ONNX Roadmap input on: Quantized Ops

Peter van Beek, Aleksandar Sutic, Thomas Gardos

Intel Corporation

Jan 2023

Background and Motivation

- Our goal:
  - Use ONNX as intermediate file format for fully quantized models
  - Produced by Neural Compressor or comparable tools with support for quantization
  - Consumed by a graph compiler/inference toolchain for further lowering steps and execution
  - Our domain of interest is embedded computer vision

- Current situation:
  - ONNX (onnx.ai): very limited set of Q-Ops
    - QLinearConv, QLinearMatMul
    - plus QuantizeLinear, DequantizeLinear, representing quantization and dequantization
  - ONNX Runtime contrib (com.microsoft): a wider set of Q-Ops
    - E.g., QLinearConcat, QLinearAveragePool, etc.
    - QLinearAdd but no QLinearSub
    - QLinearMul but no QLinearDiv

Recommendation

- Adopt a wider set of Q-Ops into ONNX (onnx.ai)
  - QLinearConcat
  - QLinearAdd/Sub/Mul/Div
  - QLinearAveragePool, QLinearGlobalAveragePool

- Robust support for Q-Ops to enable end-to-end toolchain with basic CNNs for embedded computer vision