Explaining Deep Neural Networks in medical imaging context

RGUIBI Zakaria[#1], HAJAMI AbdelMajid[#2], DYA Zitouni[#3]

[#]*Hassan First University of Settat, Faculty of Science and Technology ,Laboratory for Emerging Technologies (la-VETE), Moroccos*

[1]rguibi.fst@uhp.ac.ma, [3]zitouni.dya@uhp.ac.ma

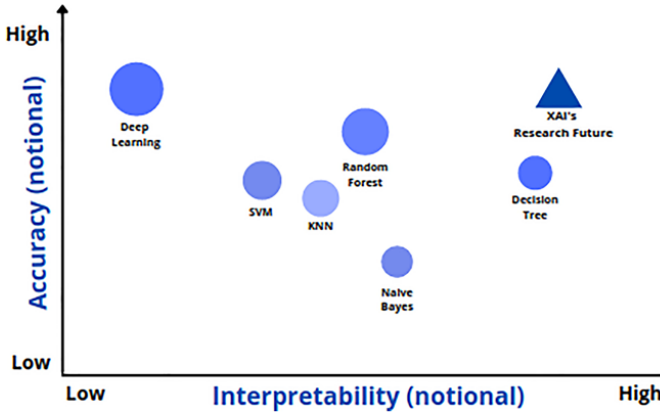
[2]abdelmajid.hajami@uhp.ac.ma

***Keywords*— Decision-making Processes, Deep Neural networks, Explaining Neural Models, Medical imaging.**

Introduction

Deep neural networks are becoming more and more popular due to their revolutionary success in diverse areas, such as computer vision, natural language processing, and speech recognition. However, the decision-making processes of these models are generally not interpretable to users. In various domains, such as healthcare, finance, or law, it is critical to know the reasons behind a decision made by an artificial intelligence system. Therefore, several directions for explaining neural models have recently been explored.

In this communication, We investigate The second major direction for explaining deep neural networks that consist of self-explanatory neural models that generate medical imaging explanations, that is, models that have a built-in module that generates explanations for the predictions of the model.



In the literature, a variety of terms exist to indicate the opposite of the “black box” nature of some of the AI and ML, and especially DL, models. We distinguish the following terms:

Interpretability: It is defined as the ability to explain or to provide the meaning in understandable terms to a human.

Explainability: Explainability is associated with the notion of expla- nation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.

Transparency: A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models in Section 3 are divided into three categories: simulatable models, decomposable models and al- gorithmically transparent models [5] .

A. Feature-based post-hoc explanatory methods

Post-hoc explanatory methods are stand-alone methods that aim to explain already trained and fixed target models. These methods can potentially develop meaningful insights about what exactly a model learnt during the training.

Most of the post-hoc models like attributions can also be seen as model agnostic as these methods are typically not dependent upon the structure of a model. However, some requirements regarding the limitations on model layers or the activation functions do exist for some of the attribution methods. There are broadly two types of approaches to explain the results of deep neural networks (DNN) in medical imaging - those using standard attribution-based methods and those using novel, often architecture or domain-specific techniques.[1]

The problem of assigning an attribution value or contribution or relevance to each input feature of a network led to the development of several attribution methods. The goal of an attribution method is to determine the contribution of an input feature to the target neuron which is usually the output neuron of the correct class for a classification problem. The arrangement of the attributions of all the input features in the shape of the input sample forms heatmaps known as the attribution maps.[1]

B. Self-explanatory neural models

Self-explanatory models are target models which incorporate an explanation generation module into their architecture such that they provide explanations for their own predictions.

