Contribution ID: **6** Type: **Talk**

Ultra Low-latency, Low-area Inference Accelerators using Heterogeneous Deep Quantization with QKeras and hls4ml

Tuesday 1 December 2020 14:52 (8 minutes)

While the quest for more accurate solutions is pushing deep learning research towards larger and more complex algorithms, edge devices demand efficient inference i.e. reduction in model size, speed and energy consumption. A technique to limit model size is quantization, i.e. using fewer bits to represent weights and biases. Such an approach usually results in a decline in performance. In this CERN-Google collaboration, we introduce a novel method for designing optimally heterogeneously quantized versions of deep neural network models for minimum-energy, high-accuracy, nanosecond inference and fully automated deployment on-chip. With a per-layer, per-parameter type automatic quantization procedure, sampling from a large base of quantizers, model energy consumption and size are minimized while high accuracy is maintained. This is crucial for the event selection procedure in proton-proton collisions at the CERN Large Hadron Collider, where resources are limited and a latency of O(1) micro second is required. Nanosecond inference and a resource consumption reduced by a factor of 50 when implemented on FPGA hardware is achieved.

Authors: AARRESTAD, Thea (CERN); SUMMERS, Sioni Paris (CERN); POL, Adrian Alan (CERN)

Presenter: AARRESTAD, Thea (CERN)