Vector Semantics and Embeddings
Tasks

- Vectors which represent words or sequences
- Dimensionality Reduction
- Recurrent Neural Network and Sequence Models
Objective

To embed: convert a token (or sequence) to a vector that represents meaning.
Objective

To embed: convert a token (or sequence) to a vector that is useful to perform NLP tasks.
Objective

port $\xrightarrow{\text{embed}}$ 

$[\text{Diagram}]$
Objective

\( port \xrightarrow{\text{embed}} \begin{pmatrix} 0 \\ \cdots \\ 0 \\ 0 \\ 1 \\ \cdots \\ 0 \end{pmatrix} \)
Objective

one-hot is sparse vector

For deep learning, in practice, they work better. Why? Roughly, less parameters becomes increasingly important when you are learning multiple layers of weights rather than just a single layer.

Prefer dense vectors

- Less parameters (weights) for machine learning model.
- May generalize better implicitly.
- May capture synonyms
Objective

Prefer dense vectors
- Less parameters (weights) for machine learning models.
- May generalize better implicitly.
- May capture synonyms.

For deep learning, in practice, they work better. Roughly, less parameters becomes increasingly important when you are learning multiple layers of weights rather than just a single layer.

one-hot is sparse vector

(Jurafsky, 2012)
Objective

Prefer dense vectors
- Less parameters (weights) for machine learning model.
- May generalize better implicitly.
- May capture synonyms

For deep learning, in practice, they work better. Roughly, less parameters becomes increasingly important when you are learning multiple layers of weights rather than just a single layer.

(Jurafsky, 2012)
Objective

To embed: convert a token (or sequence) to a vector that represents meaning.
Objective

To embed: convert a token (or sequence) to a vector that represents meaning.

Wittgenstein, 1945: “The meaning of a word is its use in the language”

Distributional hypothesis -- A word’s meaning is defined by all the different contexts it appears in (i.e. how it is “distributed” in natural language).

Firth, 1957: “You shall know a word by the company it keeps”
Objective

To embed: convert a token (or sequence) to a vector that represents meaning.

Wittgenstein, 1945: “The meaning of a word is its use in the language”

Distributional hypothesis -- A word’s meaning is defined by all the different contexts it appears in (i.e. how it is “distributed” in natural language).

Firth, 1957: “You shall know a word by the company it keeps”

The nail hit the beam behind the wall.
Distributional Hypothesis

The nail hit the beam behind the wall.
Objective

$port \xrightarrow{\text{embed}} \begin{pmatrix} 0.53 \\ 1.5 \\ 3.21 \\ -2.3 \\ 0.76 \end{pmatrix}$
Objective

port

embed

0.53
1.5
3.21
-2.3
.76

**port.n.1** (a place (seaport or airport) where people and merchandise can enter or leave a country)

**port.n.2** port wine (sweet dark-red dessert wine originally from Portugal)

**port.n.3**rn embrasure, porthole (an opening (in a wall or ship or armored vehicle) for firing through)

larboard, **port.n.4** (the left side of a ship or aircraft to someone who is aboard and facing the bow or nose)

interface, **port.n.5** ((computer science) computer circuit consisting of the hardware and associated circuitry that links one device with another (especially a computer and a hard disk drive or other peripherals))
Objective

**port**

- **port.n.1** (a place (seaport or airport) where people and merchandise can enter or leave a country)
- **port.n.2** port wine (sweet dark-red dessert wine originally from Portugal)
- **port.n.3**, embrasure, porthole (an opening (in a wall or ship or armored vehicle) for firing through)
- **larboard**, **port.n.4** (the left side of a ship or aircraft to someone who is aboard and facing the bow or nose)
- **interface**, **port.n.5** ((computer science) computer circuit consisting of the hardware and associated circuitry that links one device with another (especially a computer and a hard disk drive or other peripherals))
How?

1. One-hot representation
2. Selectors (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)

Tf-IDF: Term Frequency, Inverse Document Frequency,

PMI: Point-wise mutual information, ...etc...
How?

1. One-hot representation
2. Selectors (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)

“Neural Embeddings”:

4. Word2vec
5. Fasttext
6. Glove
7. Bert
How?

1. One-hot represent
2. **Selectors** (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)
4. Word2vec
5. Fasttext
6. Glove
7. Bert

..., word1, word2, **bill**, word3, word4, ...

