On the Downstream Performance of Compressed Word Embeddings

Avner May, Jian Zhang, Tri Dao, Chris Ré
Stanford University
Word Embeddings

Word embeddings take a lot of memory
Word Embedding Compression

Critical for deployment *under memory constraints*

- Deep compositional code learning (DCCL) \(^1\)
- Kmeans \(^2\)
- Uniform quantization \(^3\)
- Dimension reduction (e.g. PCA) \(^4\)

1. Shu et al. 2017  
2. Andrews et al. 2015  
3. Gersho et al. 1977  
4. Pearson et al. 1901
What determines the *model accuracy* attained by different *compressed word embeddings*?

Can the insights guide the selection of *compressed word embeddings* under *memory constraints*?
Existing quality measures

Can’t explain the relative model accuracy across compression methods

1. Yin et al. 2018
Setting to derive a new quality measure

Word vector as feature

Compressed word embedding

Regression label

\( X \in \mathbb{R}^{n \times d} \)

\( y \in \mathbb{R}^n \)

Model accuracy

Test mean square error (MSE) rel. to uncompressed embedding
In the setting of linear regression

Fixed design linear regression (simple and classic setup):\(^{1,2,3}\)
Same set of data points for train and test; noisy training label; noiseless test label

\[
\text{Test time prediction} = UU^T y
\]

Compressed word embedding \(X \in \mathbb{R}^{n \times d}\)
SVD \(X = U\Lambda V^T\)
Training label \(y \in \mathbb{R}^n\)

Observation
Prediction highly depends on \(U, the left singular vectors\)

1. Avron et al. 2018  
2. Bach et al. 2013  
3. Cortes et al. 2010
A new quality measure of compression word embedding

Eigenspace overlap (EO)

\[ \mathcal{E}(X, \tilde{X}) := \frac{1}{\max(d,k)} \| U^T \tilde{U} \|^2_F \]

Compressed \( X \in \mathbb{R}^{n \times d} \) uncompressed \( \tilde{X} \in \mathbb{R}^{n \times k} \)

SVD \( X = U \Lambda V^T \), \( \tilde{X} = \tilde{U} \tilde{\Lambda} \tilde{V}^T \)

**Intuition**

More *similar left singular vectors*, better model acc. relative to uncompressed embeddings
In the setting of *linear regression*

\[
\mathbb{E}_Y \left[ R_Y(\hat{X}) - R_Y(X) \right] = O\left(1 - \mathcal{E}(X, \hat{X})\right)
\]

Target label vector sampled from \( \text{Span}(U) \)

Uncompressed embedding \( X \)

Compressed embedding \( \hat{X} \)

**Theory connection** (sketch)

*Model acc.* can be bounded in terms of *eigenspace overlap*
Empirical correlation beyond the regression setting

Empirical correlation
EO attains better correlation with downstream model acc.
What determines the model accuracy attained by different compressed word embeddings?

Can the insights guide the selection of compressed word embeddings under memory constraints?
Eigenspace overlap as a selection criterion

Selecting the right embedding → better model acc. under memory budgets

Case study

Eigenspace overlap vs. PIP loss → higher acc. at 32X compression

Lower quality

32X compression

3.5 pt. higher acc.
Eigenspace overlap as a selection criterion

Selection error

Fraction of cases when *failing to select* the embedding with *better model acc.*

Utility under memory budgets

Up to 2X lower selection error at up to 32X compression

1. Yin et al. 2018
2. Avron et al. 2018
Summary

Theoretical connection in a regression setting

Empirical correlations in a wide range of models / tasks

Guide the selection of compressed word embeddings

Left singular vector is important, EO captures it

Utility under memory constraints
THANK YOU!

Spotlight: Thursday, Dec 12, 4:05 pm  
Poster: Thursday, Dec 12, 5-7 pm