The Strange Geometry of Skip-Gram with Negative Sampling: A story of geometric observations

David Mimno and Laure Thompson
Cornell University
Ideally…
Ideally...

Words span the entire K-dimensional space
Ideally...

Words span the entire K-dimensional space

Distance reflects semantic relationships
Word Embeddings

Dense Vectors
Skip-Gram with Negative Sampling (SGNS)
Skip-Gram with Negative Sampling (SGNS)
SGNS: Skip-Gram Model

The brown fox jumps over the lazy dog.
SGNS: Skip-Gram Model

The brown fox jumps over the lazy dog.
SGNS: Skip-Gram Model

The brown fox jumps over the lazy dog.

Context Window Size = 2
SGNS: Skip-Gram Model

The brown fox jumps over the lazy dog.

Context Window Size = 2

jumps $\rightarrow$ \{ brown, fox, over, the \}
**SGNS**: Negative Sampling

Co-occurrence jumps, fox:
SGNS: Negative Sampling

Co-occurrence jumps, fox:

\[ W_{\text{jumps}} \]
\[ C_{\text{fox}} \]
\[ C_{\text{neg}} \]

\[ S \]

\[ W_{\text{jumps}} \rightarrow C_{\text{fox}} \]

\[ W_{\text{jumps}} \rightarrow C_{\text{neg}} \]
Experimental Setup

Corpus

Embeddings
word2vec
GloVe

Parameters
Vector Size: 50
Window Size: 5

Mean word vector

\[ \bar{w} = \text{mean} \left( \begin{array}{c}
\end{array} \right) \]
Observation #1:

SGNS vectors arrange along a primary axis
SGNS vectors point toward mean word vector

$$w_k \cdot \bar{w}$$

Closer to $$\bar{w}$$
Artifact of word frequency?

4 Frequency Levels:
  • Ultra-high (1–100)
  • High (101–500)
  • Moderate (501–5000)
  • Low (5001+)

• Sample 100 from each
• Use sample mean vector $\hat{w}$ instead of global mean $\bar{w}$
...true for all frequency classes

\[ w_k \cdot \hat{w} \]

Closer to mean word vector \( \hat{w} \)
...and away from the context vectors

\[ c_k \cdot \hat{w} \]

Further away from \( \hat{w} \)
Not true for GloVe

\[ w_k \cdot \hat{w} \quad \text{and} \quad c_k \cdot \hat{w} \]

Closer to \( \hat{w} \)
Visualization Disconnect

TSNE vs PCA
Observation #2:

SGNS vectors are mostly non-negative
Latent dimensions skew “positive”
...inefficient use of K-dimensional space?

Preserve Semantic Properties:

1. Dropping “negative” entries
   \[ w'_{k} = \max(0, w_{k} \times \text{sign}(\bar{w}_{k})) \]

2. Subtracting mean vector
   \[ w' = w - \bar{w} \]
Observation #3:

Negative sampling affects SGNS geometry
More negative samples, better alignment

$Avg(w_k \cdot \bar{w})$ vs. # of Negative Samples
…this is not seen in other parameters

\[ \text{Avg}(w_k \cdot \bar{w}) \text{ vs. Vector Size} \]
...this is not seen in other parameters

\[ \text{Avg}(w_k \cdot \bar{w}) \text{ vs. Window Size} \]
Thank You!