[Diagram showing word embeddings]
How?

1. One-hot represent
2. Selectors (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)

“Neural Embeddings”:

4. Word2vec
5. Fasttext
6. Glove
7. Bert
How?

1. One-hot represent
2. Selectors (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)

“Neural Embeddings”:

4. Word2vec
5. Fasttext
6. Glove
7. Bert
How?

1. One-hot represent
2. Selectors (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)

“Neural Embeddings”:

4. Word2vec
5. Fasttext
6. Glove
7. Bert
SVD-Based Embeddings

Singular Value Decomposition...
Concept, In Matrix Form:

- **Columns:** $p$ features
- **Rows:** $N$ observations

Matrix:

- $f_1, f_2, f_3, f_4, \ldots, f_p$
- $o_1, o_2, o_3, \ldots, o_N$
SVD-Based Embeddings

context words are features
f1, f2, f3, f4, ...

co-occurrence counts are cells.

target words are observations
SVD-Based Embeddings

context words are features
f1, f2, f3, f4, ...

co-occurrence counts are cells.

o1
o2
o3
...

oN

Dimensionality reduction -- try to represent with only p' dimensions

c1, c2, c3, c4, ...

cp'

o1
o2
o3
...

oN

target words are observations
Data (or, at least, what we want from the data) may be accurately represented with less dimensions.
Dimensionality Reduction

Rank: Number of linearly independent columns of A. (i.e. columns that can’t be derived from the other columns through addition).

Q: What is the rank of this matrix?

\[
\begin{pmatrix}
1 & -2 & 3 \\
2 & -3 & 5 \\
1 & 1 & 0
\end{pmatrix}
\]
Dimensionality Reduction

Rank: Number of linearly independent columns of A.
(i.e. columns that can’t be derived from the other columns).

Q: What is the rank of this matrix?

A: 2. The 1st is just the sum of the second two columns...

… we can represent as linear combination of 2 vectors:
Dimensionality Reduction - PCA

Linear approximates of data in \( r \) dimensions.

Found via *Singular Value Decomposition*:

\[
X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T
\]

\( X \): original matrix, \( U \): “left singular vectors”,
\( D \): “singular values” (diagonal), \( V \): “right singular vectors”
Dimensionality Reduction - PCA

Linear approximates of data in \( r \) dimensions.

Found via *Singular Value Decomposition*:

\[
X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T
\]

- \( X \): original matrix,
- \( U \): “left singular vectors”,
- \( D \): “singular values” (diagonal),
- \( V \): “right singular vectors”
Dimensionality Reduction - PCA - Example

\[ X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T \]

Word co-occurrence counts:

\[
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
= \begin{bmatrix}
0.13 & 0.02 & -0.01 \\
0.41 & 0.07 & -0.03 \\
0.55 & 0.09 & -0.04 \\
0.68 & 0.11 & -0.05 \\
0.15 & -0.59 & 0.65 \\
0.07 & -0.73 & -0.67 \\
0.07 & -0.29 & 0.32
\end{bmatrix}
\times \begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{bmatrix}
\times
\begin{bmatrix}
0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\
0.40 & -0.80 & 0.40 & 0.09 & 0.09
\end{bmatrix}

Dimensionality Reduction - PCA - Example

\[ X_{[nxp]} \approx U_{[nxr]} D_{[rxr]} V_{[pxr]}^T \]

Observation: "beam."
- \( \text{count(beam, hit)} = 100 \) -- horizontal dimension
- \( \text{count(beam, nail)} = 80 \) -- vertical dimension
Dimensionality Reduction - PCA

Linear approximates of data in $r$ dimensions.

Found via *Singular Value Decomposition*:

$$X_{[nxp]} \approx U_{[nxr]} D_{[rxr]} V_{[pxr]}^T$$

$X$: original matrix, $U$: “left singular vectors”,
$D$: “singular values” (diagonal), $V$: “right singular vectors”

Projection (dimensionality reduced space) in 3 dimensions:

$$(U_{[nx3]} D_{[3x3]} V_{[px3]}^T)$$

To reduce features in new dataset, $A$:

$$A_{[m \times p]} VD = A_{small[ m \times 3 ]}$$
Dimensionality Reduction - PCA

Linear approximates of data in $r$ dimensions.

