Fast Machine Learning for Science
Real-time and accelerated ML for fundamental sciences

Imperial College London
25-28 September 2023

Contribution ID: **30**

Type: **Standard Talk**

PolyLUT: Learning Piecewise Polynomials for Ultra-Low Latency FPGA LUT-based Inference

Wednesday 27 September 2023 15:00 (15 minutes)

Field-programmable gate arrays (FPGAs) are widely used to implement deep learning inference. Standard deep neural network inference involves the computation of interleaved linear maps and nonlinear activation functions. Prior work for ultra-low latency implementations has hardcoded the combination of linear maps and nonlinear activations inside FPGA lookup tables (LUTs). Our work is motivated by the idea that the LUTs in an FPGA can be used to implement a much greater variety of functions than this. In this paper, we propose a novel approach to training neural networks for FPGA deployment using *multivariate polynomials* as the basic building block. Our method takes advantage of the flexibility offered by the soft logic, hiding the polynomial evaluation inside the LUTs with zero overhead.

Our aim is to enable applications that require ultra-low latency real-time processing and highly lightweight on-chip implementations. We show that by using polynomial building blocks, we can achieve the same accuracy using considerably fewer layers of soft-logic than by using linear functions, leading to significant latency and area improvements. We demonstrate the effectiveness of this approach in three different tasks: network intrusion detection, handwritten digit recognition using the MNIST dataset, and jet identification at the CERN Large Hadron Collider. Compared to prior works, for similar accuracies, our method achieves significant latency improvements in these tasks, with reductions of up to $2\times$, $19.38\times$, and $3.57\times$, respectively.

Authors: ANDRONIC, Marta (Imperial College London); CONSTANTINIDES, George A. (Imperial College London)

Presenter: ANDRONIC, Marta (Imperial College London)

Session Classification: Contributed Talks

Track Classification: Contributed Talks