On the Downstream Performance of Compressed Word embeddings

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

Word embedding

Pretraining

Downstream models

1

Apple	2.09	-0.98	1.48	0.09
Pear	0.05	-0.14	-1.08	2.12
Cat	-0.91	1.92	0.68	-1.03
Dog	1.87	0.45	1.53	1.49

Word embeddings

2

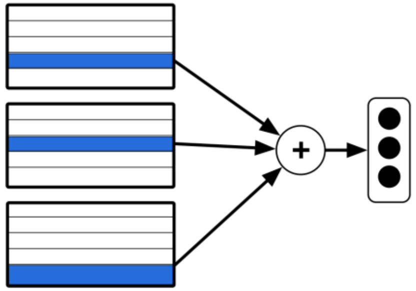
Word embedding is a memory-intensive feature representation

1. Freepik 2. slideshare

Word embedding compression

Compression is critical for deployment under memory budget

- Deep compositional code learning (DCCL) [1]



- Kmeans [2]

2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

3	0	2	1
1	1	0	3
0	3	1	0
3	0	2	2

Centroid index

1.96
1.48
-0.04
-0.97

Centroid

- Dimension reduction (e.g. PCA) [3]

- Uniform quantization [4]

2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

3	0	2	1
1	1	0	3
0	3	1	0
3	0	2	2

Centroid index

2.00
0.67
-0.67
-2.00

Centroid

Equal interval

1. Shu et al. 2017   2. Andrews et al. 2015   3. Pearson et al. 1901   4. Pearson et al. 1901

Key research questions

What determines the ***model accuracy*** of models trained with ***Compressed word embeddings?***

&

How to optimize the ***model accuracy*** under ***memory budgets*** for the ***compressed word embeddings?***

A new quality measure of compression word embedding

Eigenspace overlap (EO)

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

Uncompressed and compressed embedding $X \in \mathbb{R}^{n \times d}$ $\tilde{X} \in \mathbb{R}^{n \times k}$

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

Intuition

More similar ***spans of left singular vectors***,
better model acc. relative to uncompressed embeddings

In the context of ***fixed design linear regression***

Test MSE of fixed design regressors

$$\mathbb{E}_{\bar{y}}\left[\mathcal{R}_{\bar{y}}(\tilde{X}) - \mathcal{R}_{\bar{y}}(X)\right] = \mathcal{O}\Big(1 - \mathcal{E}(X, \tilde{X})\Big)$$

