Post training 4-bit quantization of convolutional networks for rapid-deployment
Ron Banner, Yury Nahshan and Daniel Soudry

- 4-bit Post training quantization of weights and activations
  - No retraining
  - No data set

<table>
<thead>
<tr>
<th>Method</th>
<th>VGG</th>
<th>VGG-BN</th>
<th>IncvY3</th>
<th>Res18</th>
<th>Res50</th>
<th>Res101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>63.2%</td>
<td>64.9%</td>
<td>30.6%</td>
<td>51.1%</td>
<td>62.0%</td>
<td>62.6%</td>
</tr>
<tr>
<td>All methods combined</td>
<td>70.5%</td>
<td>71.8%</td>
<td>66.4%</td>
<td>67.0%</td>
<td>73.8%</td>
<td>75.0%</td>
</tr>
<tr>
<td>Reference (FP32)</td>
<td>71.6%</td>
<td>73.6%</td>
<td>77.2%</td>
<td>69.7%</td>
<td>76.1%</td>
<td>77.3%</td>
</tr>
</tbody>
</table>

Analytical Clipping

Let $f(x)$ be the function to be quantized. The analytical clipping is given by:

$$
\text{Analytical Clipping}
$$

$$
\sum_{q=0}^{2^B-1} \int_{-\alpha}^{\alpha} f(x) \cdot (x - q) dx
$$

$$
\int_{-\alpha}^{\alpha} f(x) \cdot (x + a) dx
$$

Bit-allocation

$$
M_i = \left[ \log_2 \left( \frac{\alpha_i}{\sum_i \alpha_i} \cdot B \right) \right]
$$

Bias-correction

$$
\mu_c = \mathbb{E}(W_c) - \mathbb{E}(W^q_c)
$$

$$
\xi_c = \frac{||W_c - \mathbb{E}(W_c)||_2}{||W^q_c - \mathbb{E}(W^q_c)||_2}
$$

$$
w \leftarrow \xi_c (w + \mu_c), \quad \forall w \in W^q_c
$$