ABSTRACT

Neural Networks are ubiquitous in high energy physics research. However, these highly nonlinear parameterized functions are treated as black boxes - whose inner workings to convey information and build the desired input-output relationship are often intractable. Explainable AI (xAI) methods can be useful in determining a neural model’s relationship with data toward making it interpretable by establishing a quantitative and tractable relationship between the input and the model’s output. In this letter of interest, we explore the potential of using xAI methods in the context of problems in high energy physics.
1 Introduction

The evolution of machine learning models in high energy physics (HEP) research can be attributed to three factors- (i) growing intricacy in detector design and operation, (ii) big data of unprecedented volumes, and (iii) availability of fast computing resources. Over the years, simpler and interpretable regression and classification models have been replaced by intractable, black-box-like deep neural networks. Recent progress in the field of explainable Artificial Intelligence (xAI) have made it possible to investigate the relationships between an AI model’s inputs, architecture, and predictions. Some methods remain model agnostic, while other methods have been developed to interpret computer vision models where intuitive reasoning can be extracted from human-annotated datasets to validate xAI techniques. However, in other data structures, like large tabular data or relational data constructs like graphs, use of xAI methods are still quite novel. The scope of xAI in high energy physics has been rather limited but not without success- showing promise, for example, in learning parton showers at the LHC and jet reconstruction using particle flow algorithms.

2 The Scope of xAI in High Energy Physics

Explainability may not be indispensable for many ML tasks. However, in most such cases, either the model itself is simple enough that explanations are rather trivial or the impact of such models remains marginal in a broader problem setting. If we are to understand, trust, and manage AI models, it is important that a humanly intelligible interpretation should exist. Experimental high energy physics deals with incredibly complex detectors and very large datasets. Neural networks and deep learning models already play very important roles in analysis of detector simulation and performance, reconstruction of physics objects from tracks and energy deposits in the detectors, modeling of parton distribution and showers, and dedicated jet tagging to identify boosted, heavy quarks. Given the large computational expense of traditional detector simulation algorithms, generative models for accelerated simulation have been explored. The particle physics community has also started exploring the implementation of pre-trained DNN models on FPGA devices for online and offline trigger systems, object reconstruction, tracking and tagging. The spectrum of ML applications in LHC research is broad and ever-increasing. Given their intimate usage with online detector operations, it becomes increasingly more important to be able to explain and interpret these models.

2.1 Generative models for physics and detector simulation

Using generative models for detector simulation has shown initial success in speeding up the process of simulation with commendable accuracy. However, generative models are known to have some limitations. For instance, Generative Adversarial Networks (GANs) are known to cause mode collapse where a large subspace of latent space embedding maps to a relatively small subset of the feature space. On the other hand, Variational Auto
Encoders (VAEs) often lose sharpness in the generated model space [22]. Application of such models for LHC models hence requires rigorous scrutiny since physics of interest are rare and their latent space embeddings need to be well understood to make sure physics details and their resolutions are not lost in the process. How the latent space embeddings rely on input features, how well the network recognizes and manages correlated input features—these questions need to be understood in the context of such machine learning models.

2.2 Machine learning for object reconstruction

Combining millions of detector signals to reconstruct objects of interest, such as electrons, muons, and jets relies on dedicated algorithms for track identification and track matching with calorimeter signatures. Recent developments have shown that ML models can successfully recreate and sometimes outperform traditional algorithms [11]. How these models evaluate the input data and weigh the relative importance of these features need to be understood. For instance, the feature importance for MLPFlow algorithm has been explored by using the Layerwise Relevance Propagation (LRP) algorithm [9].

2.3 Modeling of parton distribution and showers

Effective modeling the partonic structure of colliding hadrons, as done by the NNPDF collaboration for instance, has successfully employed DNN techniques [12,15]. The underlying training code and methods have been recently made public [23]. DNNs have also been used to investigate parton showers dictated by non-perturbative QCD [24]. Recent results with xAI methods have shown that such models are capable of not only correctly predicting the final distribution of particles but also learning the underlying physics [8]. Extensive efforts in understanding ML models’ capability of learning QCD physics to correctly understand physics models for HL-LHC and future detectors can significantly improve discovery potential and analysis sensitivity.