Label vector sampled from $\mathrm{Span}(U)$

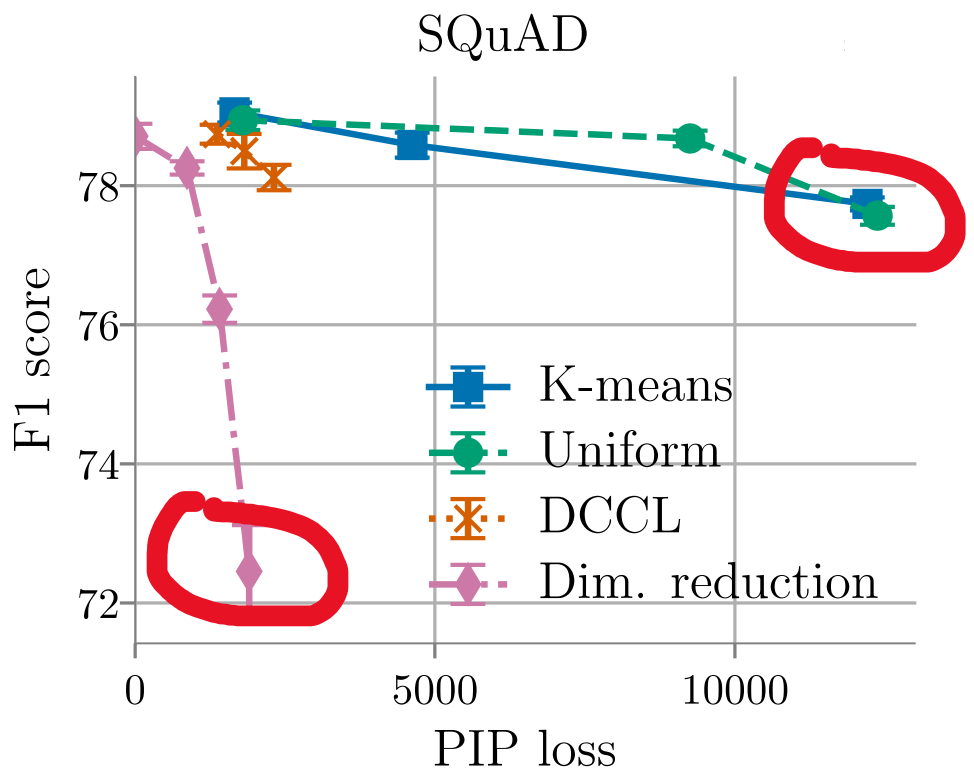
Uncompressed embedding X

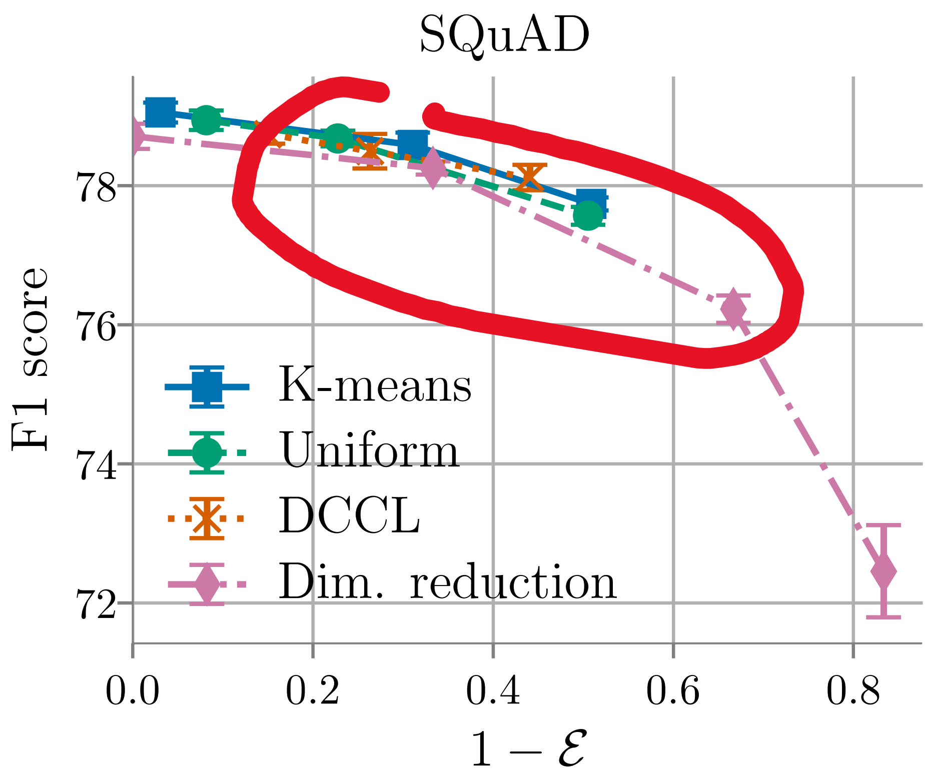
Compressed embedding $\tilde{X}$

Theory sketch

Model acc. can be bounded in terms of ***eigenspace overlap***

Beyond ***fixed design regression***

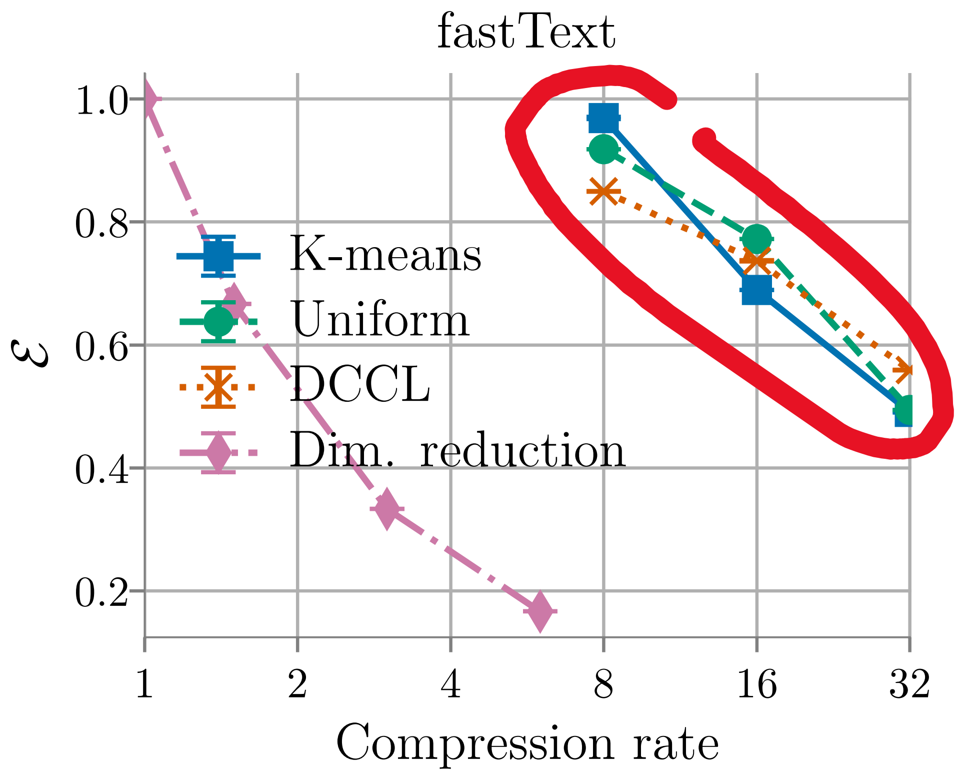




Empirical observation

EO attains ***better correlation*** with downstream ***model acc.***

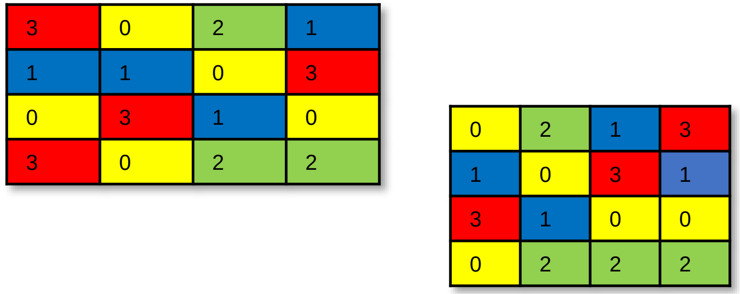
Beyond ***fixed design regression***



Empirical observation

EO explains the ***strong performance*** *of* simple ***uniform quantization***

Eigenspace overlap as an embedding selection criterion



Which compressed word embedding attains better model accuracy?

Table 1. Selection error rate of quality measures as embedding selection criteria

