Contribution ID: **16** Type: **Contributed Talk**

Deep neural networks resizing for online event selection in future collider experiments

Wednesday, 7 September 2022 12:00 (30 minutes)

Deep Learning algorithms are widely employed at the LHC for event processing and have proven to be highly effective. Nonetheless, the massive volume of data generated at the LHC makes it difficult to keep existing trigger schemes up to date and retain data for offline analysis. It is therefore becoming vital to run algorithms with higher selection capability online, making use of low-latency technology with high parallelization while not exceeding the available memory. Future high-rate collider experiment trigger solutions will inevitably use FPGA implementations of Deep Neural Networks. Deep learning algorithm design is complicated by the sub-microsecond latency requirements of FPGA-based trigger and data acquisition systems. Resource optimization is particularly important since models must be compressed and reshaped suitably before being implemented on FPGAs. For this task, iterative sampling of the hyperparameter space or grid search are widely used, resulting frequently in sub-optimal and time-consuming solutions. Here we present a mathematically sound and quicker strategy for optimizing Neural Networks under latency and size limitations. Our method works by creating a shadow network on top of the one that has to be optimized. Throughout the training, the combined optimization of shadow and standard networks shows the optimal network structure for the considered task, as described by the loss function and available data. This approach selects relevant input features while pruning unnecessary nodes, resulting in a smaller network with user-defined dimensions. Our method is a new pruning methodology for Deep Neural Networks that shows to be useful in real-time inference applications for trigger purposes, picking the optimum network design from an infinite number of choices that are compatible with the FPGA resources available. Our pruning process can be used on the entire network or just a portion of it, it is easy to integrate into existing Deep Fully Connected Neural Network classifiers, and it allows for the selection of the best-performing Fully Connected Neural Network tagger. We will demonstrate how our method ensures equal-performance pruned networks and sensitive performance increases with newly found lightweight models.

Primary authors: 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)); 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: Electroweak and High Energy Physics