EUROPEAN AI FOR
FUNDAMENTAL PHYSICS
CONFERENCE
EuCAIFCon 2025

Contribution ID: **110**

Type: **Poster + Flashtalk**

Investigating Explainable Jet Tagging with Pretrained Vision Transformers and Attention Mechanisms

Background: In High Energy Physics (HEP), jet tagging is a fundamental classification task that has been extensively studied using deep learning techniques. Among these, transformer networks have gained significant popularity due to their strong performance and intrinsic attention mechanisms. Furthermore, pre-trained transformer models are available for a wide range of classification tasks.

Aim of the work: In this work, we investigate the use of a pre-trained Vision Transformer (ViT) for jet tagging, leveraging reconstructed images from the JETCLASS dataset. Leveraging on its attention mechanism and analyzing attention maps, we provide insights into the model's decision-making process, addressing the challenge of interpretability in deep learning models often seen as "black boxes."

Methods: Two jet tagging tasks were selected: the first distinguishing between two very different processes with different number of jets in the final space and the second differentiating between two very similar processes with the same number of jets and similar properties in the final state.

To assess the generalization of the pretrained model, the fine-tuning was performed on a small dataset of 300 images per class, updating only the final two encoder layers while freezing the others. This approach refined high-level features while preserving pretrained representations.

Cumulative attention maps were generated by averaging attention weights across all heads and layers, incorporating residual connections for normalization. Model interpretability was assessed by computing the centroids of each jet within the image and defining regions of interest (ROIs) around these centroids. The attention fraction, quantifying the proportion of attention concentrated within the ROI relative to the total attention across the entire map, was computed to analyze the model's decision-making process.

Conclusions: This work aims to evaluate whether pretrained networks can optimize computational resources without compromising performance and whether attention-based methods enhance interpretability and explainability analysis. Additionally, it explores the potential of the HEP domain as a framework for technical validation, leveraging high-quality data and well-understood causal structures that could be applied to other scientific fields.

AI keywords

Explainability, Vision Transformers, Physics-informed AI, Jet Tagging, Attention Mechanism

Primary authors: MONTELEONE, Mariagrazia (Politecnico di Milano); CAMPONOVO, Federico (INFN sezione di Milano –Bicocca, Milano, Italy); CEDERLE, Lorenzo (Istituto Nazionale di Fisica Nucleare); CAMAGNI, Francesca (Istituto Nazionale di Fisica Nucleare); GOVONI, Pietro (Istituto Nazionale di Fisica Nucleare); GENNAI, Simone (Istituto Nazionale di Fisica Nucleare); Prof. PAGANELLI, Chiara (Dipartimento di Elettronica,Informazione e Bioingegneria, Politecnico di Milano)

Presenter: MONTELEONE, Mariagrazia (Politecnico di Milano)

Track Classification: Explainability & Theory