Found via *Singular Value Decomposition*:

$$X_{[nxp]} \approx U_{[nrx]} D_{[rxr]} V_{[pxr]}^T$$

$X$: original matrix, $U$: “left singular vectors”, $D$: “singular values” (diagonal), $V$: “right singular vectors”

To check how well the original matrix can be reproduced:

$$Z_{[nxp]} = U D V^T$$, How does $Z$ compare to $X$?

To reduce features in new dataset, $A$:

$$A_{[m \times p]} V D = A_{\text{small}[m \times 3]}$$
Dimensionality Reduction - PCA

Linear approximates of data in $r$ dimensions.

Found via *Singular Value Decomposition*:

$$X_{[n \times p]} \cong U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T$$

$X$: original matrix, $U$: “left singular vectors”,
$D$: “singular values” (diagonal), $V$: “right singular vectors”

To check how well the original matrix can be reproduced:

$$Z_{[n \times p]} = U D V^T$$, How does $Z$ compare to original $X$?

To reduce features in new dataset:

$$A_{[m \times p]} V D = A_{small[ m \times 3 ]}$$
Dimensionality Reduction - PCA

X: original matrix
D: “singular values”
U: “left singular vectors”
V: “right singular vectors”

Goal: Minimize the sum of reconstruction errors:

$$\sum_{i=1}^{N} \sum_{j=1}^{D} \| x_{ij} - z_{ij} \|^2$$

where $x_{ij}$ are the “old” and $z_{ij}$ are the “new” coordinates

To check how well the original matrix can be reproduced:

$$Z_{n \times p} = U D V^T$$

How does $Z$ compare to original $X$?

To reduce features in new dataset:

$$A_{m \times p} V D = A_{\text{small}}$$
Dimensionality Reduction - PCA

**Goal:** Minimize the sum of reconstruction errors:

\[ \sum_{i=1}^{N} \sum_{j=1}^{D} \left\| x_{ij} - z_{ij} \right\|^2 \]

- where \( x_{ij} \) are the “old” and \( z_{ij} \) are the “new” coordinates

To check how well the original matrix can be reproduced:

\( Z_{[n \times p]} = U D V^T \), How does \( Z \) compare to original \( X \)?

To reduce features in new dataset:

\( A_{[m \times p]} VD = A_{\text{small}[m \times 3]} \)
Dimensionality Reduction - PCA

Linear approximates of data in \( r \) dimensions.

Found via *Singular Value Decomposition*:

\[
X_{[n \times p]} \approx U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T
\]

U, D, and V are unique

D: always positive
How?

1. One-hot represent
2. Selectors (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)

“Neural Embeddings”:

4. Word2vec
5. Fasttext
6. Glove
7. Bert
How?

1. One-hot represent
2. Selectors (represent context by “multi-hot” representation)
3. From PCA/Singular Value Decomposition
   (Know as “Latent Semantic Analysis” in some circumstances)

“Neural Embeddings”:

4. Word2vec
5. Fasttext
6. Glove
7. Bert
Word2Vec

Principal: Predict missing word.

Similar to language modeling but predicting context, rather than next word.

\[ p(\text{context} \mid \text{word}) \]
Word2Vec

Principal: Predict missing word.

Similar to language modeling but predicting context, rather than next word.

\[ p(\text{context} \mid \text{word}) \]

To learn, maximize
Word2Vec

Principal: Predict missing word.

Similar to language modeling but predicting context, rather than next word.

$$p(\text{context} \mid \text{word})$$

To learn, maximize.
In practice, minimize

$$J = 1 - p(\text{context} \mid \text{word})$$
Word2Vec: Context

2 Versions of Context:
- Continuous bag of words (CBOW): Predict word from context
- Skip-Grams (SG): predict context words from target
2 Versions of Context:

- Continuous bag of words (CBOW): Predict word from context
- Skip-Grams (SG): predict context words from target

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples
3. Use logistic regression to train a classifier to distinguish those two cases
4. Use the weights as the embeddings

(Jurafsky, 2017)
Word2Vec: Context  
\[ p(\text{context} \mid \text{word}) \]

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the weights as the embeddings.

