Explaining Deep Neural Networks in a 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), Morocco*

[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 first major direction for explaining deep neural networks direction consists of feature-based post-hoc explanatory methods, that is, methods that aim to explain an already trained and fixed model (post-hoc), and that provide explanations in terms of input features, such as superpixels for images (feature-based).

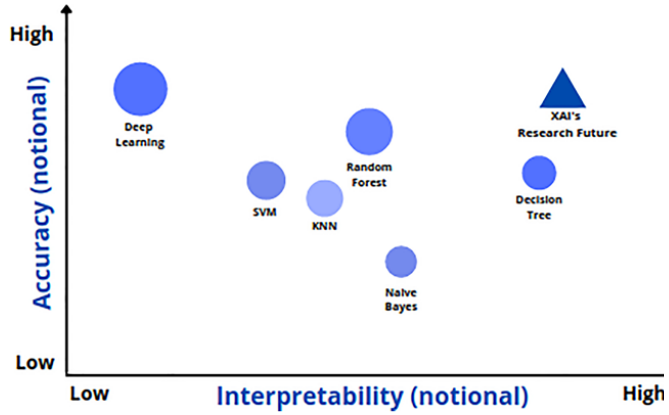

figure1: WIREs Data Mining Knowl Discov, Volume: 11, Issue: 5, First published: 12 July 2021, DOI: (10.1002/widm.1424)

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 explanation 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 algorithmically transparent models.

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 learned 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]

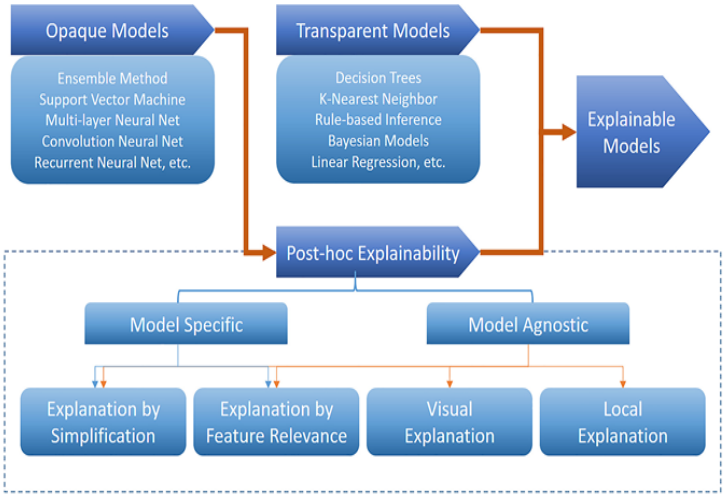

figure2: WIREs Data Mining Knowl Discov, Volume: 11, Issue: 5, First published: 12 July 2021, DOI: (10.1002/widm.1424)

1. Explanation by simplification:

refer to the techniques that approximate an opaque model using a simpler one, which is easier to interpret. The main challenge comes from the fact that the simple model has to be flexible enough so it can approximate the complex model accurately. In most cases, this is measured by comparing the accuracy (for classification problems) of these two models.

2. Explanation by feature Relevance:

attempt to explain a model's decision by quantifying the influence of each input variable. This results in a ranking of importance scores, where higher scores mean that the corresponding variable was more important for the model. These scores alone may not always constitute a complete explanation, but serve as a first step in gaining some insights about the model's reasoning.

3. Visual Explanation:

aim at generating visualizations that facilitate the understanding of a model. Although there are some inherit challenges (such as our inability to grasp more than three dimensions), the developed approaches can help in gaining insights about the decision boundary or the way features interact with each other. Due to this, in most cases, visualizations are used as complementary techniques, especially when appealing to a non-expert audience.

4. Local Explanation :

attempt to explain how a model operates in a certain area of interest. This means that the resulting explanations do not necessarily generalize to a global scale, representing the model's overall behavior. Instead, they typically approximate the model around the instance the user wants to explain, in order to extract explanations that describe how the model operates when encountering such instances.

