On the Downstream Performance of Compressed Word Embeddings

NeurIPS 2019 Spotlight!

Avner May, Jian Zhang, Tri Dao, Christopher Ré
{avnermay, zjian, trid, chrismre}@cs.stanford.edu

arXiv: https://arxiv.org/abs/1909.01264
GitHub: https://github.com/HazyResearch/smallfry

Overview

Word embeddings:

Important for strong NLP performance

Take a lot of memory

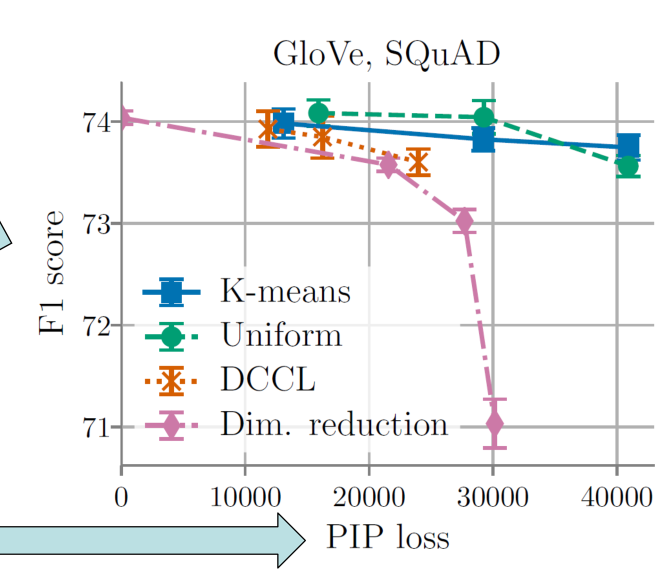
Common solution: Compression.

Motivating question:
When does a compressed embedding perform well on downstream tasks?

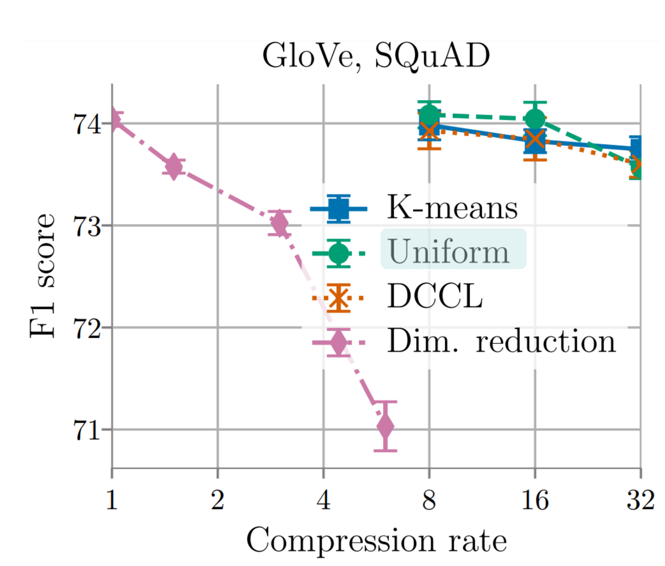
Contribution:
We propose a new theoretically grounded metric to explain the downstream performance of compressed embeddings.

Motivating Observations

1. Downstream performance does not correlate well with existing metrics.



2. Simple compression methods (e.g., uniform quantization) can match more complex ones (e.g., DCCL, k-means).



Our Metric: The Eigenspace Overlap Score (EOS)

Definition: Embedding matrix $X = USV^T \in \mathbb{R}^{n\times d}$, compressed embedding matrix $\tilde{X} = \tilde{U}\tilde{S}\tilde{V}^T \in \mathbb{R}^{n\times k}$. We define the *eigenspace overlap score (EOS)* between X and $\tilde{X}$ as

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

Intuition: $\mathrm{span}(U)$ determines linear regression performance.
This metric measures similarity between $\mathrm{span}(U)$ and $\mathrm{span}(\tilde{U})$.

Theoretical Results

Theorem 1 (informal): Generalization Bound

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

$$\mathbb{E}_{\bar{y}}\left[\mathcal{R}_{\bar{y}}(\tilde{X}) - \mathcal{R}_{\bar{y}}(X)\right] = \frac{d}{n}\cdot\left(1-\mathcal{E}(X,\tilde{X})\right) - \frac{d-k}{n}\sigma^2.$$

Generalization error

Label noise

The generalization performance of the compressed embedding is determined by the EOS.

Theorem 2 (informal): Bound for Uniform Quant.

Let $X \in \mathbb{R}^{n\times d}$ be a bounded embedding matrix $\left(X_{ij} \in \left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]\right)$, and let $\tilde{X}$ be a b-bit uniform quantization of X. Then

$$\mathbb{E}\left[1-\mathcal{E}(X,\tilde{X})\right] \le O\left(2^{-2b}\right).$$

Uniform quantization can attain high EOS at low precision.