*The nail hit the beam behind the wall.*

(Jurafsky, 2017)
Word2Vec: Context

\( p(\text{context} \mid \text{word}) \)

\[ x = (\text{hit, beam}), y = 1 \]
\[ x = (\text{the, beam}), y = 1 \]
\[ x = (\text{behind, beam}), y = 1 \]
\[ \ldots \]

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the weights as the embeddings.

*The nail hit the beam behind the wall.*

(Jurafsky, 2017)
Word2Vec: Context

\[ p(\text{context} \mid \text{word}) \]

1. Treat the target word and a neighboring context word as positive examples.

2. Randomly sample other words in the lexicon to get negative samples

3. Use logistic regression to train a classifier to distinguish those two cases

4. Use the weights as the embeddings

*The nail hit the beam behind the wall.*

(Jurafsky, 2017)
Word2Vec: Context

\[ p(\text{context} \mid \text{word}) \]

2 Versions of Context:

- Continuous bag of words (CBOW): Predict word from context
- Skip-Grams (SG): predict context words from target

\( x = (\text{hit, beam}), y = 1 \)
\( x = (\text{the, beam}), y = 1 \)
\( x = (\text{behind, beam}), y = 1 \)

\( \ldots \)
\( x = (\text{happy, beam}), y = 0 \)
\( x = (\text{think, beam}), y = 0 \)

\( \ldots \)

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples
3. Use logistic regression to train a classifier to distinguish those two cases
4. Use the weights as the embeddings

*The nail hit the beam behind the wall.*

(Jurafsky, 2017)
**Word2Vec: Context**

\[ p(\text{context} \mid \text{word}) \]

\( k \) negative example (\( y=0 \)) for every positive.

**How?** Randomly draw from unigram distribution

\[
P(w) = \frac{\text{count}(w)}{\sum_w \text{count}(w)}
\]

1. Treat the target word and a neighboring context word as positive examples.

2. **Randomly sample other words in the lexicon to get negative samples**

3. Use logistic regression to train a classifier to distinguish those two cases

4. Use the weights as the embeddings

---

*The nail hit the beam behind the wall.*

\( x = (\text{hit, beam}), y = 1 \)
\( x = (\text{the, beam}), y = 1 \)
\( x = (\text{behind, beam}), y = 1 \)
...

\( x = (\text{happy, beam}), y = 0 \)
\( x = (\text{think, beam}), y = 0 \)
...

(Jurafsky, 2017)
Word2Vec: Context

\[ p(\text{context} \mid \text{word}) \]

\[ x = (\text{hit, beam}), y = 1 \]
\[ x = (\text{the, beam}), y = 1 \]
\[ x = (\text{behind, beam}), y = 1 \]
\[ \ldots \]
\[ x = (\text{happy, beam}), y = 0 \]
\[ x = (\text{think, beam}), y = 0 \]
\[ \ldots \]

\( k \) negative example (\( y=0 \)) for every positive.

**How?** Randomly draw from unigram distribution adjusted:

\[ P_\alpha(w) = \frac{\text{count}(w)^\alpha}{\sum_w \text{count}(w)^\alpha} \]

\( \alpha = 0.75 \)

1. Treat the target word and a neighboring context word as positive examples.

2. **Randomly sample other words in the lexicon to get negative samples**

3. Use logistic regression to train a classifier to distinguish those two cases

4. Use the weights as the embeddings

\[ \text{The nail hit the beam behind the wall.} \]

(Jurafsky, 2017)
Word2Vec: Context

\[ p(\text{context} \mid \text{word}) \]

x = (hit, beam), y = 1
x = (the, beam), y = 1
x = (behind, beam), y = 1
...
x = (happy, beam), y = 0
x = (think, beam), y = 0
...

$k$ negative example (y=0) for every positive.

