Investigating Explainable Jet Tagging with Pretrained Vision Transformers and Attention Mechanisms

Mariagrazia Monteleone[1], F. Camponovo [2], L. Cederle[1], Francesca Camagni[1], Pietro Govoni [2,3], Simone Gennai [2], Chiara Paganelli[1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
[2] INFN - Sezione di Milano Bicocca, Italy
[3] Università degli Studi Milano Bicocca, Milano, Italy

Aim of the Work

Jet tagging is a fundamental classification task in High Energy Physics (HEP), increasingly tackled using deep learning. Transformer networks, with their powerful **attention mechanisms**, have shown strong performance in this domain.

In this work, we explore the use of a **pre-trained Vision Transformer** (ViT) on reconstructed jet images from the **JETCLASS dataset** (A. Dosovitskiy et al., 2020) going beyond performance and focusing on **interpretability**.

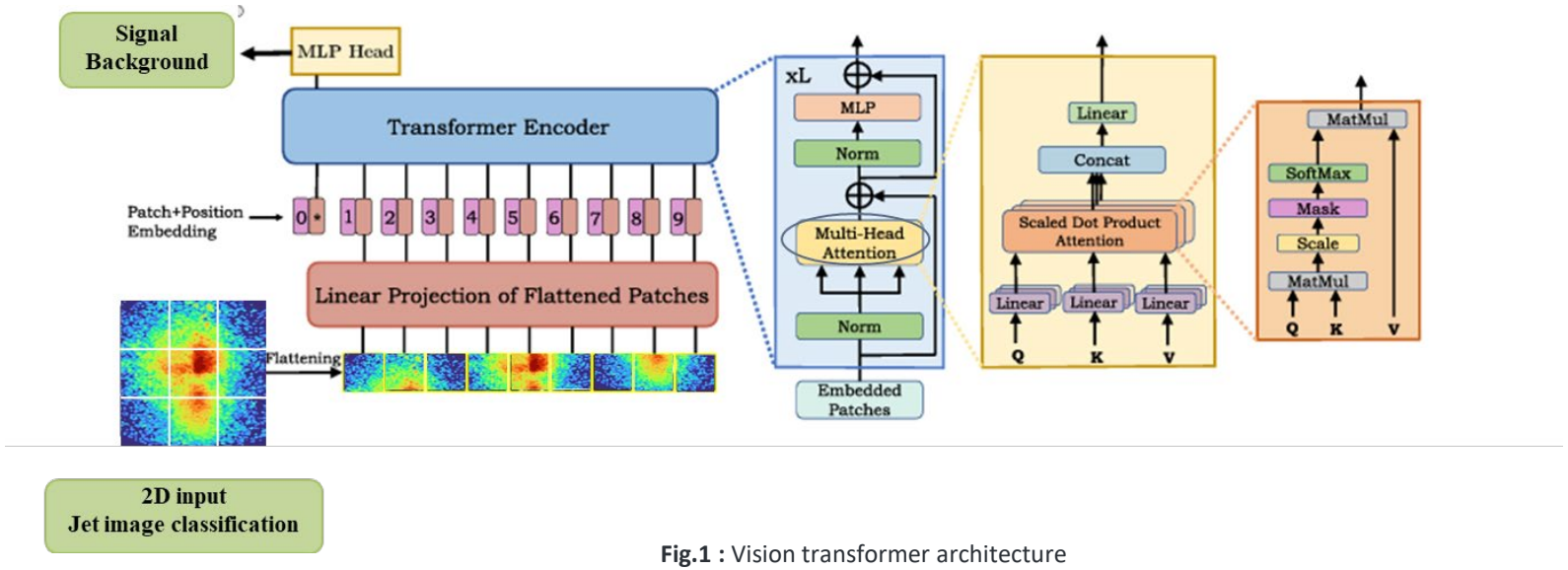

Fig.1 : Vision transformer architecture

We analyzed the attention maps to gain insights into the model's decision-making. This positions HEP as an **ideal test bed** for developing and **evaluating explainable AI (XAI) solutions**.

