A Compressed Sensing View of Unsupervised Text Embeddings, Bag-of-n-Grams, and LSTMs

Sanjeev Arora

Mikhail Khodak

Nikunj Saunshi

Kiran Vodrahalli

Overview

- The success of modern NLP is based around ***distributed representations*** - low-dimensional semantic text embeddings that are used and produced by neural networks.

- Our goal is to reason formally about distributed representations:
  - What information do they encode?
  - How will they perform on downstream tasks?

- We prove that LSTMs can compute compressed representations of older (but very effective) sparse feature representations (e.g. Bag-of-Words) that are approximately as powerful for linear document classification.

- We also observe empirically that word embeddings provide a surprisingly effective design matrix for sparse recovery of Bag-of-Words.

Setting

- Assume a distribution **D** of documents, each a sequence of at-most **T** words $\mathbf{w_1, \ldots, w_T}$ drawn from a vocabulary of size **V**.

- We are interested in real-vector document representations over which we can learn a high-accuracy binary linear classifier.

Classical Approach: Bag-of-n-Grams

- Bag-of-Words: represent each document by a vector counting the number of times each word appears.

- Bag-of-n-Grams: represent each document by a vector counting the number of times each unigram, bigram, …, n-gram appears.
  - Surprisingly effective (Wang & Manning 2012).

Distributed Approach: Hidden State of an RNN

- Assign to each word **w** a real vector $\mathbf{v_w}$ and use them as inputs to an LSTM that computes a hidden state vector $\mathbf{h_t}$ at each word in document $\mathbf{w_1, \ldots, w_T}$

$$h_t = f(v_{w_t}, h_{t-1}) \circ h_{t-1} + i(v_{w_t}, h_{t-1}) \circ g(v_{w_t}, h_{t-1})$$

- Represent the document as the last state $\mathbf{h_T}$.

- Use supervised or unsupervised training to learn the LSTM parameters.

Linear Scheme

- Between sparse and neural representations are linear embedding schemes: taking the sum over the word embeddings in a document.

- Empirically shown to be effective on some tasks (Wieting et al. 2016, Arora et al. 2017)

- Can be viewed as a linear compression **Ax** of the BoW vector **x**.

Related Work on BonG Compression

- Past work has shown how to construct compressed representations from which the original BonG vectors can be recovered:

  - Plate (1995): represent objects (words) using low-dimensional random vectors, compose objects (n-grams) using circular convolution, and represent collections of items (documents) using summation.

  - Paskov et al. (2013): use a LZ77-inspired approach to reduce the number of features; good classification performance but still quite high-dimensional.

- Our work is the first to analyze performance on downstream tasks.

Main Theorem

Theorem [**AKSV'18**]**:** Let w_0 be the optimal linear classifier for BonGs for some convex Lipschitz loss ℓ. Then we can initialize a $\mathcal{O}(nd)$-memory LSTM and learn a linear classifier $\hat{w}$ so that with probability $1 - \delta$

$$\ell(\hat{w}) \leq \ell(w_0) + \mathcal{O}\left(\|w_0\|_2 \sqrt{\varepsilon + \frac{1}{m} \log \frac{1}{\delta}}\right)$$

for $d = \tilde{\Omega}\left(\frac{T}{\varepsilon^2} \log \frac{nV}{\delta}\right)$. Here T is the maximum document length, V is the vocabulary size, and m is the number of samples.

Proof Outline

- Design an RIP matrix **A** such that a low-memory LSTM can compute a document representation **Ax**, where **x** is a BonG vector.
- Show that learning is possible under compression: a linear classifier learned over $\{\mathbf{Ax}_i\}$ is almost as good as a linear classifier learned over $\{\mathbf{x}_i\}$ if the vectors $\mathbf{x}_i$ are sparse and **A** satisfies an RIP condition.

Restricted Isometry Property (RIP):
A is (k, ε)-RIP if $(1-\varepsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1+\varepsilon)\|x\|_2$ for all k-sparse x.

Assumptions

- n-grams are order-invariant ((a,b) ~ (b,a))
  - reasonable - performance is about the same
- no word occurs in any n-gram more than once (no (a,a), (a,b,a))
  - violated in real documents, but can be removed by a preprocessing step

Proof Outline

- Design an RIP matrix **A** such that a low-memory LSTM can compute a document representation **Ax**, where **x** is a BonG vector.
- Show that learning is possible under compression: a linear classifier learned over $\{\mathbf{A}\mathbf{x}_i\}$ is almost as good as a linear classifier learned over $\{\mathbf{x}_i\}$ if the vectors $\mathbf{x}_i$ are sparse and **A** satisfies an RIP condition.

