Contribution ID: **75** Type: **Lightning 5 min talk + poster**

S-QUARK: A Scalable Quantization-Aware Training Framework for FPGA Deployment based on Keras-v3

Tuesday 15 October 2024 15:30 (5 minutes)

In this work, we present the Scalable QUantization-Aware Real-time Keras (S-QUARK), an advanced quantization-aware training (QAT) framework for efficient FPGAs inference built on top of Keras-v3, supporting all Tensorflow, JAX, and PyTorch backends.

The framework inherits all perks from the High Granularity Quantization (HGQ) library, and extends it to support fixed-point numbers with different overflow modes and different parametrization of the fixed-point quantizers. Furthermore, it extends the HGQ library to support bit-accurate softmax and multi-head attention layers. Bit-exact minifloat quantizer with differentiable mantissa and exponent bits, as well as the exponent bias, are also supported.

On the TensorFlow and JAX backend, all layers provided by the framework support JIT compilation, which can significantly speed up the training process when the training process is io-bound. The speedup ranges from 1.5x to more than 3x compared to the HGQ framework, and has 10% to 100% overhead in training performance over the native TensorFlow or JAX with Keras implementation, depending on the exact model, dataset, and the hardware used.

The library is available under the LGPLv3 license at https://github.com/calad0i/s-quark.

Focus areas

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

Co-authors: NGADIUBA, Jennifer (FNAL); Prof. SPIROPULU, Maria (California Institute of Technology); LONCAR, Vladimir (Massachusetts Inst. of Technology (US))

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

Session Classification: Lighting talks