Contribution ID: **792** Type: **Parallel Talk**

Interpretability of an Interaction Network for identifying $H \to b\bar{b}$ jets

Saturday, 9 July 2022 15:30 (15 minutes)

Multivariate techniques and machine learning models have found numerous applications in High Energy Physics (HEP) research over many years. In recent times, AI models based on deep neural networks are becoming increasingly popular for many of these applications. However, neural networks are regarded as black boxes- because of their high degree of complexity it is often quite difficult to quantitatively explain the output of a neural network by establishing a tractable input-output relationship and information propagation through the deep network layers. As explainable AI (xAI) methods are becoming more popular in recent years, we explore interpretability of AI models by examining an Interaction Network (IN) model designed to identify boosted $H \to b\bar{b}$ jets amid QCD background. We explore different quantitative methods to demonstrate how the classifier network makes its decision based on the inputs and how this information can be harnessed to reoptimize the model- making it simpler yet equally effective. We additionally illustrate the activity of hidden layers within the IN model as Neural Activation Pattern (NAP) diagrams. Our experiments suggest NAP diagrams reveal important information about how information is conveyed across the hidden layers of deep model. These insights can be useful to effective model reoptimization and hyperparameter tuning.

In-person participation

No

Primary author: ROY, Avik (University of Illinois at Urbana-Champaign)

Co-author: NEUBAUER, Mark (University of Illinois at Urbana-Champaign)

Presenter: ROY, Avik (University of Illinois at Urbana-Champaign)

Session Classification: Computing and Data handling

Track Classification: Computing and Data handling