A survey of Explainable AI and Out-of-Distribution detection methods

Jonatan Hoffmann Hanssen
Master Student, Dept. of Informatics
The Faculty of Mathematics
and Natural Sciences
Oslo, Norway
jonatahh@ifi.uio.no

***Abstract*—This essay covers a wide range of contemporary methods within the fields of Explainable Artificial Intelligence (XAI) and Out-of-Distribution (OOD) detection.**

***Index Terms*—explainable artificial intelligence, out of distribution detection, data outlier detection**

I. Introduction

Machine Learning generally, and Deep Learning specifically, have seen a tremendous increase in performance in recent years, performing comparable to humans in many tasks, for example image classification, speech recognition and many others. CITE SOMEONE. In medicine, deep learning has the potential to provide faster and more accurate detection of diseases by being trained on thousands of previous patients. CITE.

However, deep learning methods are not without their flaws. Firstly, deep neural networks are inherently unexplainable, due to the large number of parameters that any non-trivial network has. State of the art models will perform millions of operations to evaluate a single data point, and it is therefore impossible for humans to comprehend and explain the process which lead the model to make a particular decision. In medicine, this is a major limitation of deep learning methods, as both doctors and patients expect to be able to understand why a decision was made. Secondly, although neural networks may attain high accuracy on test data and appear to have learned great insights about the tasks they are employed in, they often lack robustness and can suffer large drops in performance on data points which are slightly different from the training data. As AUTHOR has shown, it is possible to create data points which are imperceptibly different from normal data points, yet still fool otherwise high performing models.

These two problems lead to the two fields of machine learning which I shall discuss in this essay. These are Explainable Artificial Intelligence (XAI), and Out-of-Distribution (OOD) detection.

II. Explainable AI

Below follows a thourough introduction to XAI, as well as detailed look at some important methods for explainability for neural networks applied to images.

A. The motivation for XAI

Given the impressive performance of DL methods, one might be convinced that these models do not need to be explainable or interpretable, and that we instead should just place our faith in the model without knowing exactly how it came to a decision. However, as [Doshi-Velez and Kim, 2017] points out, ”a single metric, such as classification accuracy, is an incomplete description of most real-world tasks”. Even if your model attains a 99.9% accuracy on your test data, this does not mean that it will perform as well whenever it is deployed in a real setting. Small differences between the data distribution when the test data was collected and when the model is deployed may have a large impact on the model's performance, or the model may have learned artifacts or specificities in the training dataset which were also present in the test dataset, leading to a false belief that the model has gained generalizable knowledge about the problem at hand. By using explainable methods, we may reveal these shortcomings.

XAI is also important whenever the model is used in settings where its decisions have a high impact. If a model is used by a hospital in disease detection, both the patient and doctor will probably want to be able to understand why the model has found that a disease is present. As [van der Velden et al., 2022] states, ”for the regulated healthcare domain, it is utmost important to comprehend, justify, and explain the AI model predictions for a wider adoption of automated diagnosis”. Furthermore, the right to an explanation of an automated decision affecting a person is included in the EU's General Data Protection Regulation, which states that ”In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right [...] to obtain an explanation of the decision reached after such assessment and to challenge the decision.” [European Union, 2016].

B. Overview of methods

1) Intrinsically explainable models: Intrinsically explainable models are models which have sufficiently low complexity, such that it is feasible for a human to understand them without further modifications. Examples of such methods are linear regression, logistic regression and decision trees [Molnar, 2022].

2) Post hoc methods: Post hoc methods are methods which are applied to the model after training. These methods do not aim to constrain the model to be interpretable, but inspect the model after training.

3) Model agnostic / model dependent methods: Model agnosticity/dependence denotes whether an XAI method uses

C. Specific methods

1) Class Activation Mapping (CAM): CAM [**?**] is a model dependent, post hoc XAI method, which is used on Convolutional Neural Nets (CNNs). For a specific output node of a model (for example, the one denoting the presence of a specific class, such as "cat"), CAM outputs a heat map showing which areas of the input image contributed to this node. In this way, CAM gives a visual explanation to which parts of an image the model focused on when making a decision to classify an image to a specific class. This method is model dependent, because it requires a specific architecture in the final layers of the network to work.

CAM is a relatively simple method to understand.

2) Gradient Class Activation Mapping (Grad-CAM):

3) Layer-Wise Relevance Propagation:

4) :

III. Out-of-Distribution Detection

References

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

[European Union, 2016] European Union (2016). Article 71: European data protection board. Accessed: February 13, 2024.

[Molnar, 2022] Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

[van der Velden et al., 2022] van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470.