Faster and Stronger Lossless Compression with Optimized Autoregressive Framework

Yu Mao*, Jingzong Li*, Yufei Cui†, and Jason Chun Xue*

*Department of Computer Science, City University of Hong Kong
†School of Computer Science, McGill University
Why Data Compression?

Projection by IBISWorld by 1990

Total Internet Traffic Volume (Exabytes per month)

20.9% growth

Observation by Cloudflare in 2020

40% growth in 3 months

January 2020
April 2020

[https://blog.cloudflare.com/recent-trends-in-internet-traffic/]

We need a stronger compression algorithm to deal with the rapid growing trend.
### Lossless compressors

#### General-Purpose Lossless Compressors

<table>
<thead>
<tr>
<th>Traditional</th>
<th>Deep-learning based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gzip, 7z, Zstandard</td>
<td>Cmix, NNCP, Dzip, TRACE</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Compression Ratio comparisons between traditional methods and deep-learning methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Homogeneous Data</th>
<th>Heterogeneous Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enwik9  Book Sound Image Float</td>
<td>Silesia Backup</td>
</tr>
<tr>
<td>Gzip</td>
<td>3.09  2.77  1.37  1.14  1.06</td>
<td>3.10  1.28</td>
</tr>
<tr>
<td>7z</td>
<td>4.35  3.80  1.59  1.38  1.14</td>
<td>4.25  1.56</td>
</tr>
<tr>
<td>Zstd-19</td>
<td>4.24  3.73  1.40  1.16  1.10</td>
<td>3.97  1.36</td>
</tr>
<tr>
<td>Dzip</td>
<td>4.47  3.95  2.04  1.72  1.26</td>
<td>4.78  1.78</td>
</tr>
<tr>
<td>TRACE</td>
<td>5.29  4.58  2.16  1.81  1.28</td>
<td>4.63  1.78</td>
</tr>
<tr>
<td>OREO</td>
<td>5.68  4.94  2.25  1.86  1.28</td>
<td>4.86  1.87</td>
</tr>
<tr>
<td>PAC</td>
<td>5.97  5.05  2.25  1.96  1.29</td>
<td>4.99  1.92</td>
</tr>
</tbody>
</table>

- Deep-learning based compressor can obtain **much higher performance** than traditional methods, but with **significantly slow compression speeds**.
To Compress 1GB Data:

None Deep-learning compressor needs Several tens of seconds

- Gzip → 33s
- 7z → 60s
- Zstd-19 → 321s
- Zstd-FPGA → 1.28s
- LPAQ → 85s

Deep-learning compressor needs several several days

- Cmix → 25 days
- Tensorflow-compress → 8 days
- NNCP → 8 days
- Dzip → 1.7 days
- TRACE → 14h
- OREO → 10h
- PAC → 5h (2080Ti)
Two blind spots of current NN-based compressors

\[ P_e(x_{i-2})P_e(x_{i-1})P_e(x_i)P_e(x_{i+1}) \]

- Duplicated processing problem
- In-batch distribution variation problem
Duplicated processing problem

PE: Probability Estimation,
FE: Feature Extraction,
FM: Feature Mixture

Previous framework

Our framework

GPU

CPU
In-batch distribution variation problem

Batch-aware model design: N set of parameter in individual layer corresponds to four item in a batch.
Learned Ordered Mask

A trainable 1D vector is introduced to dynamically learn the order information. The ordered importance is modeled as:

\[ F(x_{i-1}), \ldots, F(x_{i-k}) = W \ast \{F(x_{i-1}), \ldots, F(x_{i-k})\} \]

where \( F(x_{i-1}) \) is the extracted feature of \( x_{i-1} \) and \( W \) is learned ordered importance.

This makes PAC’s probability estimator a pure MLP architecture, which gives possibility for current general-purpose compressor to implement on other hardware.
Performance Evaluation

<table>
<thead>
<tr>
<th>Compressor</th>
<th>Peak GPU Memory Usage (GB)</th>
<th>Inference (ms)</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNCP</td>
<td>7.75</td>
<td>95.67</td>
<td>$15.83 \times 10^{10}$</td>
</tr>
<tr>
<td>Dzip</td>
<td>6.39</td>
<td>5.82</td>
<td>$7.48 \times 10^{10}$</td>
</tr>
<tr>
<td>TRACE</td>
<td>2.02</td>
<td>2.08</td>
<td>$0.34 \times 10^{10}$</td>
</tr>
<tr>
<td>OREO</td>
<td>1.18</td>
<td>1.54</td>
<td>$0.12 \times 10^{10}$</td>
</tr>
<tr>
<td>PAC</td>
<td>1.07</td>
<td>1.54</td>
<td>$0.1 \times 10^{10}$</td>
</tr>
</tbody>
</table>

(a) Host-GPU data transmit time. (b) Ordered mask generation time. (c) Individual layer inference time.
Future Direction

- Future directions in the field of compression could include:
  - 1. Specialized Hardware Acceleration
  - 2. Hybrid Compression Approaches
  - 3. Fast Image Compression—We are working on it!
Our works


Thank you!

JULY 9–13, 2023
MOSCONET WEST CENTER
SAN FRANCISCO, CA, USA