DSConv: Efficient Convolution Operator

Marcelo Gennari do Nascimento
University of Oxford
Active Vision Lab
marcelo@robots.ox.ac.uk

Roger Fawcett
Intel Corporation
https://www.omnitek.tv/about
roger.fawcett@intel.com

Victor Adrian Prisacariu
University of Oxford
Active Vision Lab
victor@robots.ox.ac.uk

1. Derivation of Equation (2)

The bit-size of the normal convolution is simply the number of floating point numbers in its tensor:

$$s_c = 32 \cdot C_o \cdot C_i \cdot K^2 \tag{1}$$

The bit-size of the *DSConv* is the sum of the VQK and the KDS:

$$s_d = b \cdot C_o \cdot C_i \cdot K^2 + 32 \cdot C_o \cdot \lceil \frac{C_i}{B} \rceil \cdot K^2 \tag{2}$$

The ratio is:

$$p = \frac{b \cdot C_o \cdot C_i \cdot K^2 + 32 \cdot C_o \cdot \lceil \frac{C_i}{B} \rceil \cdot K^2}{32 \cdot C_o \cdot C_i \cdot K^2} \tag{3}$$

$$p = \frac{b}{32} + \frac{\lceil \frac{C_i}{B} \rceil}{C_i} \tag{4}$$

2. Derivation of Equation (9)

The number of operations needed for normal convolution is:

$$T_{conv} = C_i \cdot C_o \cdot K^2 \cdot T_{FP} \tag{5}$$

Supposing that the `MAX`, `SHIFT` and `MASK` operations take ηT_{int}, where η is the ideality factor. The number of operations needed for *DSConv* is:

$$T_{ds} = C_i \cdot C_o \cdot K^2 \cdot T_{int}(1+\eta) + \lceil \frac{C_i}{B} \rceil \cdot C_o \cdot K^2 \cdot T_{FP} \tag{6}$$

So the time taken for the *DSConv* should be less than normal convolution:

$$T_{ds} \leq T_{conv} \tag{7}$$

From which we find that:

$$T_{int} \leq T_{FP} \frac{C_i - \lceil \frac{C_i}{B} \rceil}{C_i(1+\eta)} \tag{8}$$

3. Additional Results

Table 1 shows additional results of quantizing the more compact networks MobileNetV1, MobileNetV2, ShuffleNetV2, and a more accurate version of GoogleNet. Notice that as expected, more compact networks (particularly the ones using depth-wise convolution) are prone to higher accuracy loss, as a result of lower redundancy.

Block	-	256	128	16	4	128	64	64	16
Bit (W/A)	32/32	8/32	6/32	4/32	2/32	8/8	6/6	5/5	4/8
MNetV1	69.6	69.4	66.2	48.6	4.1	69.5	66.3	50.7	48.4
	89.1	88.9	86.8	72.8	11.4	88.9	86.8	74.9	72.7
MNetV2	71.9	71.8	71.2	63.5	4.3	71.7	70.7	65.8	50.3
	90.3	90.3	89.9	85.0	12.3	90.3	89.5	86.7	73.8
SNetV2	69.3	69.3	68.7	62.0	27.4	69.3	68.4	64.7	61.9
	88.3	88.3	87.8	83.2	48.5	88.3	87.6	85.3	83.1
GNet	69.8	69.8	68.9	65.3	31.2	69.8	69.0	66.7	65.3
	89.5	89.5	89.2	86.4	54.3	89.5	89.1	87.6	86.5

Table 1. Results of MobileNetV1 (MNetV1), MobileNetV2 (MNetV2), ShuffleNetV2 x1.0 (SNetV2) and GoogLeNet (GNet) after quantization. For each architecture, the top result is the Top1 and the bottom is the Top5 result (all in %).

Figure 1 shows the distribution of weight values for the VQK, KDS and resulting tensor compared to the original Weights of the first layer of ResNet50.

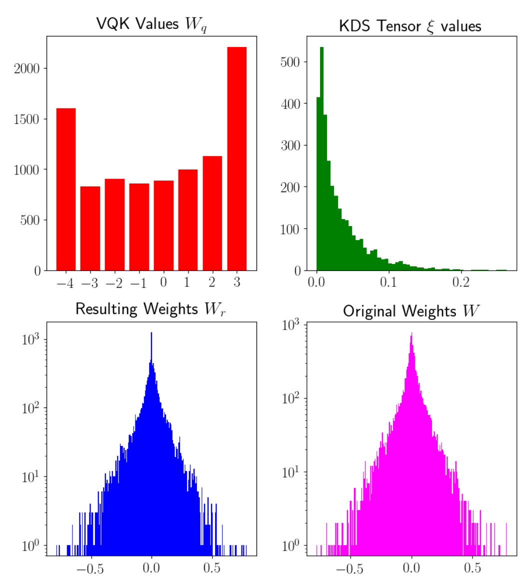

Figure 1. Result of quantizing the first layer of Resnet50 using $b = 3$ and $B = 128$ compared to original weights.