emerging field of Explainable Artificial Intelligence (XAI) [4]. XAI aims to create explainable models that are able to explicitly describe the inner workings of a black box machine learning system, and thus provide the essential explanations to their users. Examples of such XAI systems are rule extraction algorithms, that create sets of rules for the users that describe the reasoning of black box systems [6], or glass-box systems, in which the users are able to control what the systems learn [5]. A recent deep learning system that is used to diagnose and refer in retinal diseases (and outperforms medical experts), is divided into a framework of smaller systems; one for each stage of the diagnostic process [2]. This makes it easier for the clinicians to investigate and understand the reasoning of the system. The overall goal of XAI systems is therefore to yield accurate explanations of the reasoning of the systems, without sacrificing performance.

Whether or not the explanation that an XAI system gives would make sense, however, is not certain. The internal rationale that a black box system uses may yield high performance results, even though that rationale is not sound [1]. A good example of this phenomenon is the use of adversarial examples in image classification, in which an input image is altered ever so slightly using a perturbation [7]. To the human eye, there is no apparent difference between the original input image and its altered version, while a machine learning system will make a completely different classification after the image is altered. This suggests that the reasoning that the systems employ differs wildly from the reasoning that humans would use, despite the fact that the system usually makes the correct decisions. It performs well, but for the wrong reasons. In and of itself, this is impressive, as the machine learning system is able to extrapolate structures from the data that we as humans are not able to perceive, and consequently achieve a high performance. The result, however, is that when an explainable model

This is an extended abstract of a paper invited for submission to the post-proceedings (see https://www.ai.rug.nl/~verheij/publications/bnaic2019.htm).

is generated from these systems using XAI techniques, the explanation that is generated for the user is incorrect as well. Because an unsound explanation cannot provide the user with the desired information and trust, systems are needed that can be explained in a way that is aligned with a user's needs.

The aim of this study is to investigate machine learning techniques in terms of how well they are able to learn the structure that defines the dataset, rather than in terms of a general performance accuracy. Artificial datasets are generated from a set of predetermined conditions on which machine learning systems are trained. The rationale of the trained systems is compared to the original set of conditions in order to measure how well the systems are able to extrapolate the correct conditions from the data. We find that when multiple conditions define the dataset, the system does not learn each of the individual conditions correctly, but instead learns a completely different confounding structure that accurately maps the input to the output. This interaction effect is further investigated.

In real world AI systems, machine learning algorithms that only execute a task with a high performance do not always suffice; an explanation of their decision making will be required [4]. This explanation, however, is dependent on the reasoning of a system; if the reasoning is unsound, the explanation will be unsound as well. This study shows that machine learning algorithms do not always internalize the structure of their training data as we would expect. With regards to the emerging XAI techniques and the explainable models that they generate, the unsound rationales of machine learning systems can form a hindrance in creating an understandable explanation.

References

1. Bench-Capon, T.J.M.: Neural networks and open texture. In: Proceedings of the Fourth International Conference on Artificial Intelligence and Law, pp. 292–297. ACM Press, New York (New York) (1993)
2. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., other authors: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine **24**, 1342–1350 (2018)
3. Edwards, L., Veale, M.: Slave to the algorithm? Why a 'Right to explanation' is probably not the remedy you are looking for. Duke Law & Technology Review **16**, 18–84 (2017)
4. Gunning, D.: Explainable artificial intelligence (XAI) (2017), `https://www.darpa.mil/attachments/XAIProgramUpdate.pdf`
5. Holzinger, A., Plass, M., Holzinger, K., Crisan, G.C., Pintea, C., Palade, V.: A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop (2017), `https://arxiv.org/abs/1708.01104`
6. Lu, H., Setiono, R., Liu, H.: Neurorule: A connectionist approach to data mining. In: Proceedings of the 21th International Conference on Very Large Data Bases (VLDB '95), pp. 478–489. Morgan Kaufmann, San Francisco (California) (1996)
7. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems **30**(9), 2805–2824 (2017)