ARTIFICIAL INTELLIGENCE AND THE UNCERTAINTY CHALLENGE IN FUNDAMENTAL PHYSICS
27 NOV - 1 DEC 2023
AI S2AI
cnrs
IN2P3

ID de Contribution: **32**

Type: **Non spécifié**

Explainable AI for Interpretability of Deep Neural Networks : the High Energy Physics perspective

mardi 28 novembre 2023 09:45 (35 minutes)

Explainable AI (xAI) represents a set of processes and methods that allows human users to comprehend results created by machine learning algorithms. In the context of applications of AI to science, we need to look beyond standard metrics of prediction performance such as accuracy to ensure that AI models are robust to noise and adversarial samples, fair to biases in data populations, and generate trustworthy explanations of their predictions. A challenge is that xAI is hard to define and even harder to evaluate. There is no universal definition of what it means for an AI model to be explainable nor well-defined metrics to evaluate the "goodness" of explanations generated for AI models. Despite of these challenges, current xAI tools and methods are powerful allies for physicists. They have great utility in aiding in the interpretation deep neural networks (DNNs) and this information can be used to create better algorithms.

In this talk, I will discuss these aspects of xAI and the application of xAI methods to DNN models used in jet tagging. In our case study of jets coming from top quark decay in the high energy proton-proton collisions at the Large Hadron Collider, we use XAI to help identify which features play the most important roles in identifying the top jets, how and why feature importance varies across different XAI metrics, and how latent space representations encode information as well as correlate with physical quantities. We additionally illustrate the activity of hidden layers as Neural Activation Pattern (NAP) diagrams to understand how DNNs relay information across the layers and how this understanding can help us to make such models significantly simpler by allowing effective model re-optimization and hyperparameter tuning.

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

Classification de Session: Explainable AI