Contribution ID: **1337** Type: **Parallel Talk**

Optimal size constraining of Deep Neural Network Models for FPGA implementation in trigger systems of experiments at future colliders

Friday, 8 July 2022 18:15 (15 minutes)

Trigger strategies for future high-rate collider experiments invariably envisage implementations of Neural Networks on FPGAs. For the HL-LHC case, as well as for FCC-ee and ILC, triggerless approaches are explored where the event selection will be largely committed to Machine Learning models directly interfaced with detector's front-end readout. Even for the huge amounts of data produced at the FCC-hh (O(TBytes/s) expected), Machine Learning or Artificial Intelligence algorithms will play an important role to make intelligent decisions as close to the detector as possible and to provide at least O(10) data reduction factors after front-end readout.
Fulfilling the requirement of 1 μs latency maintaining the size of the TDAQ system affordable in terms of procurement and maintenance costs turns out to be extremely difficult. It is unanimously considered one of the most difficult challenges for the next generation detectors at colliders.
In this respect, resource optimization is an indispensable part of the design of TDAQ algorithms, where models must be suitably compressed, quantized, and parallelized before implementation on ASICs and FPGAs. Reshaping and compression are the first and most critical steps to take towards Neural Network optimization. The strategy is often based on sampling the hyperparameter space with an iterative procedure or on a grid search; in both cases, it is not unlikely to get sub-optimal implementations.

Is it possible to optimally tune neural networks under constraints of latency and size?
Is iterative pruning the only option? Does there exist a mathematically robust method to choose the best network architecture among all the (infinite) ones compatible with the FPGA resources available?

In the last two years, the authors have developed and tested a novel strategy to prune Deep Neural Networks. The method is based on overlying a shadow network on the one to be optimized. During the training, the combined optimization of the shadow and standard networks highlights the optimal network layout for the specific task, expressed by the loss function and the available data. This system proves to be effective for real-time inference applications for trigger purposes.
With this approach, significant input features are selected while irrelevant nodes are pruned, reducing the overall network's size to dimensions ultimately decided by the designer. The pruning process is controlled down to a mean error of ± 1 on the number of active nodes and it can be applied to the entire network or only to a portion of it. The tool is straightforward to integrate into already-developed Deep Fully Connected Neural Network classifiers, allowing to reduce the amount of input data and network size without significant performance losses.

Herein we present the tool for the first time. To show its potential for trigger purposes, we constructed a simulated dataset, using Pythia and Delphes to emulate high-level trigger observables. The simulation was used to benchmark the pruner performance on a DNN tagger for the identification of jets containing b quarks arising from Higgs boson decays. As an example of unconstrained application, we will show how to easily design the best-performing Fully Connected Neural Network tagger. Concerning constrained applications, we will show how we obtained equal-performance pruned networks using down to 50% of the available nodes, and achieved performance gains of up to 25% with models up to 30% lighter.

In-person participation

No

Primary author: IUPPA, Roberto (Istituto Nazionale di Fisica Nucleare)

Co-authors: DI LUCA, Andrea (Istituto Nazionale di Fisica Nucleare); MASCIONE, Daniela (Istituto Nazionale di Fisica Nucleare); CRISTOFORETTI, Marco (Istituto Nazionale di Fisica Nucleare); FOLLEGA, Francesco Maria (Università di Trento)

Presenter: IUPPA, Roberto (Istituto Nazionale di Fisica Nucleare)

Session Classification: Detectors for Future Facilities, R&D, novel techniques

Track Classification: Detectors for Future Facilities, R&D, novel techniques