Contribution ID: **184** Type: **Oral**

End-to-end hardware-aware model compression and deployment with PQuant and hls4ml

Tuesday 9 September 2025 14:50 (20 minutes)

Machine learning model compression techniques—such as pruning and quantization—are becoming increasingly important to optimize model execution, especially for resource-constrained devices. However, these techniques are developed independently of each other, and while there exist libraries that aim to unify these methods under a single interface, none of them offer integration with hardware deployment libraries such as hls4ml. To address this, we introduce PQuant, a Python library that simplifies the training and compression of machine learning models by providing an interface for applying a variety of pruning and quantization methods. PQuant is designed to be accessible to users without specialized knowledge of compression algorithms, while still offering deep configurability. It integrates with hls4ml, allowing compressed models to be directly utilized by FPGA-based accelerators. This makes it a valuable tool for both researchers comparing compression strategies and practitioners targeting efficient deployment on edge devices and custom hardware.

Significance

We present a Python library for training pruned and quantized machine learning models. The library includes multiple pruning methods, quantization and high-granularity quantization support, and it integrates with hls4ml for hardware deployment.

References

NextGen Triggers Technical Workshop at CERN [link: https://nextgentriggers.web.cern.ch/nextgen-triggers-technical-workshop-kicks-off-at-cern/] Presentation link: https://indico.cern.ch/event/1421629/contributions/6136755/

Experiment context, if any

This work focuses on creating an experiment agnostic, hardware-aware compression library for ML models, following the goals of the Next-Gen Triggers Project at CERN

Author: NIEMI, Roope Oskari

Co-authors: SUN, Chang (California Institute of Technology (US)); PETROVYCH, Anastasiia (CERN); Dr LUPI, Enrico (CERN, INFN Padova (IT)); DANOPOULOS, Dimitrios (CERN); DITTMEIER, Sebastian (Ruprecht-Karls-Universitaet Heidelberg (DE)); KAGAN, Michael (SLAC National Accelerator Laboratory (US)); LONCAR, Vladimir (CERN)

Presenter: NIEMI, Roope Oskari

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research