GE Energy Efficient Machine Learning at the Edge

Ji In Choi, Madeleine Georges, Julia Shin, Olivia Wang, Tiffany Zhu
Mentor: Tapan Shah
Overview

1. Introduction
2. Base Quantization Methods
3. Stochastic Quantization Methods
4. Next Steps
Introduction

● How can we make AI carbon efficient?
● Low Precision ML
  ○ Can we learn a model from training data with 1-8 bit precision (as compared to 32-64 bits)?
Quantization

What is Quantization?

Convert data from $q$-bits to $p$-bits where $q > p$

\[ [0, 3.5, 3.8, 4.8, 5.5, 5.6, 7.5, 8] \]

\[ [0, 2, 2, 4, 4, 4, 6, 6] \]

Why Quantize?

Reduce storage by 20 times
Base Quantization Methods

- Simple Quantizer
- Simple Quantizer by Column
- Quantile Quantizer
- Quantile Quantizer by Column
Base Quantization Methods

Simple Quantizer
- Use minimum and maximum values in the entire dataset
- Create $2^p$ bins with uniform range (i.e. equal bin widths)
- E.g.

Simple Quantizer By Column
- Same as Simple Quantizer but for each column
Base Quantization Methods

Quantile Quantizer
- Use different quantiles over the entire dataset
- Create $2^p$ bins with *unequal* range (i.e. unequal bin widths)
- May help account for distribution of dataset
- E.g.

Quantile Quantizer by Column
- Same as Quantile Quantizer but for each column
Base Quantization Methods - Results

- Ran the 4 base quantizers on datasets from OpenML
  - QuantileQuantizerByCol may be the best quantizer
Base Quantization Methods - Results

- Using paired t-test
  \[ H_0 : a_q - a_0 = 0 \]
  \[ H_1 : a_q - a_0 \neq 0 \]

- Final results:
  1. By **column** quantizers work better
  2. **Quantile** quantizers work better than the Simple quantizers
  3. **QuantileQuantizerByColumn** is best of the base quantizers

<table>
<thead>
<tr>
<th>Bits</th>
<th>QuantileQuantizerByColumn vs Simple Quantizer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t-Statistic</td>
</tr>
<tr>
<td>1</td>
<td>5.9405</td>
</tr>
<tr>
<td>2</td>
<td>6.1571</td>
</tr>
<tr>
<td>3</td>
<td>5.5912</td>
</tr>
<tr>
<td>4</td>
<td>4.2122</td>
</tr>
<tr>
<td>5</td>
<td>4.2644</td>
</tr>
<tr>
<td>6</td>
<td>2.5539</td>
</tr>
<tr>
<td>7</td>
<td>3.1717</td>
</tr>
<tr>
<td>8</td>
<td>2.4289</td>
</tr>
</tbody>
</table>
The process of using a stochastic method is:

1. Given dataset $X$, create bins using a deterministic base quantizer
   - We use `QuantileQuantizerByColumn` in our experiments

2. Apply a stochastic process to $X$ to put each data point into the bins generated from step 1.

Goal: lower the quantization error by introducing randomness

Two methods considered: Dithering and Stochastic Rounding
Stochastic Quantization Methods

Dithering
- Random noise is added to each input value
- Assuming data is scaled then for each data point $X_i$:

$$X_i \leftarrow X_i + \text{noise}$$

$\text{noise} \sim \text{Uniform}(-\text{bin}_\text{width}/2, \text{bin}_\text{width}/2)$

![Diagram showing dithering process](image)
Stochastic Quantization Methods

Stochastic Rounding

- The input value is rounded to one of bordering quantization levels with probability dependent on proximity.
- Given a point $X_i$ and its neighboring upper bin $U$ and lower bin $L$:

$$\text{round}(X_i) = \begin{cases} U & \text{with probability } \frac{(X_i-L)}{(U-L)} \\ L & \text{with probability } \frac{(U-X_i)}{(U-L)} \end{cases}$$
Stochastic Quantization Method - Results

Classification Data
- 55 classification datasets from OpenML
- Logistic Regression
Stochastic Quantization Method - Results

Classification Data

- Dithering is better for 1 bit
- Stochastic Rounding is better overall

- No clear pattern between improvement in accuracies and dataset attributes
Stochastic Quantization Method - Results

Regression Data
- 30 regression datasets from OpenML
- Ridge Regression
Stochastic Quantization Method - Results

Regression Data

- Both dithering and Stochastic Rounding work better on regression data

- No clear pattern between improvement in accuracies and dataset attributes
Conclusions

- **Base Quantizers:**
  - Using *quantiles* for each *column* is the best of the base methods

- **Stochastic Quantizers:**
  - Dithering and stochastic rounding can further improve quantization
  - Dithering and stochastic rounding work better on *regression* datasets than classification datasets
  - *Stochastic rounding* works better than dithering
Next Steps

**Mini-batch Learning**
Leveraging reinforcement learning to iteratively update quantizer

**Precision Reallocation**
Using linear programming to reallocate available bits to each feature

**Deep Learning Models**
Evaluate Quantization Methods on more complex, non-linear models
References

Thanks!

Questions?

jic2124
mg4128
js5569
yw3324
tz2196
@columbia.edu