Dataset	SQuAD		SST-1		MNLI	QQP
Embedding	GloVe	fastText	GloVe	fastText	BERT WordPiece	BERT WordPiece
PIP loss [1]	0.32	0.37	0.32	0.40	0.31	0.32
Δ [2]	0.34	0.58	0.39	0.57	0.32	0.33
$1-\mathcal{E}$	**0.17**	**0.11**	**0.19**	**0.20**	**0.10**	**0.10**

Utility

Up to **2X lower selection error** than existing quality measures

1. Yin et al.

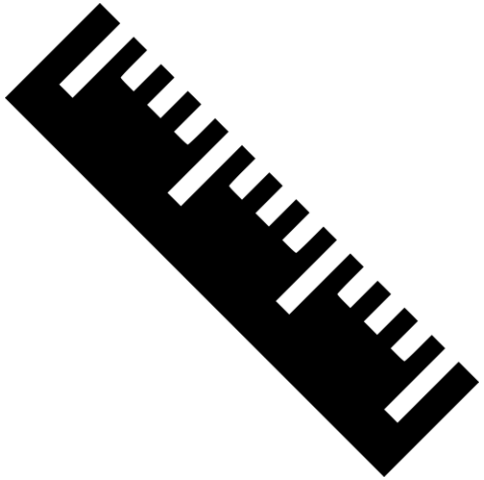
2. Avron et al.

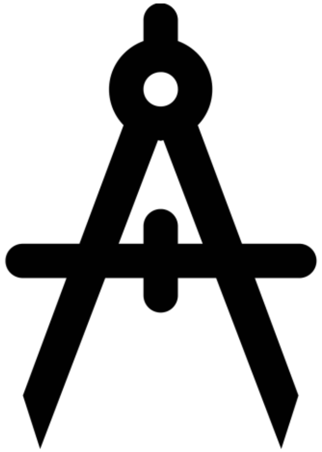
Summary

Theoretical connection b/w ***eigenspace overlap*** & ***model acc.*** for ***FDR*** setting

Strong empirical correlations b/w ***eigenspace overlap*** & ***model acc.*** beyond ***FDR***

Guide selection of compressed embeddings with ***improved model acc.***





Understanding the statistical performance

Optimizing under memory budgets