EUROPEAN AI FOR
FUNDAMENTAL PHYSICS
CONFERENCE
EuCAIFCon 2024

Contribution ID: **196**

Type: **Talk without Poster**

Hardware implementation of quantum machine learning predictors for ultra-low latency applications

Tuesday, 30 April 2024 13:42 (20 minutes)

Tensor Networks (TNs) is a computational paradigm used for representing quantum many-body systems. Recent works show how TNs can be applied to perform Machine Learning (ML) tasks, yielding comparable results to standard supervised learning techniques. In particular [1] leveraged Tree Tensor Networks (TTNs) to achieve the classification of particle flavor state in the context of High Energy Physics.

In this work, we want to analyze the use of TTNs in high-frequency real-time applications like online trigger systems. Indeed, TTN-based algorithms can be deployed in online trigger boards, by exploiting low latency hardware like FPGA. Besides, FPGAs are known to be suitable for inherently concurrent tasks like matrix multiplications. When implementing biologically inspired neural network on FPGA the goal is to keep the design as small as possible to cope with the resource limitations. Pruning is the primary technique adopted for removing parameters that do not substantially contribute to the performance of the ML task. On the other hand, quantum features of the TTN like quantum correlations or entanglement entropy can be mapped to properties of the network that can help in the pruning process identifying unnecessary features or nodes [2]. This makes TTNs good candidates for efficient hardware implementation.

We will show different implementations of a TTNs classifier on FPGA capable of performing inference on a classical dataset used for ML benchmarking. A preparatory analysis of bond dimensions, features ordering, and weight quantization, done in the training phase, will lead to the choice of the TTN architecture. The generated TTNs will be deployed on hardware accelerator. Using an FPGA integrated in a server we will completely offload the inference of the TTN. Finally, a projection of the needed resources for the hardware implementation of a classifier for the application in HEP will be provided by comparing how different degrees of parallelism obtained in hardware affect physical resources and latency.

[1] Timo Felser et al. "Quantum-inspired machine learning on high-energy physics data". In: npj Quantum Information (2021).
[2] Yoav Levine et al. "Deep Learning and Quantum Entanglement: Fundamental Connections with Implications to Network Design". In: International Conference on Learning Representations (2018)

Primary authors: BORELLA, Lorenzo (University of Padova); COPPI, Alberto (University of Padova); PAZZINI, Jacopo (University of Padova); STANCO, Andrea (University of Padova); TRIOSSI, Andrea (University of Padova); ZANETTI, Marco (University of Padova)

Presenter: TRIOSSI, Andrea (University of Padova)

Session Classification: 1.4 Hardware acceleration & FPGAs