Contribution ID: **242** Type: **Oral**

Pruning and resizing deep neural networks for FPGA implementation in trigger systems at collider experiments

Thursday 27 October 2022 15:50 (20 minutes)

Deep Learning algorithms are widely used among the experimental high energy physics communities and have proved to be extremely useful in addressing a variety of tasks. One field of application for which Deep Neural Networks can give a significant improvement is event selection at trigger level in collider experiments. In particular, trigger systems benefit from the implementation of Deep Learning models on FPGAs. However, this task poses specific challenges to Deep Learning algorithm design, due to the microsecond latency requirements and limited resources of FPGA-based trigger systems. Before being implemented on an FPGA, Neural Networks may need to be appropriately compressed in order to reduce the number of neurons and synapses. A widespread technique to reduce the size of Deep Neural Networks is pruning. Numerous approaches have been developed to create a pruned model from an untrained one. Nearly all of them use a similar procedure, according to which the network is first trained to convergence, then single weights are removed on the basis of a particular ranking. To recover from accuracy loss, pruned networks are finally retrained. The pruning and retraining process is repeated iteratively, shrinking the network's size. This procedure however can be quite long and resource demanding. Moreover, the relative importance of parameters changes along iterations and this may lead to converging to sub-optimal configurations.
Here we propose a different pruning strategy, which proved to be a mathematically rigorous and faster method for optimizing Neural Networks under size constraints. Our approach works by overlaying a shadow network on the one that has to be optimized. The shadow network is very simple to incorporate into already developed Deep Neural Networks and can be used to prune the whole network or just a portion. Through the training process, the combined optimization of the shadow and standard networks takes place. As a result, the pruning procedure occurs along with the training, and not in two different phases. The proposed method performs a pruning of the nodes, rather than of the single connections, allowing for a determination of an ideal network layout, with the number of total nodes determined by the user so to match the FPGA resources available. After finding the optimal network layout, the reduced network can be retrained as a new independent model. Preliminary results will be presented, along with new developments and applications.

Significance

Here we are presenting a novel pruning strategy for compressing Deep Neural Networks for FPGA implementation. Our method is mathematically sound and time-saving with respect to standard pruning strategies. It allows to prune during the training stage, and to prune nodes rather than single connections. It provides network layouts as effective as the optimal one.

References

Experiment context, if any

The presenter and the research team are members of the ATLAS collaboration, actively working on jet flavor-tagging.

Primary authors: Mr DI LUCA, Andrea (Universita degli Studi di Trento and INFN (IT)); MASCIONE, Daniela (Universita degli Studi di Trento and INFN (IT)); FOLLEGA, Francesco Maria (Universita degli Studi di Trento and INFN (IT)); Dr CRISTOFORETTI, Marco (Universita degli Studi di Trento e INFN (IT)); IUPPA, Roberto (Universita degli Studi di Trento and INFN (IT))

Presenter: MASCIONE, Daniela (Universita degli Studi di Trento and INFN (IT))

Session Classification: Track 2: Data Analysis - Algorithms and Tools

Track Classification: Track 2: Data Analysis - Algorithms and Tools