Lossy Data Compression for Simulated CMS Pile-up Pixel Subdetector Datasets

Fellow: Paulius Balciunas
Mentors: Tomas Raila, Valdas Rapsevicius

Duration: 12 weeks

Project Description

The High-Luminosity upgrade of the LHC (HL-LHC) will present an unprecedented computational challenge for the CMS experiment, with the average number of simultaneous proton-proton interactions (pile-up) expected to reach 200 per bunch crossing. Accurately modeling this background environment requires the production of massive, high-fidelity simulated datasets at the DIGI-level (the digitized output of the detector electronics). Currently, CMS relies on the pre-mixing strategy, where large libraries of pile-up events are pre-generated and stored for later reuse. However, the large volume of these datasets is becoming a significant storage bottleneck. The long-term goal of the HEP community is to employ generative surrogate models — ML architectures capable of sampling simulated event data on-the-fly. This project focuses on an intermediate proof-of-concept step: developing high-performance lossy data compression for the CMS pixel tracker to reduce storage footprints and create the discrete latent representations applicable for future generative modeling.

The core of this approach is the application of vector quantization (VQ) techniques to the sparse, high-dimensional hit maps produced by the pixel tracker. By discretizing the detector information into a finite set of codebook indices, we can achieve substantial compression factors while maintaining control over the reconstruction error. This work involves several directions:

- Exploring and adapting various VQ techniques in order to optimize the balance between compression rate, performance and reconstruction error
- Applying ML-based VQ methods and neural network architectures, such as VQ-VAE
- Integrating the proposed data compression solution into CMSSW framework by developing relevant modules
- Extending the proof-of-concept solution to the simulated data of other CMS subdetectors

This project will address issues by optimizing the Physics analysis pipeline by reducing the size of simulated Pileup dataset.

.

Software Deliverables

1. VQ algorithm and analysis for CMS Pixel data compression.
2. VQ-VAE algorithm and analysis for CMS Pixel data compression.

3. Data analysis documentation.
4. Final Report & Presentation: A summary of the project work, results, and software, as required by IRIS-HEP.

Project Timeline (12 Weeks)

- Weeks 1-2: Getting Started, Analysis & Early Development
    - Learn the CMS Pileup dataset, formats, and tools.
    - Learn VQ and VAE algorithms for compression applications.
    - Plan further analysis.
- Weeks 3-6: VQ Algorithm Development Work
    - Implement the CMS Pixel data VQ compression algorithm.
    - Analyze compression parameters and tune up algorithm.
- Weeks 7-10: VQ-VAE Algorithm Development Work
    - Enhance VQ compression algorithm with VAE component.
    - Analyze compression parameters and tune up algorithm.
- Weeks 11-12: Writing Up & Finishing
    - Write down the technical details and how to use the new tools.
    - Complete the final report and prepare the presentation.