Document Representation

Represent each word w by an i.i.d. random Rademacher vector $v_w \in \{\pm 1/\sqrt{d}\}^d$ and each n-gram $g = (w_1, \ldots, w_n)$ as $v_g = d^{\frac{n-1}{2}} (v_{w_1} \odot \cdots \odot v_{w_n})$. Then there exists an LSTM that can compute the vector

$$h_T = \begin{pmatrix} \sum_{\text{word } w \text{ in document}} v_w \\ \vdots \\ \sum_{\text{n-gram } g \text{ in document}} v_g \end{pmatrix}$$

Then the matrix A whose columns are the n-gram vectors v_g satisfies $h_T = Ax$ for any document, where x is the document's BonG vector.

$$h_t = f(v_{w_t}, h_{t-1}) \circ h_{t-1} + i(v_{w_t}, h_{t-1}) \circ g(v_{w_t}, h_{t-1})$$

A is RIP

For random variables $x^{(1)}, \ldots, x^{(d)} \sim \mathcal{U}\{\pm 1\}^V$, $j = 1, \ldots, d$ we can write

$$\sqrt{d}A = \begin{pmatrix} \phi_1(x^{(1)}) & \cdots \\ \vdots & \\ \phi_1(x^{(d)}) & \cdots \end{pmatrix}$$

where each ϕ_i corresponds to a unique n-gram. This system has the following two properties:

1. Each ϕ_i is a mononomial in n variables and thus has norm bounded by 1.
2. Each monomial is unique $\implies \mathbb{E}\langle \phi_i(x^{(k)}), \phi_j(x^{(k)}) \rangle = 0 \; \forall \; i \neq j$.

Properties 1 and 2 imply that $\sqrt{d}A$ corresponds to a bounded orthonormal system (BOS) and so A is (k, ε)-RIP for $d = \tilde{\Omega}\left(\frac{k}{\varepsilon^2} \log \frac{V}{\delta}\right)$ with probability $1 - \delta$ (Foucart & Rauhut 2013).

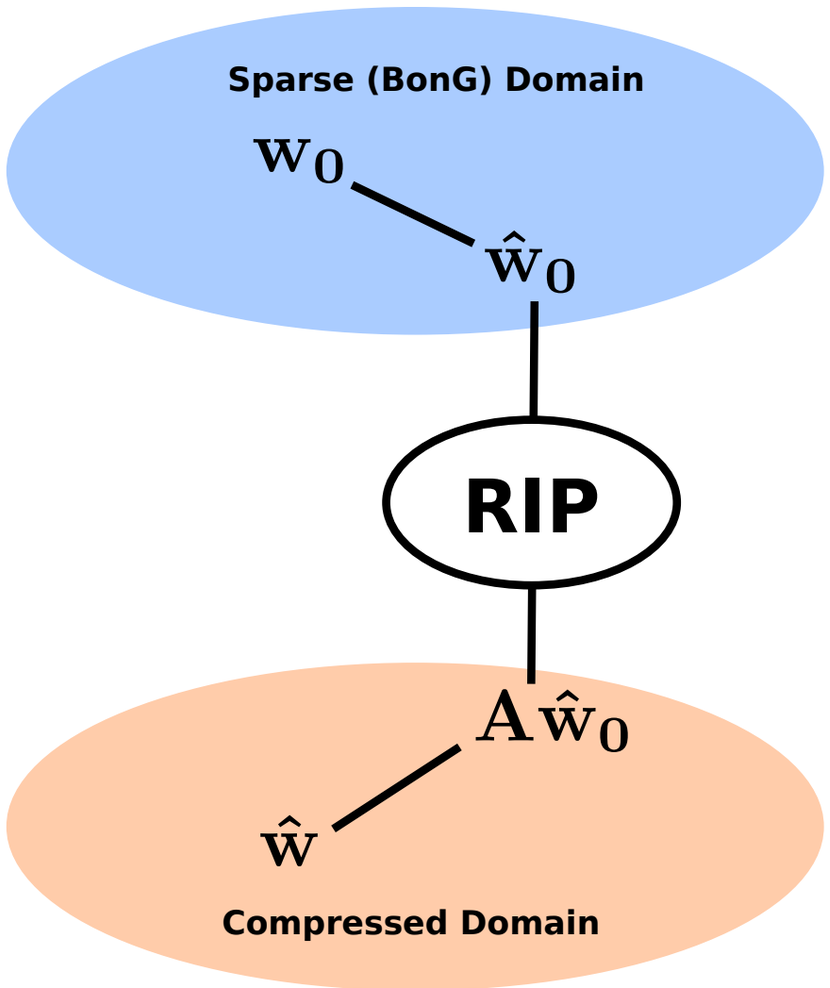
Proof Outline

- Design an RIP matrix **A** such that a low-memory LSTM can compute a document representation **Ax**, where **x** is a BonG vector.
- Show that learning is possible under compression: a linear classifier learned over $\{\mathbf{A}\mathbf{x}_i\}$ is almost as good as a linear classifier learned over $\{\mathbf{x}_i\}$ if the vectors $\mathbf{x}_i$ are sparse and **A** satisfies an RIP condition.

