Contribution ID: **28** Type: **Flashtalk with Poster**

Model compression and simplification pipelines for fast and explainable deep neural network inference in FPGAs in HEP

Tuesday, 30 April 2024 14:22 (3 minutes)

Resource utilization plays a crucial role for successful implementation of fast real-time inference for deep neural networks on latest generation of hardware accelerators (FPGAs, SoCs, ACAPs, GPUs). To fulfil the needs of the triggers that are in development for the upgraded LHC detectors, we have developed a multi-stage compression approach based on conventional compression strategies (pruning and quantization) to reduce the memory footprint of the model and knowledge transfer techniques, crucial to streamline the DNNs simplifying the synthesis phase in the FPGA firmware and improving explainability. We present the developed methodologies and the results of the implementation in a working engineering pipeline used as pre-processing stage to high level synthesis tools. We show how it is possible to build ultra-light deep neural networks in practice, by applying the method to a realistic HEP use-case: a toy simulation of one of the triggers planned for the HL-LHC. Moreover we explored an array of xAI methods based on different approaches, and we tested their capabilities in the HEP use-case., and as a result, we obtained an array of potentially easy-to-understand and human-readable explanations of models' predictions, describing for each of them strengths and drawbacks in this particular scenario, providing an interesting atlas on the convergent application of multiple xAI algorithms in a realistic context.

Primary authors: RUSSO, Graziella (Sapienza Università di Roma and INFN Roma); GIAGU, Stefano (Sapienza Università di Roma and INFN Roma, Roma, Italy)

Presenter: RUSSO, Graziella (Sapienza Università di Roma and INFN Roma)

Session Classification: 1.4 Hardware acceleration & FPGAs

Track Classification: Session A