Dataset

To evaluate the attention mechanisms, two jet-tagging tasks were used: signal vs. background **(TTBar–ZJetsToNuNu)** and signal vs. signal **(HToBB–HToCC)** using images reconstructed from differing event counts.

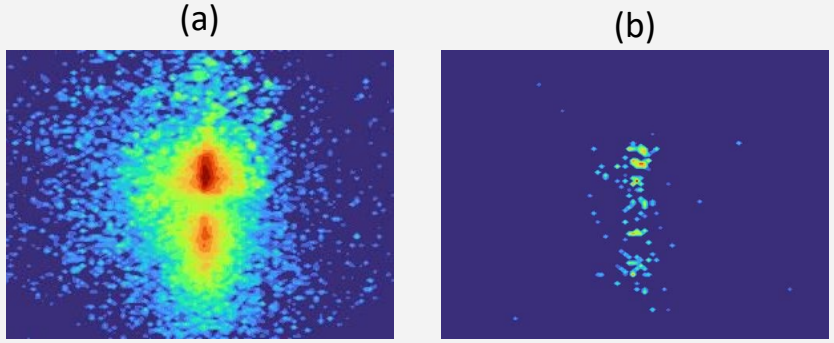

Fig. 2 HToBB jet counts 1k (a) and10 (b)

Fig. 3 HToBB jet counts 1k (a) and10 (b)

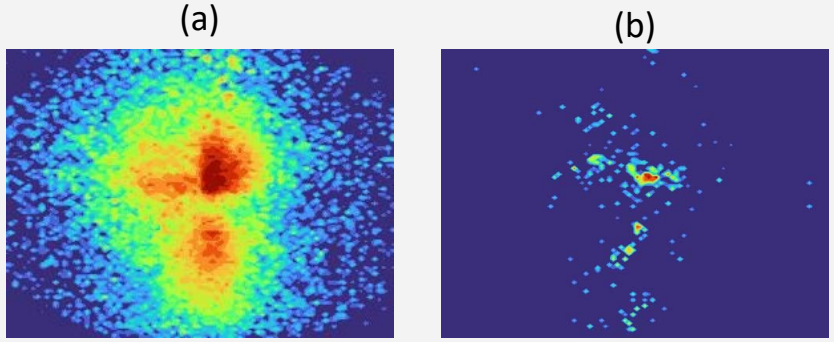

Fig. 4 TTBar jet counts 1k (a) and10 (b)

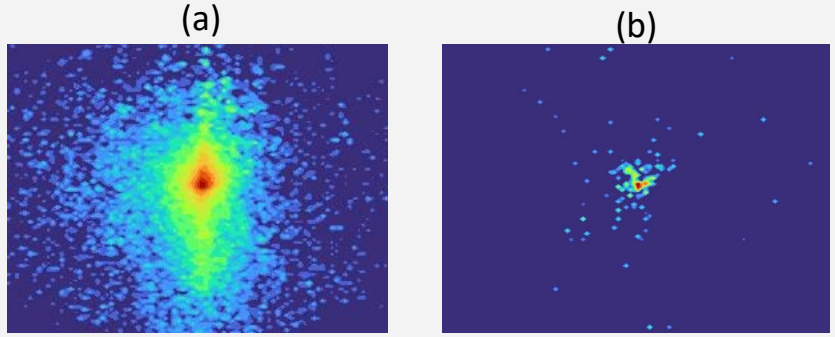

Fig. 5 ZJetToNuNu jet counts 1k (a) and10 (b)

Results

ViT was **fine-tuned** on 300 images/class by updating only layers 11–12. We analyzed **cumulative attention maps** to assess interpretability, averaging attention over all heads/layers. **Mean attention** was computed in up to three ROIs around prong centroids per event.