2.4 Object tagging and event classification

This is one of the most active areas of ML application in the experimental HEP. Identifying potential signal events from a much larger set of background events—essentially a needle in a haystack problem—makes it ideal to adapt already tested and validated models from computer vision and natural language processing. For instance, the GoogleNET model [25], a convolutional neural network (CNN) for visual image recognition, was successfully adapted for classifying neutrinoless double beta decay with the NEXT-100 [26] detector [27]. In the context of the NOvA detector [28], similar studies were done with modified CNN models [29]. While such applications of established ML models in neutrino experiments have made them increasingly more popular over the years, there are major challenges, including interpretability of these models, associated with their comprehensibility and trustworthiness [30].
On the energy frontier, identifying the origin of jets and associating them with partons is one of the areas where ML has been most widely used. Ref. [31] summarizes a wide variety of models dedicated to identify jets originating from top quarks, novel architectures like Interaction Network (IN) [32] have been developed to identify $H \rightarrow b\bar{b}$ jets from QCD background. The inputs to these models are a large set of kinematic variables associated with particle tracks, secondary vertices, and calorimeter signatures. Incorporating such large number of variables that have non-trivial, intractable correlations can lead to arbitrarily complex models while models themselves can rely on a handful of these features. For instance, we examine the change in ROC-AUC score [33] for the IN model when individual particle track and secondary vertex features are masked in Figure 1a. It reveals that the model’s performance is hardly affected when certain features are masked. Figure 1b shows the Neural Activation Pattern (NAP) diagram for the model, showing relative activation strength of each node in the activation layers of the model, each normalized with respect to the largest cumulative activation node in the same layer. The relative activation strength illustrates that only a smaller subset of the activation nodes strongly participate in information propagation across the model. These studies suggest that the model can be made significantly simpler, both in terms on the number of features is relies on as well as the number of trainable parameters.

Figure 1: (a) Change in AUC score with respect to the baseline model when each of the track and secondary vertex features is individually masked during inference with the trained baseline model. (b) 2D map of relative neural activity for different nodes of the activation layers. To simultaneously visualize the scores for QCD and $H \rightarrow b\bar{b}$ jets, we project the activity scores of the former as negative values.
2.5 DNNs built on FPGAs for Trigger applications

With DNN’s success in jet and event classification, recent work has placed emphasis on developing DNN-enabled FPGAs for trigger level applications at the LHC [15–18]. Since these inference-only dedicated hardware can play crucial role in event classification and identifying events of interest in real time, explainability becomes an important criteria to ensure these models are trustworthy. Interpretability of DNNs implemented on dedicated hardware requires simultaneous understanding of how a model’s performance varies when operating on these platforms as well as how the model’s architecture interplays with the electronic signals received from the detectors.

3 Summary and Outlook

The scope of xAI in HEP research is as broad as that of ML applications in HEP. While the existing methods of xAI can help us better understand, trust, and optimize the various ML models used in HEP research, the constraints from underlying physics may also pave ways of developing novel explainability metrics for dedicated ML models in HEP. Even beyond the usual scope of regression and classification tasks, such dedicated xAI methods can be very useful for uncertainty quantification. The interface of xAI and uncertainty quantification in ML is still quite unexplored, both as a broader statistical problem and in domain specific applications [34]. The broader scope of ML applications in HEP will require dedicated understanding of how robust as well as interpretable these models are [35]. Being a data intensive research community, methods of xAI will provide invaluable insight regarding how data relates with intricate and otherwise intractable models that represent them. Hence, it is quite fathomable that the community’s major scientific drives, both in near- and long-term, will significantly benefit from dedicated exploration of such methods.

References


