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:

\[ s_c = 32 \cdot C_o \cdot C_i \cdot K^2 \]  

(1)

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

\[ s_d = b \cdot C_o \cdot C_i \cdot K^2 + 32 \cdot C_o \cdot \left\lceil \frac{C_i}{B} \right\rceil \cdot K^2 \]  

(2)

The ratio is:

\[ p = \frac{b \cdot C_o \cdot C_i \cdot K^2 + 32 \cdot C_o \cdot \left\lceil \frac{C_i}{B} \right\rceil \cdot K^2}{32 \cdot C_o \cdot C_i \cdot K^2} \]  

\[ p = b + \frac{\left\lceil \frac{C_i}{B} \right\rceil}{C_i} \]  

(4)

2. Derivation of Equation (9)

The number of operations needed for normal convolution is:

\[ T_{conv} = C_i \cdot C_o \cdot K^2 \cdot T_{FP} \]  

(5)

Supposing that the MAX, SHIFT and MASK operations take \( \eta T_{int} \), where \( \eta \) is the ideality factor. The number of operations needed for DSConv is:

\[ T_{ds} = C_i \cdot C_o \cdot K^2 \cdot T_{int}(1+\eta) + \left\lceil \frac{C_i}{B} \right\rceil \cdot C_o \cdot K^2 \cdot T_{FP} \]  

(6)

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

\[ T_{ds} \leq T_{conv} \]  

(7)

From which we find that:

\[ T_{int} \leq T_{FP} \frac{C_i - \left\lceil \frac{C_i}{B} \right\rceil}{C_i(1+\eta)} \]  

(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.

![Figure 1](image.png)

Figure 1. Result of quantizing the first layer of ResNet50 using \( b = 3 \) and \( B = 128 \) compared to original weights.