Experiments

Correlation of EOS with Downstream Performance

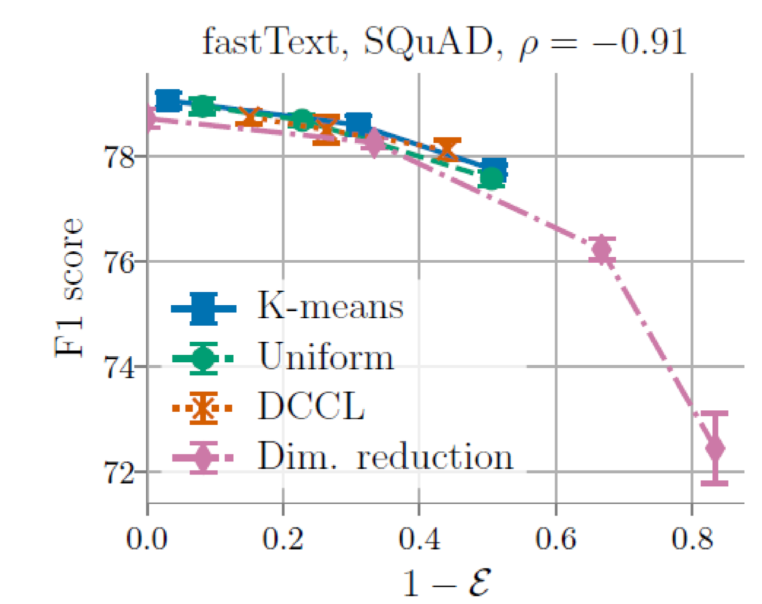
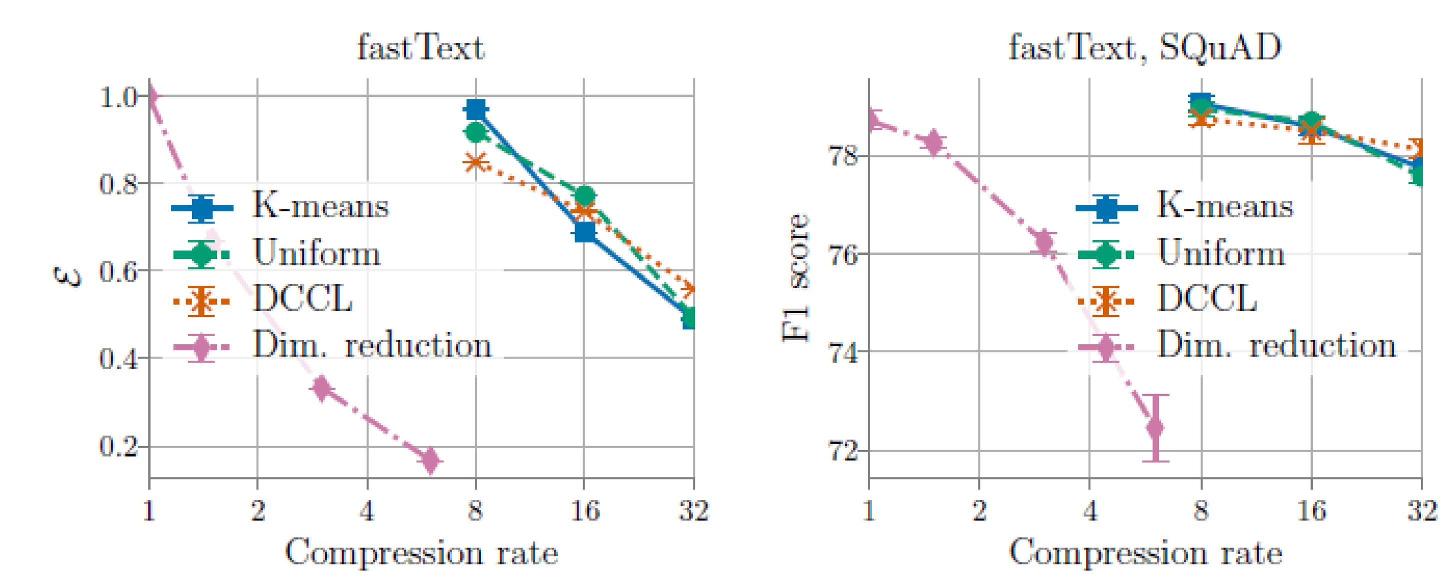


Table 1: **Selection error rates.**

Dataset	SQuAD		SST-1	
Embedding	GloVe	fastText	GloVe	fastText
PIP loss	0.32	0.37	0.32	0.40
Δ	0.34	0.58	0.39	0.57
$\Delta_{\max}$	0.28	0.22	0.30	0.27
$1-\mathcal{E}$	**0.17**	**0.11**	**0.19**	**0.20**

EOS correlates well with downstream perf.
→ can use metric as a *selection criterion* for choosing between compressed embeddings!

Uniform Quantization Performance



Uniform quantization matches or outperforms more complex methods (in EOS and downstream performance).