Modern Explainability Techniques for Today's Complex Model Architecture

Vidyalakshmi Venkataraman
http://orcid.org/0009-0001-9098-7974
Pondicherry University, India

Sukhvinder Singh
Pondicherry University, India

Ravi Singh
http://orcid.org/0009-0005-1714-5120
Pondicherry University, India

ABSTRACT

AI has grown phenomenally and has created space for itself in every industry domain with the increasing technological advances in deep learning (CNN, RNN, GAN, Variational Auto-Encoders). As the ability to solve complex problems increases, the architecture of the model also gets complicated, making it directly proportional to the ability of solving complex problems. The model turns to be a black-box not allowing humans or end users to understand the reason of why the model gave a particular output. This has led to the need to make models more trustworthy and dependable on its output which was only possible when the black-box was converted to a white box – making the human understand what the model is learning at the intermediate layers of the architecture. Explainable AI is the emerging field of study which makes the model interpretable. While there are numerous methods identified, this article presents a detailed classification and grouping of methods based on domain of use and mode of explanation and detailed explanation of 10 methods.

INTRODUCTION

In the growing age of Artificial Intelligence (AI), many new methods or algorithms or concepts have taken a prominent place in various industries ranging from healthcare to autonomous vehicles which has significantly revolutionized and has bought about innovation in various sectors. Whenever there is a growth in any industry, there is always a challenge behind that must be solved. The challenge that mainly

DOI: 10.4018/405391

an industry face is the understanding that how these machines work, what are the inner happenings of this model and how this model can bring about a prediction of this kind.

It is indispensable to know that any industry is a making a significant growth using machine learning and deep learning models. The deep learning (DL) models like convolutional neural network (CNN) and recurrent neural networks (RNN) are extremely difficult to understand as they involve huge amounts of data and has numerous neural network layers. Though these are exceptionally all-encompassing models which has the ability to learn intricate patterns and representations from datasets, it still comes with a huge cost making it difficult to understand the working of the model, which is referred to as "Black Box", which completely hides the explainability and interpretability of the model's decision-making process.

Considering the concept of Deep Networks (DN) which mainly excels in image recognition tasks (Generation/Classification) and Natural Language Processing (NLP) tasks, involving multiple hidden layers and requiring a huge data, the model architecture remains a mystery for a human to understand how the output decision is given, which parameters or the features of which hidden layer contributed to the model in prediction. This is continued by addressing the problems of Ensemble models, where more than one model is considered to improve the accuracy of the model by reducing the errors. The challenge here is to understand which model in the ensemble has contributed more on the decision making or contributing for the prediction.

Generative models like GAN, transformers, BERT, variational auto-encoders etc. are also a type of deep neural network that is mainly used to build and learn relationship between words and sentences or images and its associated features. These are the models which are mainly used in GPT. Because of their extensive parameterization and non-linearity between the words and sentences or interactions in general, they are highly difficult to interpret.

With the growing research, while we have made significant stride in getting great efficiency on model performance, the model remains prone for adversarial attacks. There is a need to make these models interpretable and ensure the output given is without any bias.

The increasing dependability on these tools and architectures make these models a black box which cannot be directly interpreted or explained for the reasons of the outcome.

Below are listed some of the challenges given by complex models making it to be termed as black box in an analogy of the flight black box:

1. **Opaque in Nature** – AI technologies being applied in any domain of work like healthcare, manufacturing, legal and finance systems would require transparency for the stakeholder using the system to believe in the decisions given by the model, but the complex architecture fails to give that understanding and transparency to the user in the intermediate layers.
2. **Trust, Accountability & Ethical Concerns** – Any AI system that make automated decisions should undergo regulatory compliance as a part of the legal framework. The regulatory requirements mandate that a decision made by the AI system should be explainable. It becomes challenging to trust or make AI system accountable for any decision that they make when the inner working of the model is not known.
3. **Trial-and-Error process for model improvement** – To enhance the accuracy and minimize errors of the AI model, it is crucial to comprehend its internal performance by identifying its strengths and weaknesses. This understanding facilitates precise debugging and optimization of the model, as a trial-and-error approach can be time-consuming and laborious.