**How?** Randomly draw from unigram distribution adjusted:

\[
P_\alpha(w) = \frac{\text{count}(w)^\alpha}{\sum_w \text{count}(w)^\alpha}
\]

\[ \alpha = 0.75 \]

1. Treat the target.
2. Randomly.
3. Use logistic.
4. Use the weight.
Word2Vec: Context

- Continuous bag of words (CBOW): Predict word from context
- Skip-Grams (SG): Predict context words from target

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the weights as the embeddings.

\[ P(y=1| c, t) = \frac{1}{1 + e^{-t \cdot c}} \]

The nail hit the beam behind the wall.

\( x = (\text{hit, beam}), y = 1 \)
\( x = (\text{the, beam}), y = 1 \)
\( x = (\text{behind, beam}), y = 1 \)
... 
\( x = (\text{happy, beam}), y = 0 \)
\( x = (\text{think, beam}), y = 0 \)
...
**Word2Vec: Context**

Word2Vec is a method for creating word embeddings that capture the context of words in a sentence. It uses two versions of context:

- **Continuous bag of words (CBOW):** Predicts a word from its context.
- **Skip-Grams (SG):** Predicts context words from the target word.

The goal is to train a model that can distinguish between positive and negative examples. The formula for calculating the probability of a positive example is:

$$ P(y=1| c, t) = \frac{1}{1+e^{-t \cdot c}} $$

where

- **$x$** is a positive context.
- **$y$** is the target word.
- **$P(y=1| c, t)$** is the probability of selecting $x$ as a positive context.
- **$\sigma(z) = \frac{1}{1+e^{-z}}$** is the logistic function.
- $t$ and $c$ are parameters.

**Steps for training:**

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish between the two cases.
4. Use the weights as the embeddings.

An example sentence is: "The nail hit the beam behind the wall."
Word2Vec: Context

The nail hit the beam behind the wall.

The intuition is that the cosine similarity $\mathbf{c} \cdot \mathbf{t}$ is a measure of similarity:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta$$

But, it is not a probability! To make it one, apply logistic activation:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the weights as the embeddings.

\[
\begin{align*}
x &= (hit, beam), y = 1 \\
x &= (the, beam), y = 1 \\
x &= (behind, beam), y = 1 \\
\vdots \\
x &= (happy, beam), y = 0 \\
x &= (think, beam), y = 0 \\
\vdots
\end{align*}
\]

\[
P(y=1 | \mathbf{c}, \mathbf{t}) = \frac{1}{1 + e^{-\mathbf{c} \cdot \mathbf{t}}}
\]

(Jurafsky, 2017)
Word2Vec: Context

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the weights as the embeddings.

Intuition: $t^\top c$ is a measure of similarity: $a \cdot b = ||a|| \cdot ||b|| \cdot \cos \theta$

But, it is not a probability! To make it one, apply logistic activation:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Single context:

$$P(y=1 | c, t) = \frac{1}{1 + e^{-t^\top c}}$$

All contexts:

$$P(y=1 | c, t) = \frac{1}{1 + e^{-t^\top c_i}} \prod_{i=1}^{k}$$

The nail hit the beam behind the wall.

$\text{(Jurafsky, 2017)}$
Word2Vec: How to Learn?

\[ P(y=1 \mid c, t) \]

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the weights as the embeddings.

The nail hit the beam behind the wall.

(Jurafsky, 2017)
Word2Vec: How to Learn?

P(y=1| c, t)
Assume 300 * |vocab| weights (parameters) for each of c and t

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the weights as the embeddings.

The nail hit the beam behind the wall.

(Jurafsky, 2017)
**Word2Vec: How to Learn?**

\[ P(y=1| c, t) \]

Assume 300 \* |vocab| weights (parameters) for each of c and t  
Start with random vectors (or all 0s)

1. Treat the target word and a neighboring context word as positive examples.  
2. Randomly sample other words in the lexicon to get negative samples.  
3. Use logistic regression to train a classifier to distinguish those two cases.  
4. Use the weights as the embeddings

*The nail hit the beam behind the wall.*  
(Jurafsky, 2017)
Word2Vec: How to Learn?

\[ P(y=1|c, t) \]

Assume 300 * |vocab| weights (parameters) for each of c and t
Start with random vectors (or all 0s)

Goal:
Maximize similarity of (c, t) in positive data (y = 1)
Minimize similarity of (c, t) in negative data (y = 0)

The nail hit the beam behind the wall.