Compressed Learning (Calderbank et al. 2009)

We examine four different classifiers:

1. the optimal sparse classifier $\mathbf{w_0}$
2. the sparse classifier $\mathbf{\hat{w}_0}$ minimizing the (regularized) loss over $\{(x_i, y_i)\}_{i=1}^m$
3. the dense classifier $\mathbf{A\hat{w}_0}$
4. the classifier $\mathbf{\hat{w}}$ minimizing the (regularized) loss over $\{(Ax_i, y_i)\}_{i=1}^m$



Bounding $\ell(\hat{w}_0)$ in terms of $\ell(w_0)$ and $\ell(\hat{w})$ in terms of $\ell(A\hat{w}_0)$ can be done using standard techniques. **We need the RIP condition on A to bound $\ell(A\hat{w}_0)$ in terms of $\ell(\hat{w}_0)$.**

Proof Sketch

Consider the following two facts:

1. The minimizer $\hat{w}_0$ of $\frac{1}{m}\sum_{i=1}^{m}\ell(w^T x_i, y_i) + \frac{1}{2C}\|w\|_2^2$ can be written as the linear combination $\hat{w}_0 = \sum_{i=1}^{m}\alpha_i y_i x_i$, with bounded coefficients $|\alpha_i| \leq \frac{\lambda C}{m}$.

2. A is $(2k,\varepsilon)$-RIP $\implies (1+\varepsilon)x^T x' - R^2\varepsilon \leq (Ax)^T(Ax') \leq (1-\varepsilon)x^T x' + R^2\varepsilon$ for k-sparse x, x' with $\|x\|_2, \|x'\|_2 \leq R$.

So for any k-sparse x with $\|x\|_2 \leq R$:

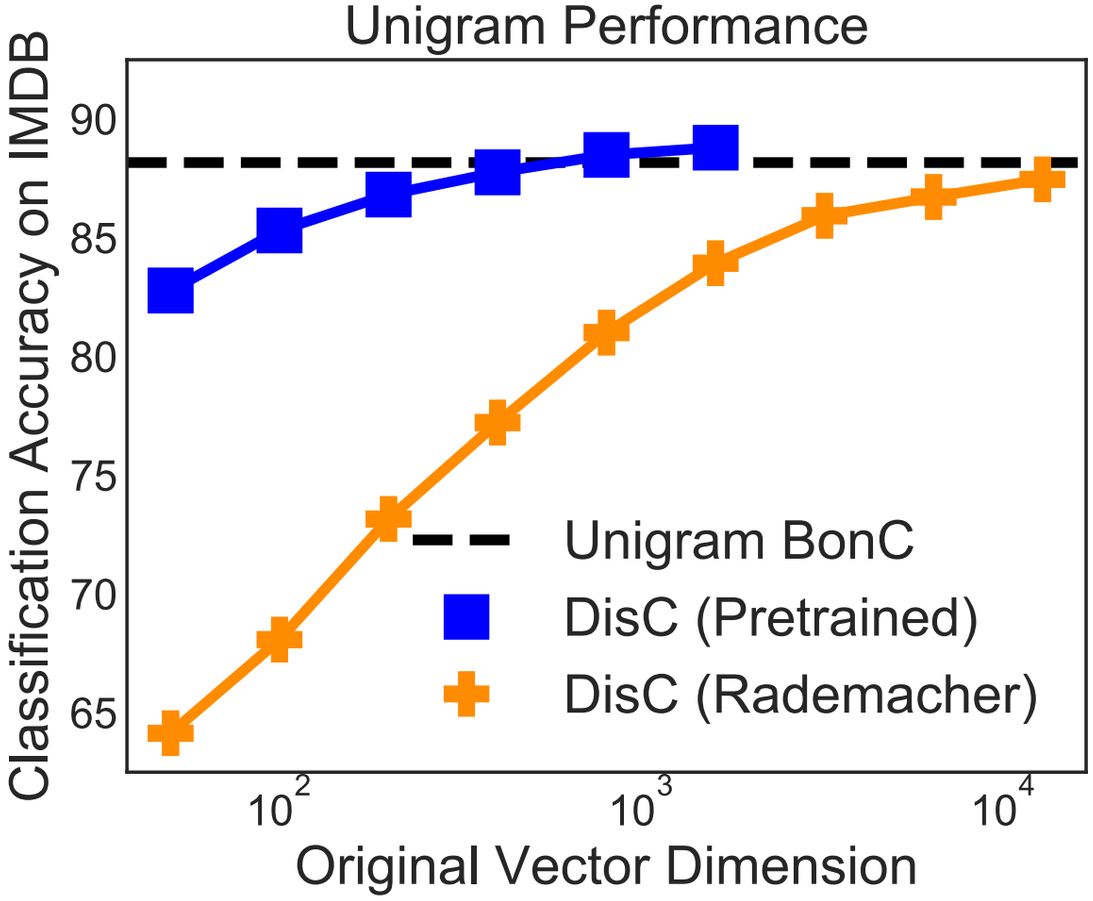
$$
\begin{aligned}
(A\hat{w}_0)^T(Ax) &= \sum_{i=1}^{m}\alpha_i y_i (Ax_i)^T(Ax) \\
&\leq \sum_{i:\alpha_i y_i \geq 0}\alpha_i y_i\left((1-\varepsilon)x_i^T x + 2R^2\varepsilon\right) + \sum_{i:\alpha_i y_i < 0}\alpha_i y_i\left((1+\varepsilon)x_i^T x - 2R^2\varepsilon\right) \\
&= \hat{w}_0^T x - \varepsilon\sum_{i=1}^{m}|\alpha_i y_i| x_i^T x + 2R^2\varepsilon\sum_{i=1}^{m}|\alpha_i y_i| \leq \hat{w}_0^T x + 3\lambda C R^2\varepsilon
\end{aligned}
$$

