Contribution ID: **21** Type: **Short talk**

Optimized Deep Learning Inference on High Level Trigger at the LHC: Computing time and Resource assessment

Thursday, 12 May 2022 16:20 (15 minutes)

We present a study on latency and resource requirements for deep learning algorithms to run on a typical High Level Trigger computing farm at a high-pT LHC experiment at CERN. As a benchmark, we consider convolutional and graph autoencoders, developed to perform real-time anomaly detection on all the events entering the High Level Trigger (HLT) stage. The benchmark dataset consists of synthetic multijet events, simulated at a center-of-mass energy 13 TeV. Having in mind a next-generation heterogeneous computing farm powered with GPUs, we consider both optimized CPU and GPU inference, using hardware-specific optimization tools to meet the constraints of real-time processing at the LHC: ONNX runtime for CPU and NVIDIA TensorRT for GPUs. We observe O(msec) latency with different event batch sizes for both CPU- and GPU-based model inference with maximal gain seen at batch size of 1 (corresponding to the typical use case of event-parallelized HLT farms). We show that these optimized workflows offer significant savings with respect to native solutions (Tensorflow 2 and Keras) both in terms of time and computing resources.

Primary authors: Ms WOZNIAK, Kinga Anna (CERN and University of Vienna); Dr PIERINI, Maurizio (CERN); Mr JAWAHAR, Pratik (Worcester Polytechnic Institute and CERN); HASAN, Syed Anwar Ul (Universita & INFN Pisa (IT)); Ms CHERNYAVSKAYA, Nadezda (CERN)

Presenter: HASAN, Syed Anwar Ul (Universita & INFN Pisa (IT))

Session Classification: Workshop