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.

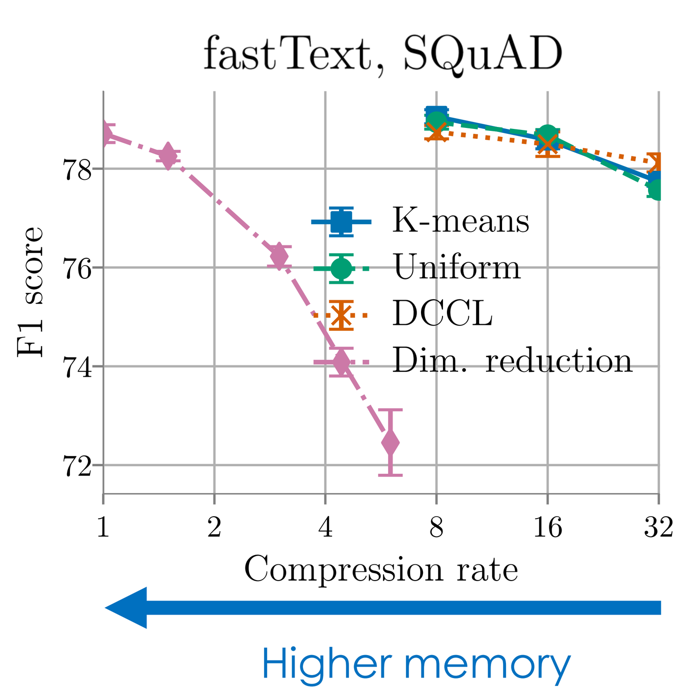
Motivating Observations

Observation #1

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

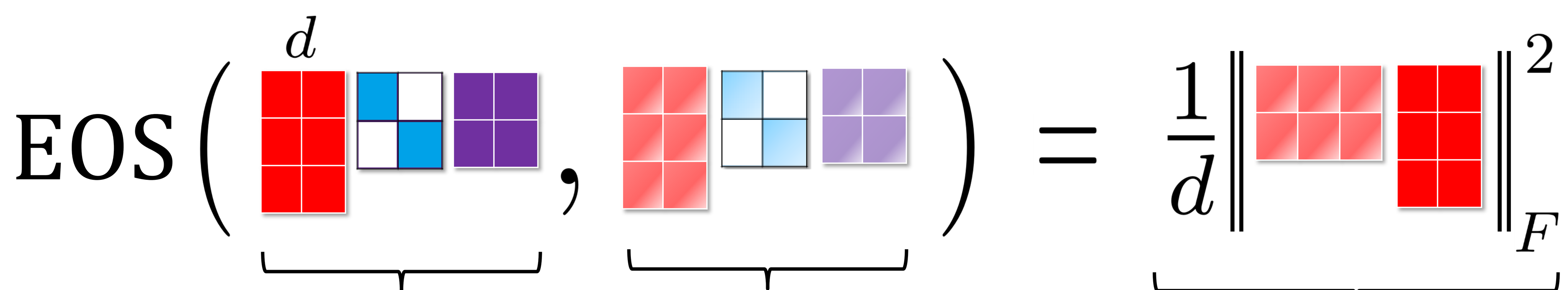
Observation #2

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



A New Quality Measure: The Eigenspace Overlap Score (EOS)

Definition



Uncompressed embedding SVD $X = USV^T$

Compressed embedding SVD $\tilde{X} = \tilde{U}\tilde{S}\tilde{V}^T$

Eigenspace overlap score

Intuition:

- Span of **left singular vectors** determines linear regression predictions.
- **EOS measures similarity** between the compressed & uncompressed embeddings' left singular vectors.

Theoretical Results

Theorem 1 (informal): Generalization & EOS

For fixed design linear regression, if $\bar{y} \in \mathbb{R}^n$ is a random label vector in $\mathrm{span}(U)$, then

$$\underbrace{\mathbb{E}_{\bar{y}}\left[\mathcal{R}_{\bar{y}}(\tilde{X}) - \mathcal{R}_{\bar{y}}(X)\right]}_{\text{Test MSE relative to uncompressed embedding}} = \mathcal{O}\Big(1 - \mathrm{EOS}(X, \tilde{X})\Big).$$

