Deriving Formula for mimicking TFLite model

1 Variables

Let f_i, f_w, f_b be the `fp32` inputs, weights, and biases (from the `fp32` model). Let q_i, q_w, q_b be the `int8/int32` inputs, weights, and biases (calculated from quantizing f_i, f_w, f_b using the respective quantization parameters). The values s_x, z_x are then the respective quantization scale and zero-point parameters.

To quantize a value:

$$q_x = f_x/s_x + z_x \tag{1}$$

To dequantize a value:

$$f_x = (q_x - z_x) * s_x \tag{2}$$

2 Dense Layer Calculation

The Dense layer calculation is then normally:

$$f_o = f_i \cdot f_w + f_b \tag{3}$$

2.1 Writing the calculation in terms of quantized values and parameters

Since `netron` shows quantized weights and biases, we assume that TFLite has quantized q_i, q_w, q_b. To recreate the TFLite model in tensorflow, we would first need to calculate q_i, q_w, q_b from quantizing f_i, f_w, f_b. Then, for f_o with type `fp32`:

$$\begin{aligned} f_o &= f_i \cdot f_w + f_b \\ &= ((q_i - z_i) * s_i) \cdot ((q_w - z_w) * s_w) + ((q_b - z_b) * s_b) \\ &= (q_i - z_i) \cdot (q_w - z_w) * s_i * s_w + ((q_b - z_b) * s_b) \end{aligned} \tag{4}$$