Contribution ID: **70** Type: **Poster**

A gradient-based hardware-aware neural architecture search framework for hls4ml

Tuesday 15 October 2024 16:20 (5 minutes)

In software-hardware co-design, balancing performance with hardware constraints is critical, especially when using FPGAs for high-energy physics (HEP) applications with hls4ml. Limited resources and stringent latency requirements exacerbate this challenge. Existing frameworks such as AutoQKeras use Bayesian optimization to balance model size/energy and accuracy, but they are time-consuming, rely on early-stage training that can lead to inaccurate configuration evaluations, and often require significant trial and error. In addition, these metrics often do not reflect actual hardware usage.

In this work, we present a gradient-based Neural Architecture Search (NAS) framework tailored for hardware-aware optimization within the hls4ml workflow. Our approach incorporates practical hardware resource metrics into the search process and dynamically adapts to different HLS designs, tool versions, and FPGA devices. Unlike AutoQKeras, our design is fully trained during the search process, requiring only minimal fine-tuning afterwards. This framework allows users to efficiently explore trade-offs between model performance and hardware usage for their specific tasks in a single shot. Key contributions include: (1) a user-friendly interface for easy customization of the search space; (2) deep integration with hls4ml, allowing users to define and experiment with their own HLS synthesis configurations for FPGA; and (3) flexibility, allowing users to define custom hardware metrics for optimization, such as combinations of multiple FPGA resources.

We demonstrate the effectiveness of our approach using a 1.8M parameter convolutional neural network for an energy reconstruction task in calorimeters. Compared to the baseline model, the searched model achieved a 48.01% reduction in parameter count, as well as reductions in LUT usage of 29.73%, FF of 31.62%, BRAM of 16.06%, and DSP of 23.92%, with only a 0.84% degradation in MAE. The entire search process took approximately 2 GPU hours, demonstrating the potential of our framework to accelerate FPGA deployment in resource-constrained environments. Furthermore, this method can be extended beyond HEP to enable more efficient and scalable FPGA deployments in various fields, such as edge computing and autonomous systems.

Focus areas

HEP

Author: CHEN, ChiJui

Co-author: LAI, Bo-Cheng

Presenter: CHEN, ChiJui

Session Classification: Lighting talks