At a high level, self-explanatory models have two interconnected modules: (i) a predictor module, i.e., the part of the model that is dedicated to making a prediction for the task at hand, and (ii) an explanation generator module, i.e., the part of the model that is dedicated to providing the explanation for the prediction made by the predictor. For example, Lei et al. [4] introduced a self-explanatory neural network where the explanation generator selects a subset of the input features, which are then exclusively passed to the predictor that provides the final answer based solely on the selected features. Their model is also regularised such that the selection is short. Thus, the selected features are intended to form the explanation for the prediction. Self-explanatory models do not necessarily need to have supervision on the explanations. [2]

In general, for self-explanatory models, the predictor and explanation generator are trained jointly, hence the presence of

the explanation generator is influencing the training of the predictor. This is not the case for post-hoc explanatory methods, which do not influence at all the predictions made by the already trained and fixed target models. Hence, for the cases where the augmentation of a neural network with an additional explanation generator results in a significantly lower task performance than that of the neural network trained only to perform the task, one may prefer to use the latter model followed by a post-hoc explanatory method. On the other hand, it can be the case that enhancing a neural network with an explanation generator and jointly training them results in a better performance on the task at hand. This can potentially be due to the additional guidance in the architecture of the model, or to the extra supervision on the explanations if available.[2]

Interpretability or lack thereof can limit the adoption of machine learning methods in decision-critical —e.g., medical or legal— domains. Ensuring interpretability would also contribute to other pertinent criteria such as fairness, privacy, or causality. Our focus in this paper is on complex self-explaining models where interpretability is built-in architecturally and enforced through regularization. Such models should satisfy three desiderata for interpretability: explicitness, faithfulness, and stability where, for example, stability ensures that similar inputs yield similar explanations.[6]

Most post-hoc interpretability frameworks are not stable in this sense. High modeling capacity is often necessary for competitive performance. For this reason, recent work on interpretability has focused on producing a posteriori explanations for performance-driven deep learning approaches. The interpretations are derived locally, around each example, on the basis of limited access to the inner workings of the model such as gradients or reverse propagation, or through oracle queries to estimate simpler models that capture the local input-output behavior [16, 2, 14]. Known challenges include the definition of locality (e.g., for structured data), identifiability and computational cost (with some of these methods requiring a full-fledged optimization subroutine). [6]

However, point-wise interpretations generally do not compare explanations obtained for nearby inputs, leading to unstable and often contradicting explanations. A posteriori explanations may be the only option for already-trained models. Otherwise, we would ideally design the models from the start to provide human-interpretable explanations of their predictions. In this work, we build highly complex interpretable models bottom up, maintaining the desirable characteristics of simple linear models in terms of features and coefficients, without limiting performance. [6]

For example, to ensure stability (and, therefore, interpretability), coefficients in our model vary slowly around each input, keeping it effectively a linear model, albeit locally. In other words, our model operates as a simple interpretable model locally (allowing for point-wise interpretation) but not globally (which would entail sacrificing capacity). We achieve this with a regularization scheme that ensures our model not only looks like a linear model, but (locally) behaves like one.[6]

C. General Conclusions and Perspectives

In the last few years, opening the "black box" is critically important not only for acceptability within the society but also for regulatory purpose. As black box Machine Learning (ML) models are increasingly being employed to make important predictions in critical contexts like healthcare, the demand for transparency is increasing from various stakeholders in AI the danger is on creating and using decisions that are not justifiable, legitimate, or that simply do not allow obtaining detailed explanations of their behavior. Explanation supporting the output of a model is crucial, e.g., in precision medicine, where experts require far more information from the model than a simply binary prediction for supporting their diagnosis. [3][5]

The most recent work on the interpretability of complex machine learning models has focused on estimating a posteriori explanations for previously trained models around specific predictions. Self-explaining models where interpretability plays a key role already during learning have received much less attention.[6]

References

[1] Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable Deep Learning Models in Medical Image Analysis. *J. Imaging* **2020**, *6*, 52. https://doi.org/10.3390/jimaging6060052.

[2] Camburu, OM. n.d. “Explaining Deep Neural Networks.” PhD thesis, University of Oxford.

[3] Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424. https://doi.org/10.1002/widm.1424

[4] Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 107–117.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, 2020, Pages 82-115, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.12.012.

[6] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 7786–7795.