(Jurafsky, 2017)
Word2Vec: How to Learn?

\[ P(y=1|c, t) \]

Assume \(300 \times |\text{vocab}|\) weights (parameters) for each of \(c\) and \(t\)
Start with random vectors (or all 0s)

Goal:
Maximize similarity of \((c, t)\) in positive data \((y = 1)\)
Minimize similarity of \((c, t)\) in negative data \((y = 0)\)

\[
\sum_{(c,t)} (y) \log P(y = 1|c, t) + (y - 1) \log P(y = 0|c, t)
\]
Word2Vec: How to Learn?

\[ P(y=1| c, t) \]

Assume 300 * |vocab| weights (parameters) for each of c and t
Start with random vectors (or all 0s)

Goal:
Maximize similarity of (c, t) in positive data (y = 1)
Minimize similarity of (c, t) in negative data (y = 0)

\[
\sum_{(c,t)} (y) \log P(y = 1|c, t) + (y - 1) \log P(y = 0|c, t)
\]

\[ 1 - P(y = 1|c, t) = \frac{e^{-t \cdot c}}{1 + e^{-t \cdot c}} \]
Word2Vec: How to Learn?

P(y=1| c, t)

Assume 300 * |vocab| weights (parameters) for each of c and t
Start with random vectors (or all 0s)

Goal:
Maximize similarity of (c, t) in positive data (y = 1)
Minimize similarity of (c, t) in negative data (y = 0)

\[ \sum_{(c,t)} (y) \log P(y = 1|c, t) + (y - 1) \log P(y = 0|c, t) \]

Optimized using gradient descent type methods.

\[ 1 - P(y = 1|c, t) = \frac{e^{-t \cdot c}}{1 + e^{-t \cdot c}} \]
Word 2 Vec

$\sum_{(c,t)} (y) \log P(y = 1|c, t) + (y - 1) \log P(y = 0|c, t)$

(Jurafsky, 2017)
Word2Vec captures analogies (kind of)

(Jurafsky, 2017)
Word2Vec: Quantitative Evaluations

Compare to manually annotated pairs of words: WordSim-353 (Finkelstein et al., 2002)

Compare to words in context (Huang et al., 2012)

Answer TOEFL synonym questions.
Current Trends in Embeddings

1. Contextual word embeddings (a different embedding depending on context):

*The nail hit the beam behind the wall.*

*They reflected a beam off the moon.*
Current Trends in Embeddings

1. Contextual word embeddings (a different embedding depending on context):
   The nail hit the beam behind the wall.
   They reflected a beam off the moon.

2. Embeddings can capture changes in word meaning.

(Kulkarni et al., 2015)
1. Contextual word embeddings (a different embedding depending on context):
   - The nail hit the beam behind the wall.
   - They reflected a beam off the moon.

2. Embeddings can capture changes in word meaning.

3. Embeddings capture demographic biases in data.

(Garg et al., 2018)
1. Contextual word embeddings (a different embedding depending on context):

   The nail hit the beam behind the wall.
   They reflected a beam off the moon.

2. Embeddings can capture changes in word meaning.

3. Embeddings capture demographic biases in data.
   a. Efforts to debias (Garg et al., 2018)
   b. Useful for tracking bias over time.
Current Trends in Embeddings

1. Contextual word embeddings (a different embedding depending on context):

   *The nail hit the beam behind the wall.*
   *They reflected a beam off the moon.*

2. Embeddings can capture changes in word meaning.

3. Embeddings capture demographic biases in data.
   a. Efforts to debias
   b. Useful for tracking bias over time.  
      (Garg et al., 2018)
Vector Semantics and Embeddings

Take-Aways

- Dense representation of meaning is desirable.
- Approach 1: Dimensionality reduction techniques
- Approach 2: Learning representations by trying to predict held-out words.
- Word2Vec skipgram model attempts to solve by predicting target word from context word:
  maximize similarity between true pairs; minimize similarity between random pairs.
- Embeddings do in fact seem to capture meaning in applications
- Dimensionality reduction techniques just as good by some evaluations.
- Current Trends: Integrating context, Tracking changes in meaning.