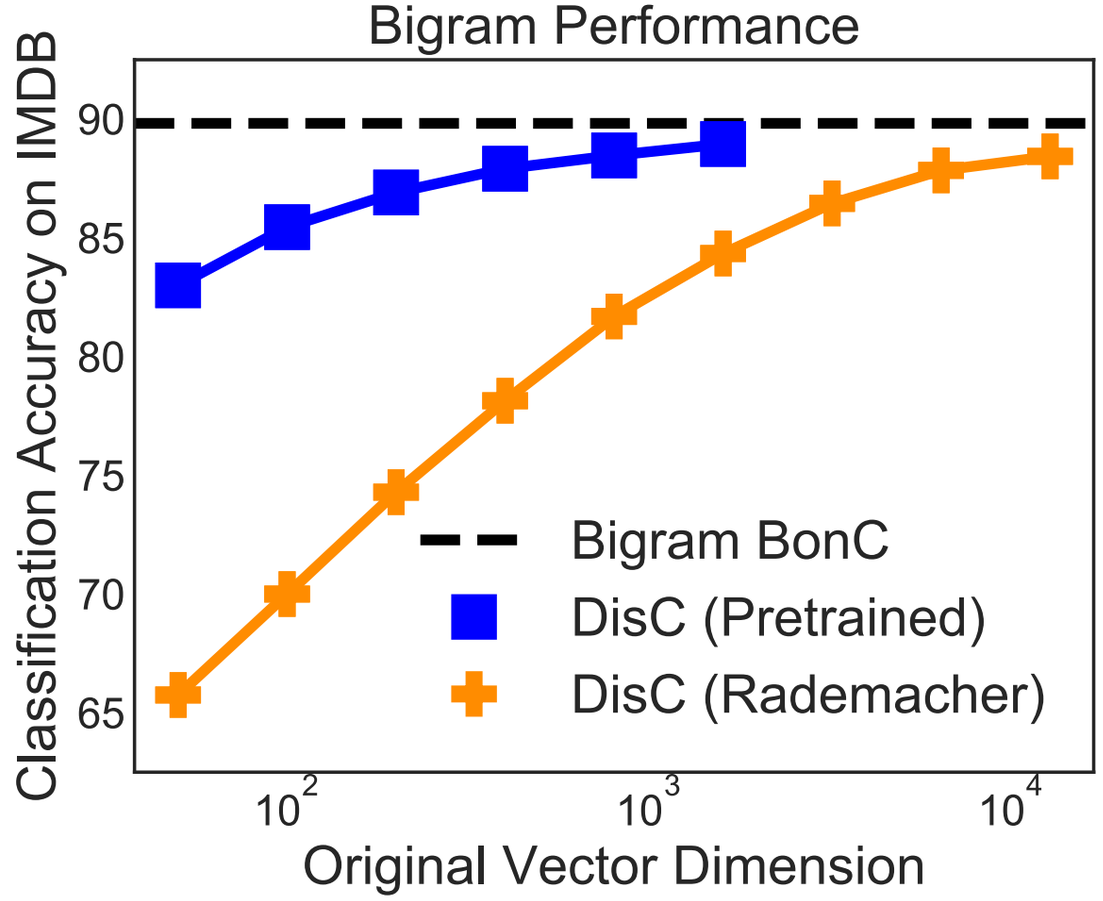
Similarly $(A\hat{w}_0)^T(Ax) \geq \hat{w}_0^T x - 3\lambda C R^2\varepsilon$.