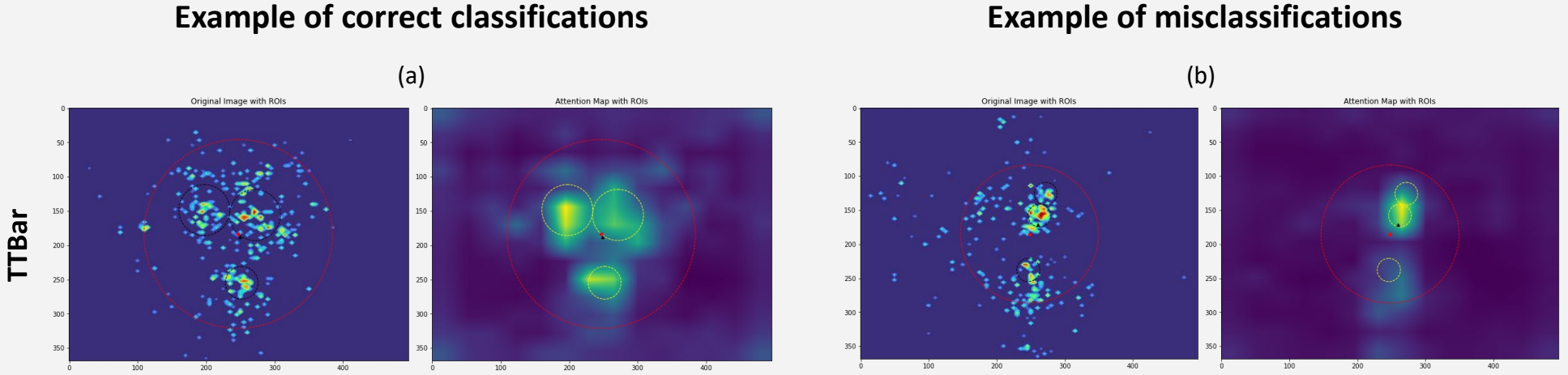

Fig. 5 Example of TTBar original image and its relative attention map correctly classified (a) and not (b).

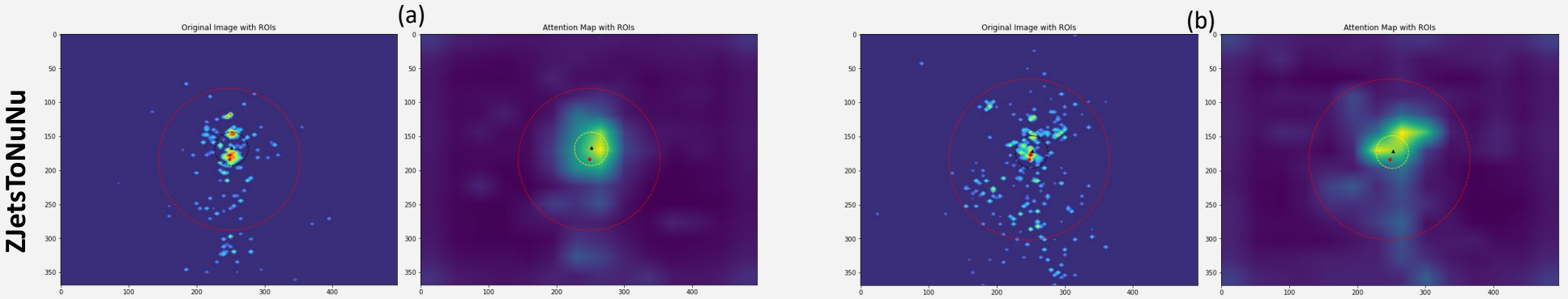

Fig. 6 Example of ZJetsToNuNu original image and its relative attention map correctly classified (a) and not (b).

Stop by my poster during poster session B Thursday, 12:00 and 15:00, to learn more!

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE
EuCAIFCon 2025