On the Downstream Performance of Compressed Word Embeddings

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

- Important for strong NLP performance
- Take a lot of memory
Word Embedding Compression
What determines whether a compressed embedding matrix will perform well on downstream tasks?
Motivating Observation

**Existing ways** of measuring compression quality often **fail to explain** relative downstream performance.

Better compression quality measure  \(\uparrow\)  Worse downstream performance  \(\downarrow\)
Our Contributions: Outline

1. Define a **new measure** of compression quality.
2. Prove **generalization bounds** using this measure.
3. Show strong **empirical correlation** w. downstream performance.
4. Use measure to **select** compressed embeddings.

Up to **2x lower** selection error rates than the next best measure.
Observation:
Predictions are a projection onto the span of left singular vectors.
Defining the Measure: Eigenspace Overlap Score (EOS)

**Intuition:**
Measures similarity between the span of *left singular vectors*.

\[
E(d, \mathcal{F}) = \frac{1}{d} \| F \|^2
\]
Theoretical Results: Linear Regression

**Theorem (informal):**
Expected difference in *test* mean-squared error attained by *compressed* vs. *uncompressed* embeddings is *determined by EOS*.

Better EOS → Better downstream performance
**Empirical Correlation:** Beyond Linear Regression

EOS attains **strong correlation** with downstream **model accuracy**.

1 - EOS

Lower model acc.

Lower quality

Empirical Correlation: Beyond Linear Regression

EOS attains *up to 2x lower selection* error rates than 2nd best.

![Graph showing selection error rates for NLP tasks with Bars for PIP Loss [1], Δ [2], Δ_max [3], and Eigenspace Overlap.](image)

Our Contributions: Summary

1. Defined a **new measure** of compression quality.

2. Proved **generalization bounds** using this measure.

3. Showed strong **empirical correlation** w. downstream perf.

4. Used measure to **select** compressed embeddings.
THANK YOU!

Poster #185, 5-7 pm Dec. 12!

Code: https://github.com/HazyResearch/smallfry
E-mail: avnermay@cs.stanford.edu