Taking expectations over $x \sim \mathcal{D}$ yields $\ell(A\hat{w}_0) \leq \ell(\hat{w}_0) + 3\lambda^2 C R^2\varepsilon$.

Classification Performance

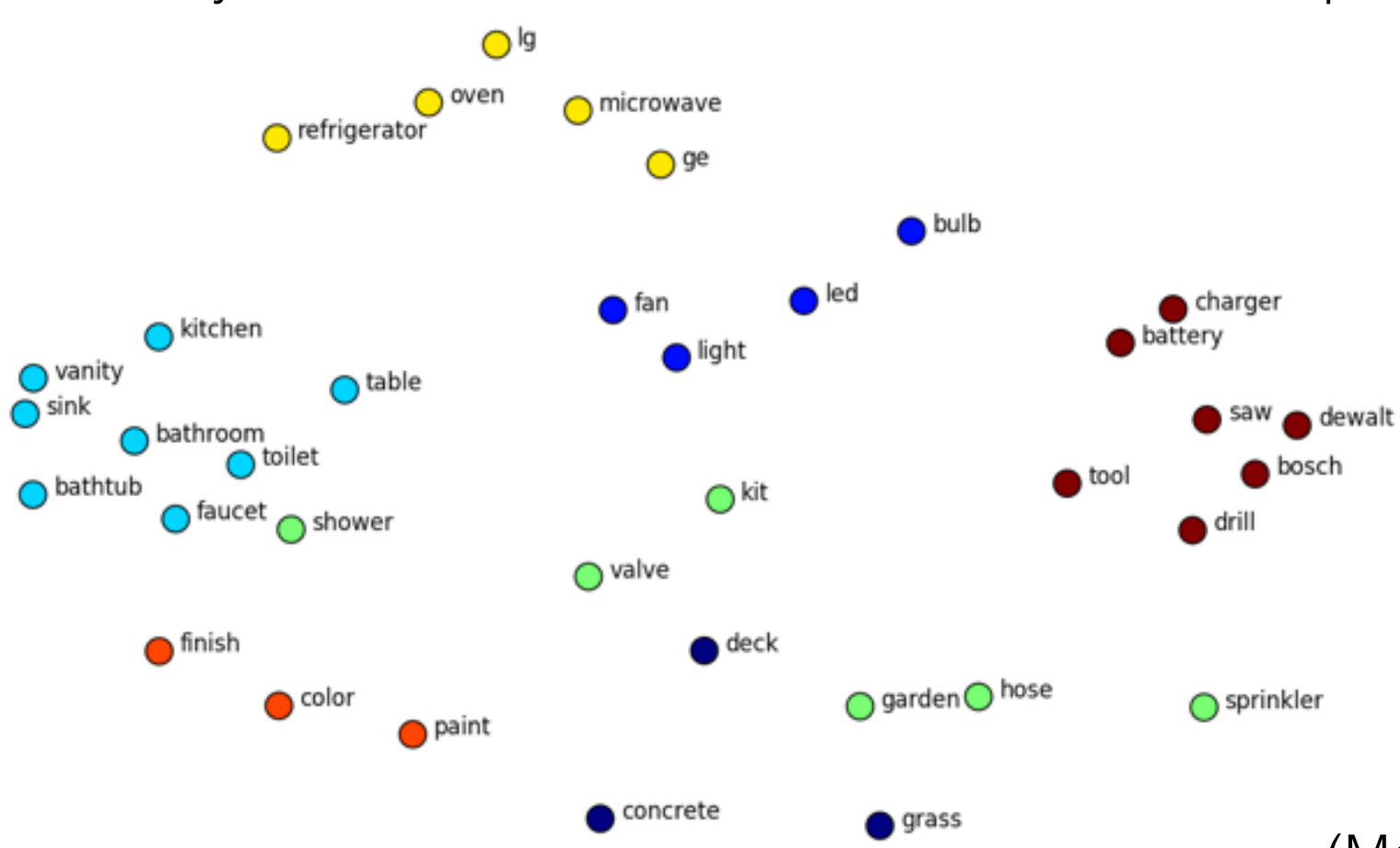
$$\ell(\hat{w}) \leq \ell(w_0) + \mathcal{O}\left(\|w_0\|_2 \sqrt{\varepsilon + \frac{1}{m} \log \frac{1}{\delta}} \right)$$





