Contribution ID: **57** Type: **Talk without Poster**

Studying Adversarial Deep Learning techniques in the context of High-Energy Physics

Tuesday, 30 April 2024 17:22 (20 minutes)

Adversarial deep learning techniques are based on changing input distributions (adversaries), with the goal of causing false classifications when input to a deep neural network classifier. Adversaries aim to maximize the output error while only exerting minimal perturbations to the input data. Moreover, various techniques to defend against such attacks have been developed in the past. While rooted in AI Safety, adversarial deep learning offers a range of techniques that could potentially enhance high-energy physics deep learning models. Additionally, it might provide new opportunities to gather insights into the systematic uncertainties of deep neural networks. While adversarsial deep learning has triggered immense interest in the recent years in all kind of fields, its possible applications in the context of high-energy physics (HEP) have not yet been studied in detail.

In this work, we employ adversarial deep learning techniques on multiple neural networks from within the high-energy physics domain, all reconstructed using publicly available data from the CERN Open Data portal to ensure reproducibility. Through the utilization of adversarial attacks and defense techniques, we not only assess the robustness of these networks but additionally aim for the construction of HEP networks portraying larger robustness and better generalization capabilities.

Primary author: SAALA, Timo

Co-author: SCHOTT, Matthias (Physikalisches Institut der Universität Bonn)

Presenter: SAALA, Timo

Session Classification: 3.3 Hardware acceleration, FPGAs & Uncertainty quantification