Contribution ID: **115** Type: **Oral**

Leveraging Large-Scale Pretraining for Efficient Jet Classification: An Evaluation of Transfer Learning, Model Architectures, Dataset Scaling, and Domain Adaptation in Particle Physics

Monday 11 March 2024 15:10 (20 minutes)

In particle physics, machine learning algorithms traditionally face a limitation due to the lack of truth labels in real data, restricting training to only simulated samples. This study addresses this challenge by employing self-supervised learning, which enables the utilization of vast amounts of unlabeled real data, thereby facilitating more effective training.

Our project is particularly motivated by the need for improved data-Monte Carlo (MC) agreement in CMS analyses, seeking to bridge the gap between simulation and real-world data. We employ contrastive learning to leverage the JetClass dataset for large-scale pretraining, aiming to capture generalizable features about jets. These features can then be fine-tuned for downstream classification tasks such as Top Tagging and $H \to b\bar{b}$ vs QCD, with minimal additional effort.

The research explores several key questions: the scalability of dataset size in pretraining, the comparative analysis of contrastive learning techniques like SimCLR and VICReg, the effectiveness of the ParticleTransformer architecture over conventional transformer models, and whether self-supervised pretraining on unlabeled data combined with fine-tuning on labeled simulation aids the model in adapting to the data domain.

By investigating these aspects, we aim to provide insights into the impact of dataset size on pre-training, evaluate the strengths and weaknesses of various contrastive learning methods, assess the architectural advantages of ParticleTransformer in jet classification, and facilitate the domain adaptation of machine learning algorithms for enhanced applicability in particle physics.

This study significantly contributes to the field of machine learning in particle physics, demonstrating the immense potential of self-supervised learning in utilizing real, unlabeled data for more efficient and accurate jet classification.

Significance

This research introduces a groundbreaking approach in experimental particle physics by utilizing self-supervised learning, particularly contrastive learning, to exploit large datasets of unlabeled real data. It explores novel aspects such as the scalability of pre-training dataset size, the comparative effectiveness of contrastive learning techniques, and the potential of the ParticleTransformer architecture over conventional models. These contributions represent substantial progress beyond standard practices, providing practical advancements in machine learning applications within experimental particle physics.

References

Experiment context, if any

Primary authors: PAREJA, Carlos (University of California, San Diego); MOKHTAR, Farouk (Univ. of California San Diego (US)); LI, Haoyang (Univ. of California San Diego (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); KANSAL, Raghav (Univ. of California San Diego (US)); ZHAO, Zihan (Univ. of California San Diego (US))

Presenter: ZHAO, Zihan (Univ. of California San Diego (US))

Session Classification: Track 2: Data Analysis - Algorithms and Tools

Track Classification: Track 2: Data Analysis - Algorithms and Tools