Word Embeddings

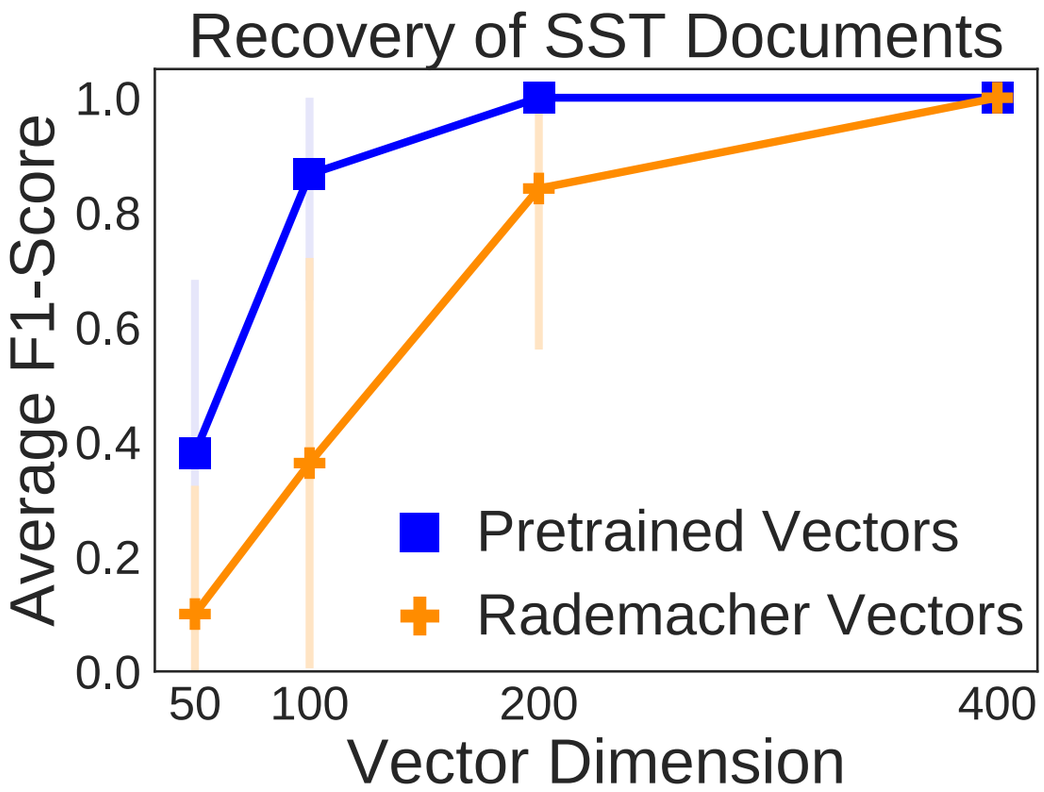
- For the main result we assumed words were represented by Rademacher random vectors.
- In practice pretrained embeddings capturing the ‘meaning’ of words are used instead.
- These vectors are trained so that similar words are closer together and thus *cannot* satisfy RIP. How can we understand their better performance?

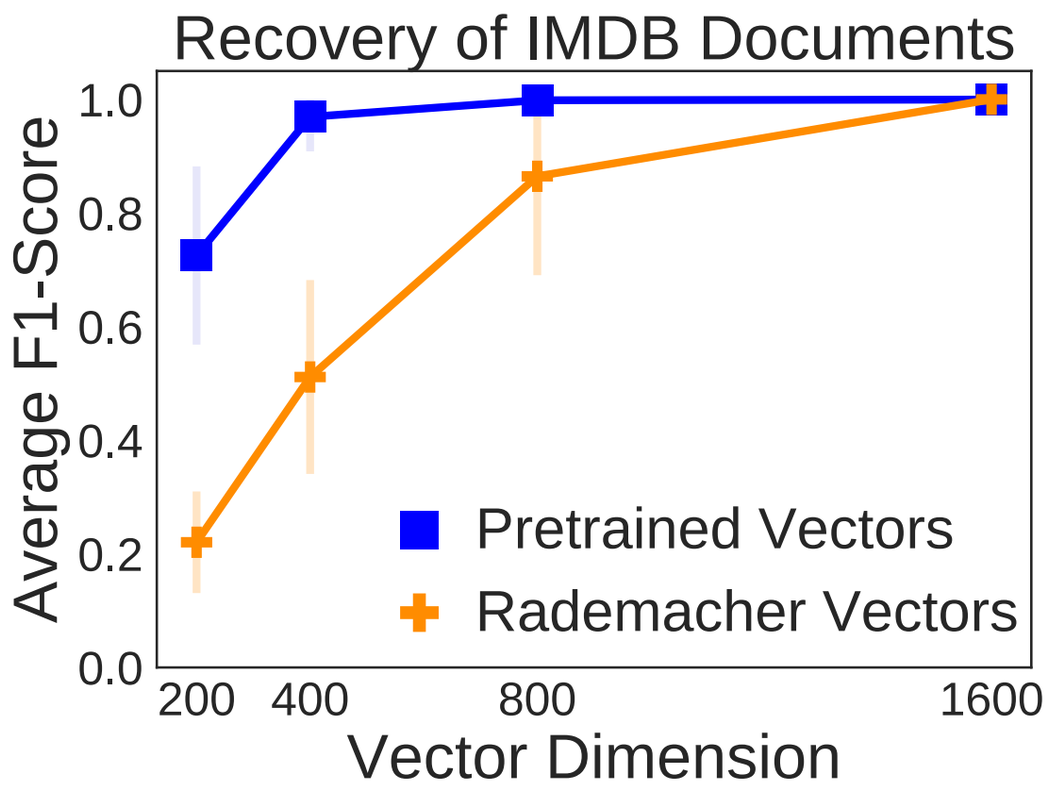


(Merentitis et al. 2016)

