Contribution ID: **1** Type: **Standard Talk**

ATHEENA: A Toolflow for Hardware Early-Exit Network Automation

Wednesday 27 September 2023 17:00 (15 minutes)

The continued need for improvements in accuracy, throughput, and efficiency of Deep Neural Networks has resulted in a multitude of methods that make the most of custom architectures on FPGAs. These include the creation of hand-crafted networks and the use of quantization and pruning to reduce extraneous network parameters. However, with the potential of static solutions already well exploited, we propose to shift the focus to using the varying difficulty of individual data samples to further improve efficiency and reduce average compute for classification. Input-dependent computation allows for the network to make runtime decisions to finish a task early if the result meets a confidence threshold. Early-Exit network architectures have become an increasingly popular way to implement such behaviour in software.
We create A Toolflow for Hardware Early-Exit Network Automation (ATHEENA), an automated FPGA toolflow that leverages the probability of samples exiting early from such networks to scale the resources allocated to different sections of the network. The toolflow uses the data-flow model of fpgaConvNet, extended to support Early-Exit networks as well as Design Space Exploration to optimize the generated streaming architecture hardware with the goal of increasing throughput/reducing area while maintaining accuracy. Experimental results on three different networks demonstrate a throughput increase of $2.00\times$ to $2.78\times$ compared to an optimized baseline network implementation with no early exits. Additionally, the toolflow can achieve a throughput matching the same baseline with as low as 46% of the resources the baseline requires.

Primary authors: BIGGS, Benjamin; Prof. BOUGANIS, Christos-Savvas (Imperial College London); Prof. CONSTANTINIDES, George (Imperial College London)

Presenter: BIGGS, Benjamin

Session Classification: Contributed Talks

Track Classification: Contributed Talks