B. example of Explaining Deep Neural Networks in a medical imaging context.

In this example, the classifier is trained on the LPBA40 dataset for single-label classification of healthy vs. pathological. The results are evaluated in a 4-fold cross-validation manner over the patients, meaning that per patient there is one healthy and four pathological images in the training and testing datasets. On test data, the classification achieves AUC of 1, which intuitively means this is a perfect classifier for this problem. Here we also create explanation maps using all methods and train the autoencoder for VAE-perturbation with the original healthy LPBA40 images. The qualitative evaluation of this examples (Fig. 5) interestingly reveals, that none of the explanation methods, is able to reliably detect Lesion 4, leading to the conclusion that the learned classification of those images is not necessarily correlated to the pathology presence. One possible explanation is that the CNN learns to classify images with stronger gray value variability as pathological. Again guided backpropagation turns out to be noisy and not class-discriminative. Note, that the same few regions are highlighted for the different lesions leading to the conclusion, that the network learns the possible location of the structures and only observes those places. This is of course based on the small variability of lesions and different images we use for this experiment. Here gradCAM generally detects the rough location of the pathologies, but its explanations have bad resolution, which in the case of the smaller lesions (e.g. Lesion 3) leads to bad results.[6]

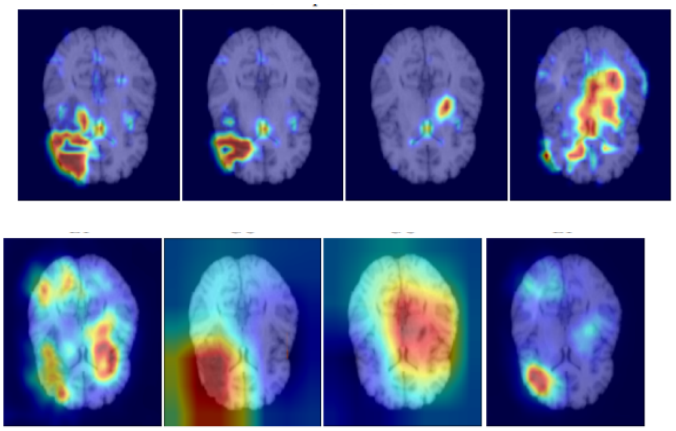

figure 3: Different explanation techniques for the single-label LPBA40 lesion classifier.

The first row of Fig. 4 shows an example of a chest X-ray truly classified as positive for tuberculosis. Te Chest X-ray shows patchy opacities in the right upper lobe with pleural apical thickening and upward deviation of the right hilum. These findings are consistent with pulmonary tuberculosis. In the saliency map, the outline of the soft tissue structures of the mediastinum are highlighted, and especially the area of the right upper lobe. This correlates perfectly with the pathological changes seen in the X-ray image.

The second row of Fig. 4 gives an example of the chest X-ray of a healthy patient, which was correctly identified as negative for tuberculosis. The saliency map (panel f) shows symmetric high lightening of the borders of the mediastinum. There is no increased signal in any of the lobes of the lung.[7]

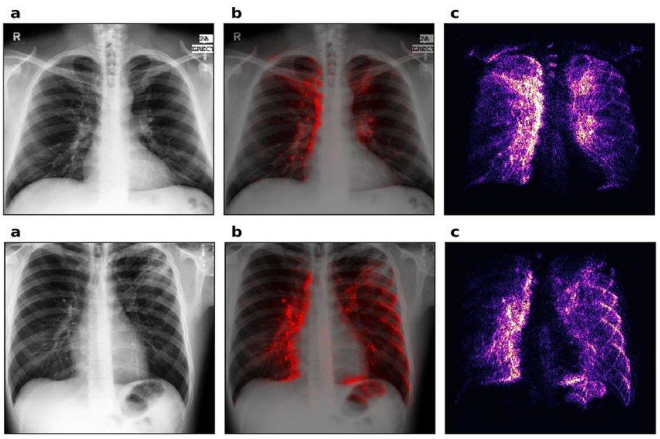

Figure 4. Saliency map with overlay for two correctly classified cases. Panels (a) and (d) show the chest images of the patients, panels (c) and (f) show the saliency maps, while panels (b) and (e) show the saliency maps overlaid on the chest images for comparison. The frst row shows a patient with tuberculosis, with an output score 0.98 (the maximum was 1). The second row shows a healthy patient with a score 0.00 (the minimum was 0). Both scores suggest high confidence in the prediction.

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 is 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 behaviour. 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]

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] Hristina Uzunova, Jan Ehrhardt, Timo Kepp, Heinz Handels, "Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders," Proc. SPIE 10949, Medical Imaging 2019: Image Processing, 1094911 (15 March 2019) https://doi.org/10.1117/12.2511964.

[7] Pasa, F., Golkov, V., Pfeiffer, F. et al. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. Sci Rep 9, 6268 (2019). https://doi.org/10.1038/s41598-019-42557-4