Abstract

We study the problem of quantizing $N$ sorted, scalar datapoints with a fixed codebook containing $K$ entries that are allowed to be rescaled. By studying the properties of the optimal quantizer, we derive an $O(N \log K)$ algorithm that is guaranteed to find the optimal quantization parameters for any fixed codebook regardless of data distribution. We apply our algorithm to synthetic and real-world neural network quantization problems and demonstrate the effectiveness of our approach.

Problem formulation

Given a sorted data vector $w$ with elements $w_1 ≤ w_2 ≤ ... ≤ w_N$ and a fixed codebook $C = \{c_1, c_2, ..., c_K\}$ such that $c_1 < c_2 < ... < c_K$ we would like to learn the optimal rescaling of the codebook and the assignments ($z$) of datapoints into the rescaled codebook defined by the following MSE problem:

$$\min_{\alpha, z} \text{Loss}(\alpha, z) = \frac{N}{K} \sum_{k} z_k w_k - \alpha c_k^2 \quad \text{s.t.} \quad z_k^2 \leq 1, \quad z_k \in \{0, 1\}^K.$$  \hspace{1cm} (1)

Below is an illustration of the problem setup:

\[\begin{array}{ccc}
C_1 & C_2 & C_3 \\
\times & \times & \times \\
\end{array}\]

\[\begin{array}{cccc}
m_1 & c_1 & c_2 & c_3 \\
\times & \times & \times & \times \\
\end{array}\]

The problem includes many important cases:
1. symmetric INT8 quantization: $C = \{0, \pm 127\}
2. powers-of-two quantization: $C = \{0, \pm 2^1, \pm 2^2, \pm 2^3\}
3. log-scale codebooks: $C = \{0, \log 2, \log 3, \log 4, ..., \log s\}
4. and many others!

Locally optimal solutions

Let us analyze the loss wrt $\alpha$ and $Z$ separately keeping other variable fixed.
- **Locally optimal scale for a fixed $Z$**:
  $$\min_{\alpha} \text{Loss}(\alpha, Z) \Rightarrow \alpha = \text{OPTSCALE}(Z)$$
- **Locally optimal assignments for a fixed $\alpha$**:
  $$\min_{Z} \text{Loss}(\alpha, Z) \Rightarrow Z = \text{OPTASSIGNMENT}(\alpha)$$

Exact forms of $\text{OPTSCALE}(Z)$ and $\text{OPTASSIGNMENT}(\alpha)$ are in the paper.

Characterization of global solution

The global solution $(\alpha^*, Z^*)$ must satisfy

$$\alpha^* = \text{OPTSCALE}(Z^*) \quad \text{and} \quad Z^* = \text{OPTASSIGNMENT}(\alpha^*) \quad \text{(2)}$$

There are many pairs that satisfy the equation above, but the one obtaining the minimal loss is the $(\alpha^*, Z^*)$-pair we are looking for.

Important result and an initial algorithm

Lemma: The number of fixed points satisfying (2) is at most $NK + 1$.

Proof idea: Optimal assignments will remain constant for certain region of $\alpha$-values (see paper)

Corollary: There are only $NK + 1$ values of $\alpha$ we need to check!

Initial $O(NK^2 \log N)$ algorithm enumerating possible $\alpha$-values.

1. function $\text{OPTQUANT}(w, C)$
2. $Z_k^* = (z_1^*, ..., z_{NK+1}^*) = (0, 0, ..., 0)$
3. $\alpha_k^* = \infty, \alpha_K^* = -\infty$ for $k = 1, \ldots, K - 1$
4. for $k = 1, \ldots, K - 1$
5. $m_n = \frac{\alpha_k + \alpha_{k+1}}{2}$
6. for $n = 1, \ldots, N$ do
7. for $k = 1, \ldots, K - 1$
8. $Z_k^* = \text{OPTASSIGNMENT}(\alpha)$
9. $\alpha_k^* = \text{OPTSCALE}(Z^*)$
10. $L = \text{Loss}(\alpha, Z^*)$
11. if $L \leq \alpha_k^*$ then
12. $\alpha_k^* = L$
13. $Z_k^* = Z_k^*$
14. return $\alpha_0^* = Z_0^*$

Experiments on synthetic data

Quantization of $N = 10000$ datapoints sampled from the Gaussian mixture distribution. The probability density function of the distribution is plotted on the left. For this particular distribution our algorithm achieves optimal MSE error which outperforms other baselines (min-max and alt-opt).

Experiments on neural network compression

IN8 quantization results, top-1 accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>FP32</th>
<th>Min-max weight with activation calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet-v1</td>
<td>71.88</td>
<td>76.16</td>
</tr>
<tr>
<td>MobileNet-v2</td>
<td>71.88</td>
<td>76.16</td>
</tr>
<tr>
<td>ResNet18</td>
<td>69.74</td>
<td>76.16</td>
</tr>
<tr>
<td>ResNet50</td>
<td>76.16</td>
<td>76.16</td>
</tr>
</tbody>
</table>

INT4 quantization results, top-1 accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>FP32</th>
<th>Min-max weight with activation calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet-v1</td>
<td>71.88</td>
<td>76.16</td>
</tr>
<tr>
<td>MobileNet-v2</td>
<td>71.88</td>
<td>76.16</td>
</tr>
<tr>
<td>ResNet18</td>
<td>69.74</td>
<td>76.16</td>
</tr>
<tr>
<td>ResNet50</td>
<td>76.16</td>
<td>76.16</td>
</tr>
</tbody>
</table>