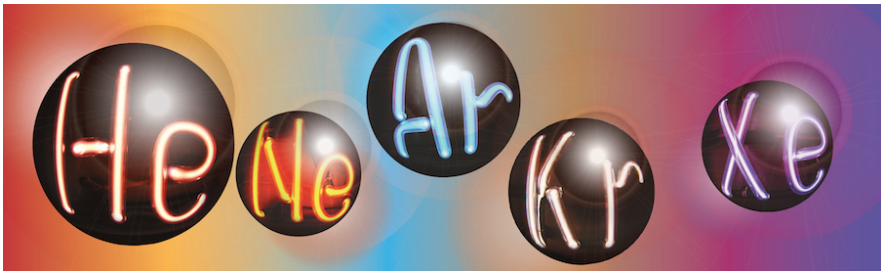

Contribution ID: **50** Type: **Talk**

Physics at the Edge: Fast, Accurate, and Low Power Consuming AI for Particle Physics

AI has had a transformative effect on particle physics, including visual pattern recognition in liquid argon time-projection chamber (LArTPC) experiments. To solve complex topological problems both at trigger level and in offline reconstruction, algorithms are currently deployed on CPUs or GPUs. The former are unspecialised chips with low speed, whilst the latter offer high speed but at the cost of very high power consumption. In this talk, we will go through the work we have been doing on new paradigms for fast, accurate, and low-power-consuming AI in LArTPC experiments using edge AI devices. In our first publication, we conducted a benchmark of different quantisation techniques and speed, accuracy, and power consumption of convolutional neural networks (CNNs) used for neutrino event recognition on CPU, GPU, and edge TPU. We showed that with the edge TPU, made by Google Coral, we obtained accuracy and latency comparable to a powerful GPU, but with two orders of magnitude less power consumption. In our current work, we are testing advanced transformer models to separate at a hit-based level track and shower-like topologies in LArTPCs. We will show that these models can be successfully quantised and deployed on state-of-the-art edge devices, with competitive speed and accuracy, but a fraction of the energy required by CPUs and GPUs. The power consumption and costs of GPUs represent a bottleneck to future AI-enabled LAr experiments, in addition to serious concerns for the environment. With this work, we aim to present a possible solution without sacrificing latency or accuracy.

Primary author: VERGANI, Stefano (King's College London)

Presenter: VERGANI, Stefano (King's College London)