A Sparse Recovery Experiment

- What do word embedding-based document representations encode?
  - Compress a BoW vector **x**: **b = Ax**
  - Recover **x** using Basis Pursuit (BP): **min $|x|_1$ s.t. Ax = b**
  - Note: RIP provides exact recovery guarantees for BP.

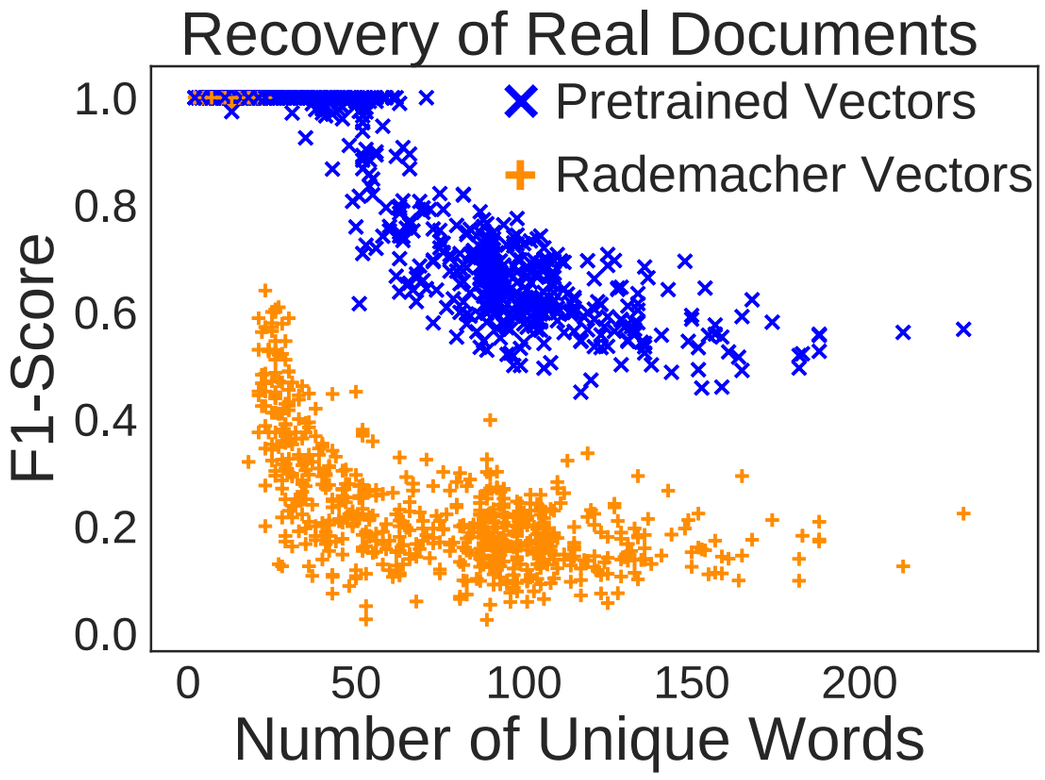


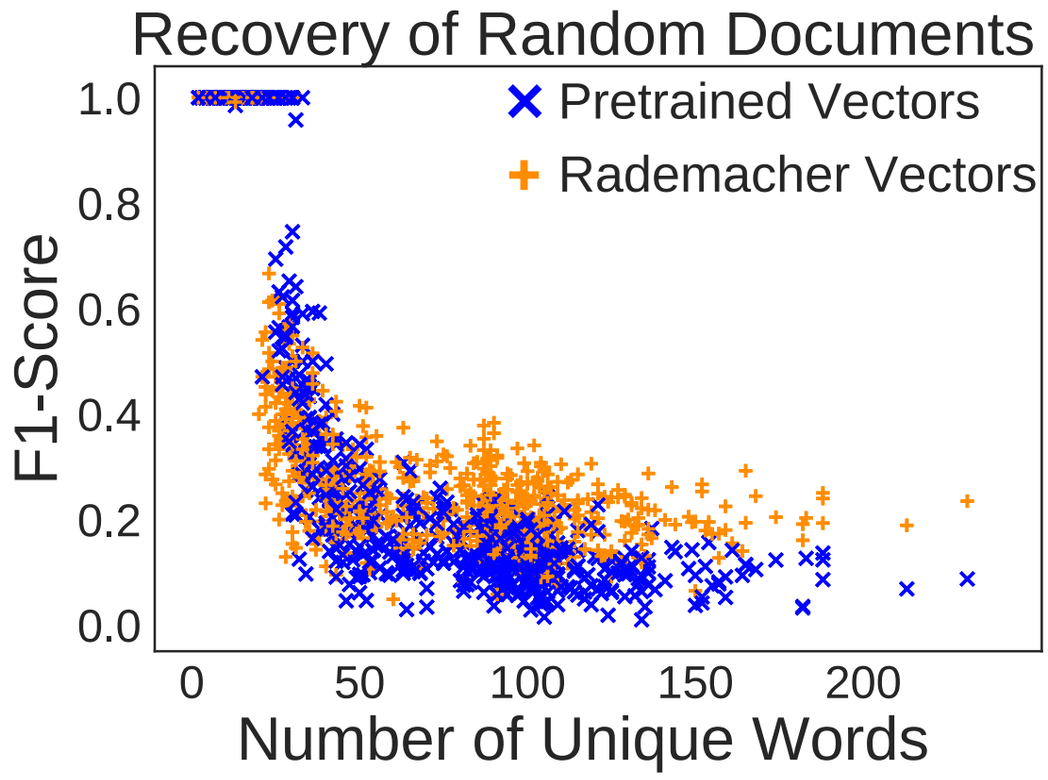


Why Are Embeddings Good for Compressed Sensing?

- RIP is a very strong condition - sufficient but not necessary

- Word embeddings only perform well when the compressed signal is a BoW vector; for random sparse vectors they perform poorly:

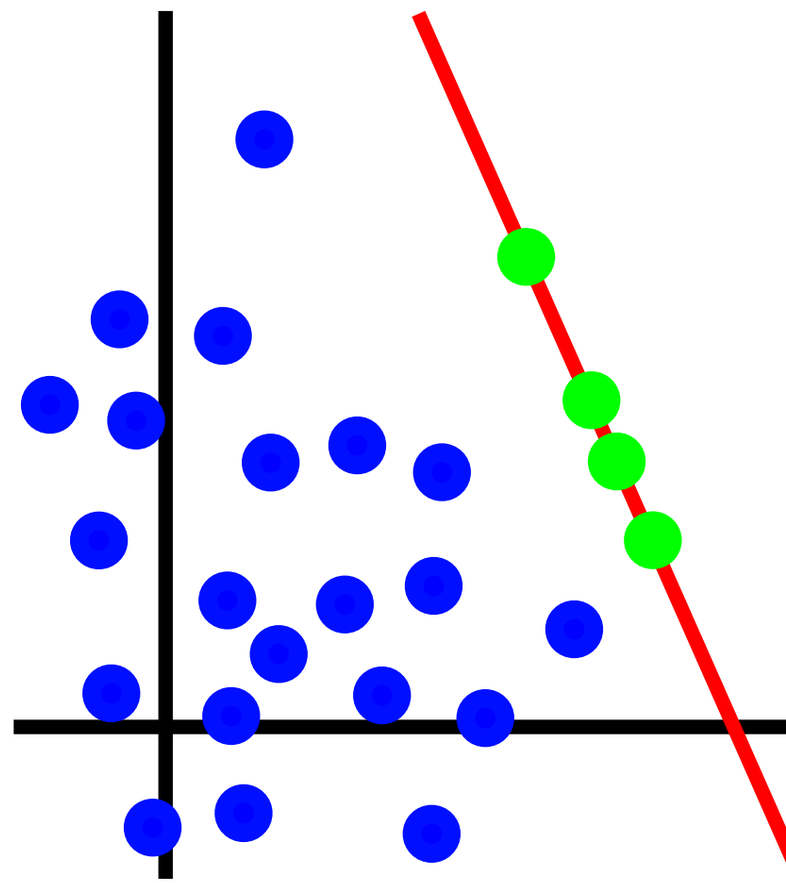




A Geometric Explanation?

- Using a theorem due to Donoho & Tanner (2005) we show that perfect recovery of a sparse signal **x** with support **S** from **Ax** is equivalent to the existence of a hyperplane going through the columns of $\mathbf{A}_S$ such that all other columns of **A** are on the same side as the origin.

- Since words occurring in the same document tend to have similar vectors, perhaps they are more likely to have a hyperplane separating them out.



Future Work

- Train an RNN initialized or regularized by the constructed linear scheme.

- Incorporate better n-gram embeddings.

- Open Problems:
  - Is compressed learning possible under weaker conditions on **A**?
  - Provide a generative or information-theoretic explanation of recovery for word embeddings.