On the Downstream Performance of Compressed Word Embeddings

Avner May, Jian Zhang, Tri Dao, Christopher Ré
Department of Computer Science, Stanford University
{avnermay, zjian, trid, chrismre}@cs.stanford.edu

Overview

Word embeddings:
- Important for strong NLP performance
- Take a lot of memory

Common Solution: Compression (e.g., 32-bit → 1-bit)

Key question: What determines the performance of downstream models trained with compressed word embeddings?

Contribution:
- A new compression quality measure which
  - Is theoretically related to downstream perf.
  - Empirically correlates with downstream perf.
  - Can efficiently identify compressed embeddings with strong downstream perf. w/o model training.

Motivating Observations

Observation #1
Existing metrics (e.g. PIP loss [1]) fail to explain relative downstream performance across compression methods.

Observation #2
A simple compression method (uniform quantization) can match more complex ones (e.g., DCCL [4], k-means [5]).

Theoretical Results

Theorem 1 (informal): Generalization & EOS
For fixed design linear regression, if \( \hat{y} \in \mathbb{R}^n \) is a random label vector in \( \text{span}(U) \), then

\[
\mathbb{E}_{\hat{y}} \left[ R_y(\hat{X}) - R_y(X) \right] = \mathcal{O}\left( 1 - \text{EOS}(X, \hat{X}) \right).
\]

The compressed embedding's model accuracy can be expressed in terms of EOS.

Theorem 2 (informal): Uniform Quantization Bound
Let \( \hat{X} \) be a \( b \)-bit uniform quant. of \( X \). To achieve \( \text{EOS} \geq 1 - \epsilon \), \( \hat{X} \) requires a logarithmic \# of bits

\[
b = \mathcal{O} \left( \log_2 \left( \frac{1}{\epsilon} \right) \right).
\]

Uniform quantization can attain high EOS with low precision.

Experiments

Correlation of EOS with Downstream Performance

Spearman rank correlation

<table>
<thead>
<tr>
<th>Embedding</th>
<th>fastText</th>
<th>BERT</th>
<th>WordPiece</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIP loss [1]</td>
<td>0.34</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>( \Delta ) [2]</td>
<td>0.31</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>( \Delta_{\text{max}} ) [3]</td>
<td>0.72</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>EOS (Ours)</td>
<td>0.91</td>
<td>0.92</td>
<td></td>
</tr>
</tbody>
</table>

EOS correlates strongly with downstream performance.

Uniform Quantization Performance

Uniform quantization matches the more complex methods.

Resources and References

Resources:
- GitHub: https://github.com/HazyResearch/smallfry

Code:
- https://github.com/HazyResearch/smallfry

References