Contribution ID: **68** Type: **not specified**

Enhancing generalization in high energy physics using white-box adversarial attacks

Wednesday 6 November 2024 11:30 (20 minutes)

Machine learning is becoming increasingly popular in the context of particle physics. Supervised learning, which uses labeled Monte Carlo simulations, remains one of the most widely used methods for discriminating signals beyond the Standard Model. However, this paper suggests that supervised models may depend excessively on artifacts and approximations from Monte Carlo simulations, potentially limiting their ability to generalize well to real data. This study aims to enhance the generalization properties of supervised models. It reviews the application of four distinct white-box adversarial attacks, in the context of classifying Higgs boson decay signals. The attacks are divided into two groups: weight space attacks, and feature space attacks. A dense network is used to compare these methods. To study and quantify the sharpness of the found local minima, this paper also presents two analysis methods: gradient ascent and reduced Hessian eigenvalue analysis. The results show that white-box adversarial attacks significantly improve generalization performance, though they also increase computational complexity.

Track

Tagging (Classification)

Author: ROTHEN, Franck (Universite de Geneve (CH))

Co-authors: Mr LEIGH, Matthew (University of Geneva); KLEIN, Samuel Byrne (Universite de Geneve (CH)); GOLLING, Tobias (Universite de Geneve (CH))

Presenter: ROTHEN, Franck (Universite de Geneve (CH))

Session Classification: Tagging