Contribution ID: **42** Type: **Contributed talk**

PQuant: A Tool for End-to-End Hardware-Aware Model Compression

Friday 23 May 2025 11:30 (20 minutes)

Machine learning model compression methods such as pruning and quantization are critical for enabling efficient inference on resource-constrained hardware. Compression methods are developed independently, and while some libraries attempt to unify these methods under a common interface, they lack integration with hardware deployment frameworks like hls4ml. To bridge this gap, we present PQuant, a Python library that streamlines the training and compression of machine learning models. PQuant offers an interface for applying diverse pruning and quantization methods, making it accessible to users without deep expertise in compression while still supporting advanced configuration. Notably, integration with hls4ml is ongoing, which will enable deployment of compressed models to FPGA-based accelerators. This will make PQuant a practical tool for both researchers exploring compression strategies and engineers aiming for efficient inference on edge devices and custom hardware platforms.

Would you like to be considered for an oral presentation?

Yes

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: Contributed Talks

Track Classification: 5 Fast ML: Application of ML to DAQ/Trigger/Real Time Analysis/Edge Computing