Machine Learning Acceleration — Quantization Process and Tools Development

Dennis Yin, Yihui Chen, Scott Hauck, Shih-Chieh Hsu, Elham E Khoda

Abstract

We are here to demonstrate the importance of quantization for machine learning model optimization and tools we implemented for doing so.

Quantization

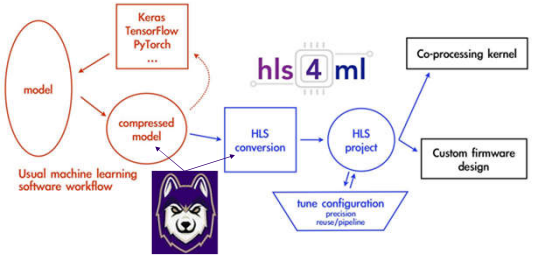

Fig1. Constantly updating supporting layers for different model

- QAT = Quantization Aware Training
- PTQ = Post Training Quantization
- Quantization: reducing bit width of numerical values

Why quantization?

1. **Faster Inference** 2. Lower Energy Consumption
3. **Compatibility with FPGA** 4. Low Memory Footprint

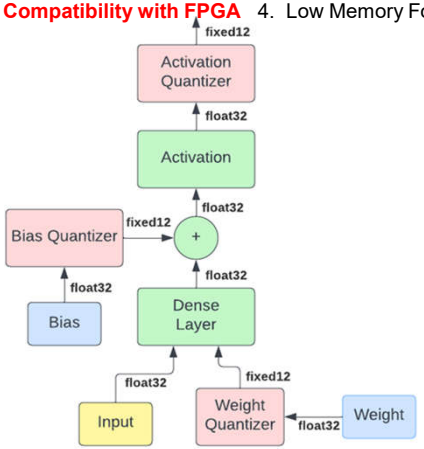

Fig2. Simulation of quantization process in QKeras

Recurrent Neural Network

Features of RNN layer:

- Process sequential data using recurrent structure
- Utilize information from previous elements
- Good for prediction

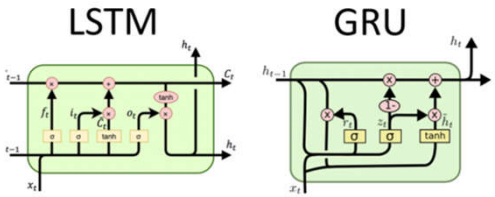

Fig3. The structure of two Different RNN layers: LSTM and GRU

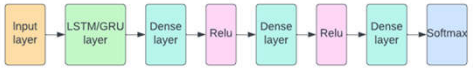

Fig4. The architecture of RNN model predicting human's quick drawing object (quickdraw dataset)

Quantization result:

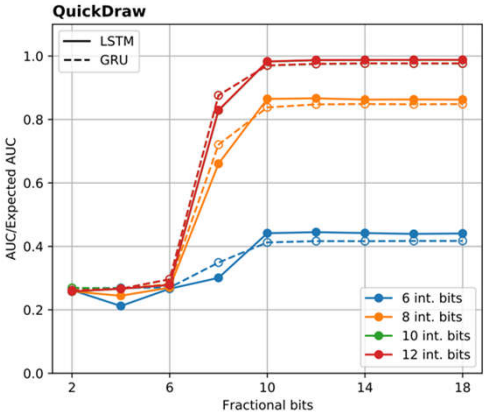

Fig5. QAT result for predicting quickdraw dataset using two different RNN layers

FPGA Latency VS GPU Latency:

Model	GPU [μs]	FPGA [μs]
GRU	NOT TESTED	35.4-164.0
LSTM	1515.15	35.9-164.0

Table1. The Throughput Latency Comparison between Nvidia Tesla V100 (GPU) and Xilinx Alveo U250 (FPGA)

Transformer Neural Network

Attention is all you need! :

- Process sequential data in parallel
- Context-aware processing
- Incorporates positional encoding
- Optimized for handling long-term dependencies

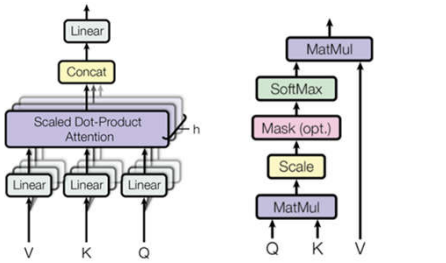

Fig6. The structure of MultiHeadAttention Layer

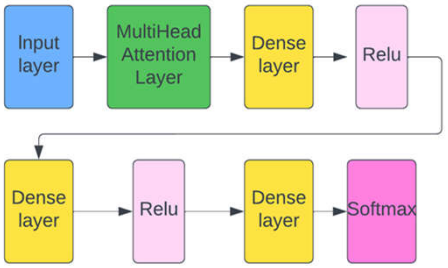

Fig7. The architecture of a Transformer model for detecting gravitational wave anomalies

Quantization result:

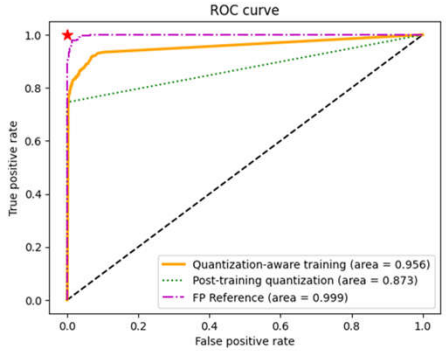

Fig8. Quantization result comparison between QAT, PTQ, and reference model based on the previous transformer model. A stronger inclination towards the red star indicates a better result.

Summary

- Quantization could lead to a reduction in model accuracy. Finding balance point is important.
- QAT is better than PTQ in general. Recommended to use QAT for final optimization.
- FPGA can have a much better latency performance compare to GPU.
- The packages are welcome for public usage.

HLS4ML

QKeras

Reference

- [1] E. E. Khoda et al., "Ultra-low latency recurrent neural network inference on FPGAs for physics applications with hls4ml," Machine Learning: Science and Technology, vol. 4, no. 2, p. 025004, 2023.
- [2] B. Moons, K. Goetschalckx, N. Van Berckelaer, and M. Verhelst, "Minimum energy quantized neural networks," in 2017 51st Asilomar Conference on Signals, Systems, and Computers, 2017: IEEE, pp. 1921-1925.
- [3] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.