Contribution ID: **76** Type: **Lightning 5 min talk + poster**

An Efficient Multiply Accumulate Tree for Real-time Quantized Neural Networks

Tuesday 15 October 2024 16:15 (5 minutes)

Neural networks with a latency requirement at the order of microseconds, like the ones used at the CERN Large Hadron Colliders, are typically deployed on FPGAs fully unrolled. A bottleneck for the deployment of such neural networks is area utilization, which is directly related to the number of Multiply Accumulate (MAC) operations in matrix-vector multiplications.

In this work, we present the Multiply Accumulate Tree (MAC tree), an algorithm that optimizes the area usage of fully parallel vector-dot products on chips by exploiting self-similar patterns in the network's weights.

We implement the algorithm with the hls4ml library, a FOSS library for running real-time neural network inference on FPGAs, and compare the resource usage and latency with the original hls4ml implementation on different networks. The results show that the proposed MAC tree can achieve a reduction of LUT utilization by up to 50% in realistic quantized neural networks, while reducing the latency by up to a few folds. Furthermore, the proposed MAC tree provides an accurate estimation of the post-P&R resource utilization (error within ˜10%) and reasonably good latency estimation, which can be used during the design phase to optimize the neural networks.

Focus areas

Author: SUN, Chang (California Institute of Technology (US))

Co-authors: NGADIUBA, Jennifer (FNAL); Prof. SPIROPULU, Maria (California Institute of Technology)

Presenter: SUN, Chang (California Institute of Technology (US))

Session Classification: Lighting talks