Contribution ID: **22** Type: **Standard Talk**

End-to-End Neural Network Compression and Deployment for Hardware Acceleration Using PQuant and hls4ml

Tuesday 2 September 2025 13:20 (20 minutes)

As the demand for efficient machine learning on resource-limited devices grows, model compression techniques like pruning and quantization have become increasingly vital. Despite their importance, these methods are typically developed in isolation, and while some libraries attempt to offer unified interfaces for compression, they often lack support for deployment tools such as hls4ml. To bridge this gap, we developed PQuant, a Python library designed to streamline the process of training and compressing machine learning models. PQuant offers a unified interface for applying a range of pruning and quantization techniques, catering to users with minimal background in compression while still providing detailed configuration options for advanced use. Notably, it features built-in compatibility with hls4ml, enabling seamless deployment of compressed models on FPGA-based accelerators. This makes PQuant a versatile resource for both researchers exploring compression strategies and developers targeting efficient implementation on edge devices or custom hardware platforms. We will present the PQuant library, the performance of several compression algorithms implemented with it, and demonstrate the conversion flow of a neural network model from an uncompressed state to optimized firmware for FPGAs.

Author: NIEMI, Roope Oskari

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

Presenter: NIEMI, Roope Oskari

Session Classification: Contributed talks