Operators & quantized op

Adding bit parameter to QuantizeLinear and DequantizeLinear operators for improved hardware compatibility

More and more hardware support low bit quantization

But Onnx only supports INT8 quantization for the QuantizeLinear and DequantizeLinear operators

Risc V - low power ML

UNTETHER AI

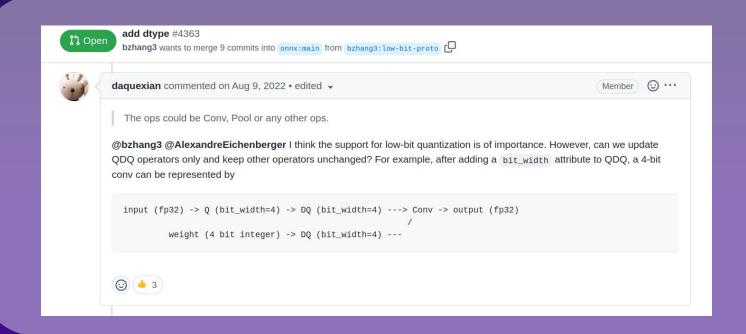


Ternary to INT8

INT4 to INT8

Binary & INT8

Already an MR that discusses what we want to do



The Advantages









Simple backward compatibility as the default would be int8

Can help in making ONNX a better framework for the tiny edges (MCU, FPGA, RISC V) that have harder requirements Not a massive MR, it won't reduce the size of the model but will help in the usability of the model

Will help in improving the experience of quantization on ONNX compare to TFLITE





- Datakalab -

MERCI!

Do you have any questions?

lf@datakalab.com