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

Theorem 2 (informal): Uniform Quantization Bound

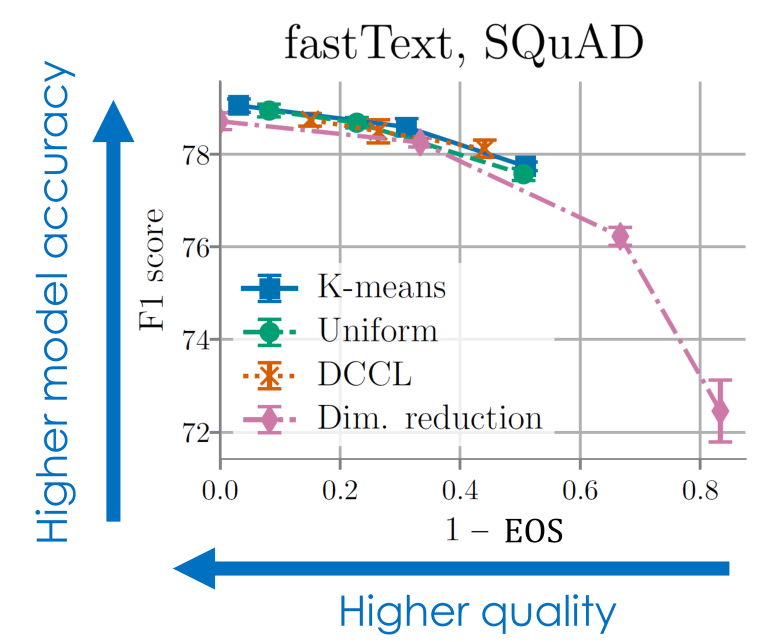
Let $\tilde{X}$ be a b-bit uniform quant. of X. To achieve $\mathrm{EOS} \geq 1 - \epsilon$, $\tilde{X}$ requires a logarithmic # of bits

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

Uniform quantization can attain high EOS with low precision.

Experiments

Correlation of EOS with Downstream Performance

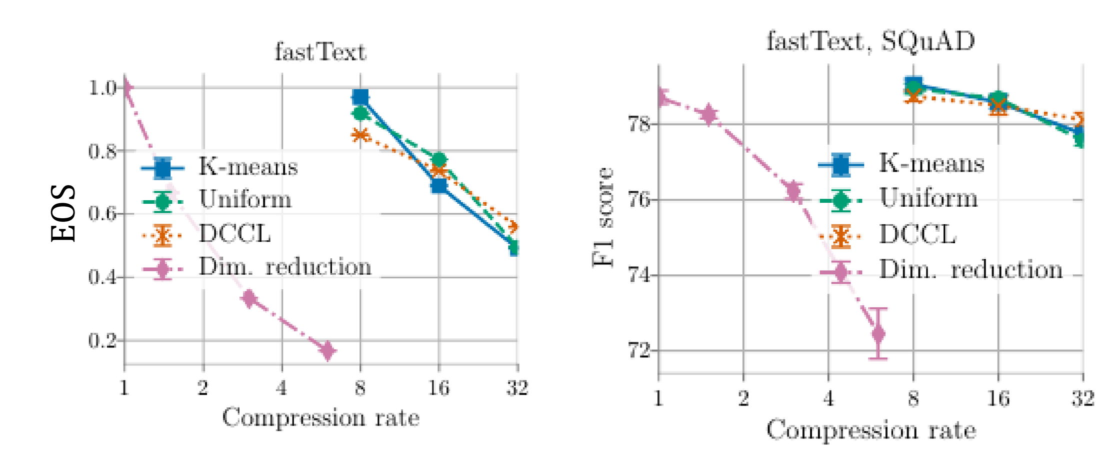


Spearman rank correlation

Dataset	SQuAD	MNLI
Embedding	fastText	BERT WordPiece
PIP loss [1]	0.34	0.45
Δ [2]	0.31	0.44
$\Delta_{\max}$ [3]	0.72	0.86
EOS (Ours)	**0.91**	**0.92**

EOS correlates strongly with downstream performance.

Uniform Quantization Performance

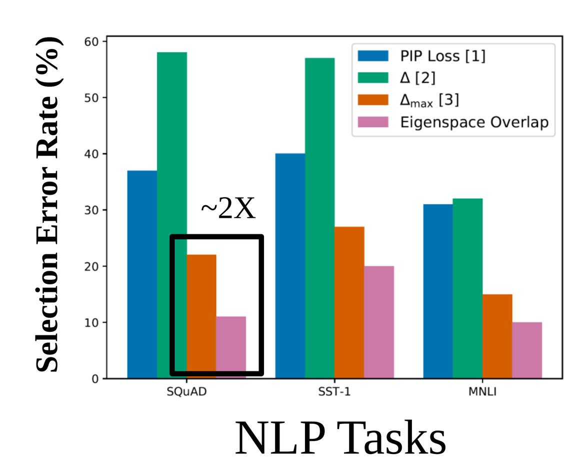


Uniform quantization matches the more complex methods.

EOS for Compressed Embedding Selection

Idea: Use EOS to efficiently select between compressed embeddings.

EOS attains up to 2x lower selection error rate than next best measure.



Resources and References

Resources

arXiv: https://arxiv.org/abs/1909.01264

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

arXiv GitHub

References

[1] Yin and Shen. On the dimensionality of word embedding. NeurIPS, 2018.
[2] Avron et al. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. ICML, 2017.
[3] Zhang et al. Low-precision random Fourier features for memory-constrained kernel approximation. AISTATS, 2019.
[4] Shu and Nakayama. Compressing word embeddings via deep compositional code learning. ICLR, 2018
[5] Andrews. Compressing word embeddings. ICONIP, 2016.