Fast Machine Learning for Science
Imperial College London
Real-time and accelerated ML for fundamental sciences
25-28 September 2023

Contribution ID: **8** Type: **Standard Talk**

Hardware-aware pruning of real-time neural networks with hls4ml Optimization API

Wednesday 27 September 2023 13:30 (15 minutes)

Neural networks achieve state-of-the art performance in image classification, medical analysis, particle physics and many more application areas. With the ever-increasing need for faster computation and lower power consumption, driven by real-time systems and Internet-of-Things (IoT), field-programmable gate arrays (FPGAs) have emerged as suitable accelerators for deep learning applications. Due to the high computational complexity and memory footprint of neural networks, various compression techniques, such as pruning, quantisation and knowledge distillation, have been proposed in literature. Pruning sparsifies a neural network, reducing the number of multiplications and memory. However, unstructured pruning often fails to capture properties of the underlying hardware, bottlenecking improvements and causing load-balance inefficiency on FPGAs.

We propose a hardware-centric formulation of pruning, by formulating it as a knapsack problem with parallelisation-aware tensor structures. The primary emphasis is on real-time inference, with latencies of order 1µs. We evaluate our method on a range of tasks, including jet tagging at CERN's Large Hadron Collider and fast image classification (SVHN, Fashion MNIST). The proposed method achieves reductions ranging between 55% and 92% in digital signal processing blocks (DSPs) and up to 81% in block memory (BRAM), with inference latencies ranging between 105ns and 205µs.

The proposed algorithms are integrated with hls4ml and open-sourced with an Apache 2.0 licence, enabling an end-to-end tool for hardware-aware pruning and real-time inference. Furthermore, the tools are readily integrated with QKeras, enabling pruning and inference of models trained with quantisation-aware training. Compared to TensorFlow Model Optimization, hls4ml Optimization API offers advanced functionality, including support for structured pruning, gradient-based ranking methods and integration with model reduction libraries, such as Keras Surgeon. Furthermore, by enabling multiple levels of pruning granularity, the software can target a wide range of hardware platforms. Through integration with hls4ml, an open-source, end-to-end system is built, allowing practitioners from a wide range of fields to compress and accelerate neural networks suited for their applications.

Authors: RAMHORST, Benjamin (Imperial College London); Prof. CONSTANTINIDES, George (Imperial College London); Dr LONČAR, Vladimir (Massachusetts Institute of Technology)

Presenter: RAMHORST, Benjamin (Imperial College London)

Session Classification: Contributed Talks

Track Classification: Contributed Talks