Contribution ID: **54** Type: **not specified**

Large-Scale Pretraining and Finetuning for Efficient Jet Classification in Particle Physics

Thursday 7 November 2024 17:20 (20 minutes)

This study introduces an innovative approach to analyzing unlabeled data in high-energy physics (HEP) through the application of self-supervised learning (SSL).
Faced with the increasing computational cost of producing high-quality labeled simulation samples at the CERN LHC, we propose leveraging large volumes of unlabeled data to overcome the limitations of supervised learning methods, which heavily rely on detailed labeled simulations. By pretraining models on these vast, mostly untapped datasets, we aim to learn generic representations that can be finetuned with smaller quantities of labeled data. Our methodology employs contrastive learning with augmentations on jet datasets to teach the model to recognize common representations of jets, addressing the unique challenges of LHC physics.

Building on the groundwork laid by previous studies, our work demonstrates the critical ability of SSL to utilize large-scale unlabeled data effectively.
We showcase the scalability and effectiveness of our models by gradually increasing the size of the pretraining dataset and assessing the resultant performance enhancements.
Our results, obtained from experiments on two datasets—JetClass, representing unlabeled data, and Top Tagging, serving as labeled simulation data—show significant improvements in data efficiency, computational efficiency, and overall performance. These findings suggest that SSL can greatly enhance the adaptability of ML models to the HEP domain. This work opens new avenues for the use of unlabeled data in HEP and contributes to a better understanding of the potential of SSL for scientific discovery.

Track

Tagging (Classification)

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

Co-authors: MOKHTAR, Farouk (Univ. of California San Diego (US)); KANSAL, Raghav (Univ. of California San Diego (US)); LI, Haoyang (Univ. of California San Diego (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US))

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

Session Classification: Foundation models