Contribution ID: **341** Type: **Poster**

Improving robustness of jet tagging algorithms with adversarial training

Tuesday, 25 October 2022 16:10 (30 minutes)

In the field of high-energy physics, deep learning algorithms continue to gain in relevance and provide performance improvements over traditional methods, for example when identifying rare signals or finding complex patterns. From an analyst's perspective, obtaining highest possible performance is desirable, but recently, some focus has been laid on studying robustness of models to investigate how well these perform under slight distortions of input features. Especially for tasks that involve many (low-level) inputs, the application of deep neural networks brings new challenges. In the context of jet flavor tagging, adversarial attacks are used to probe a typical classifier's vulnerability and can be understood as a model for systematic uncertainties. A corresponding defense strategy, adversarial training, improves robustness, while maintaining high performance. This contribution presents different approaches using a set of attacks with varying complexity. Investigating the loss surface corresponding to the inputs and models in question reveals geometric interpretations of robustness, taking correlations into account. Additional cross-checks against other, physics-inspired mismodeling scenarios are performed and give rise to the presumption that adversarially trained models can cope better with simulation artifacts or subtle detector effects.

Experiment context, if any

Context: ATLAS, CMS

References

https://arxiv.org/abs/2203.13890

Significance

Such studies are crucial to understand if potential mismodelings in simulation could lead to differences in performance in data compared to simulation. Sophisticated calibration techniques are applied which at times might still leave residual disagreement. Therefore, any technique that evades that problem during training and that probes the trade-off between performance and robustness is of importance for identification of physics objects with deep learning algorithms, especially with large numbers of (low-level) input features. In this contribution, a successful application of defense strategies for deep-learned flavor tagging algorithms is shown and is accompanied by novel insights into the neural network's properties that help explaining the observed behavior.

Primary author: STEIN, Annika (Rheinisch Westfaelische Tech. Hoch. (DE))

Presenters: STEIN, Annika (Rheinisch Westfaelische Tech. Hoch. (DE)); MONDAL, Spandan (RWTH Aachen (DE))

Session Classification: Poster session with coffee break