Global Vectors for Word Representation

GenAI Course - UM6P-CC

Abdeljalil Otman & Youness Anouar

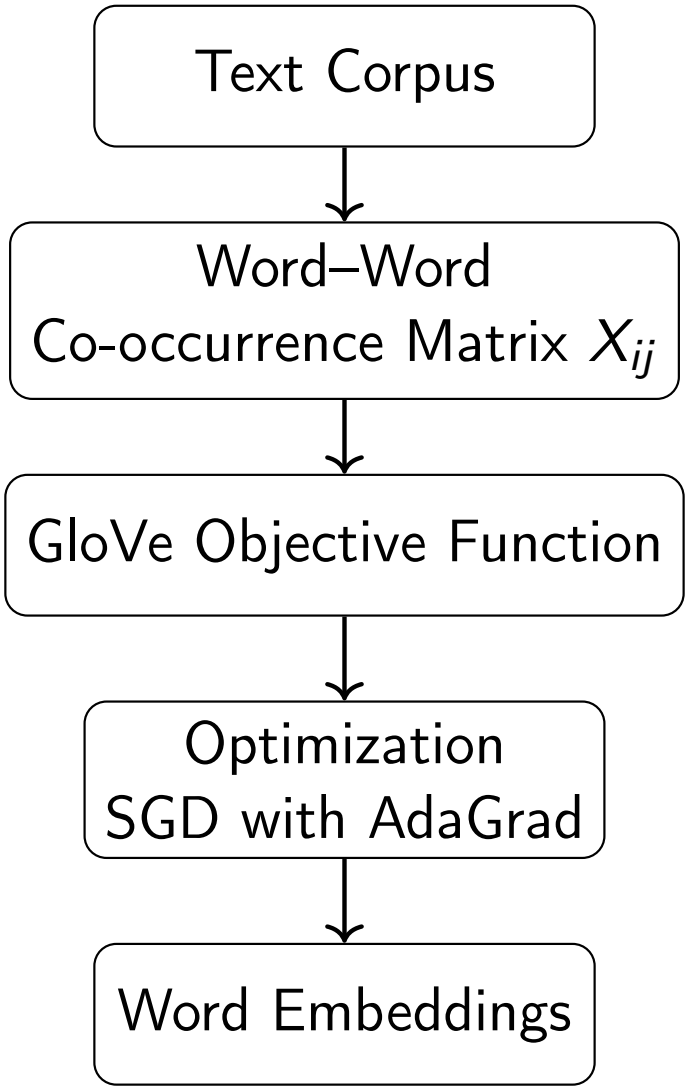
February 4, 2026

Outline

Biggest Takeaways: What is GloVe?

- **GloVe (Global Vectors)** is a word embedding model based on **global word–word co-occurrence statistics**
- Existing methods have limitations:
  - **LSA**: uses global statistics but weak at word analogies
  - **Word2Vec**: captures analogies but ignores global corpus structure
- **GloVe bridges both worlds** by combining:
  - Global matrix factorization
  - Local context window learning

GloVe: Learning Pipeline Overview



Technical Details: From Co-occurrence to Meaning

- Meaning is derived from **global word co-occurrence statistics**
- **Ratios of co-occurrence probabilities** capture semantic properties
- GloVe directly models these global statistics

Table 1: Co-occurrence probabilities for target words *ice* and *steam* with selected context words from a 6 billion token corpus. Only in the ratio does noise from non-discriminative words like *water* and *fashion* cancel out, so that large values (much greater than 1) correlate well with properties specific to ice, and small values (much less than 1) correlate well with properties specific of steam.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

- Co-occurrence matrix: X_{ij} = number of times word j appears in the context of word i
- Objective function:

GloVe Objective: Meaning of Each Term

$$\sum_{i,j} f(X_{ij}) \left(w_i^\top \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

- X_{ij}: number of times word j appears in the context of word i
- w_i: vector represent ation of the **target word** i
- $\tilde{w}_j$: vector representation of the **context word** j
- b_i, $\tilde{b}_j$: bias terms for target and context words
- $\log X_{ij}$: log co-occurrence count (stabilizes large values)
- $f(X_{ij})$: weighting function controlling the influence of rare and frequent pairs

Complexity Analysis

Worst case: $O(|V|^2)$, Loops over all word pairs in X (size $|V| \times |V|$).
Optimization: $O(|X|)$, Only consider nonzero entries of X, thanks to $f(0) = 0$ in the loss function:

$$|X| = \#\{X_{ij} \neq 0\},$$

- Assume

$$X_{ij} = \frac{k}{r_{ij}^{\alpha}},$$

  where r_{ij} is the rank of pair (i, j) in decreasing order.
- Size of corpus:

$$|C| \approx \sum_{i,j} X_{ij} = kH_{|X|,\alpha}, \quad H_{x,s} = \sum_{r=1}^{x} \frac{1}{r^s} = \frac{x^{1-s}}{(1-s)} + \zeta(s) + O(x^{-s})$$

(the expansion of generalized harmonic numbers, (Apostol, 1976))

Complexity Analysis

For large $|X|$, we can approximate:

$$|C| \sim \frac{|X|}{1-\alpha} + \zeta(\alpha)|X|^{\alpha} + O(1)$$

Hence, the number of nonzero entries:

$$|X| \sim \begin{cases} O(|C|), & \alpha < 1 \\ O(|C|^{1/\alpha}), & \alpha > 1 \end{cases}$$

With $\alpha = 1.25 > 1$:

$$|X| = O(|C|^{0.8}) \quad \text{(better than window-based } O(|C|))$$

Main Tasks: Word Analogy

Task

a is to b as c is to ..?

- **Semantic analogies:** people, countries, capitals
  - Example: Athens : Greece :: Berlin : ?
- **Syntactic analogies:** tense, plurality, adjective forms
  - Example: dance : dancing :: fly : ?

Word Analogy: Evaluation

Evaluation Method

1. Compute prediction vector:

$$v_{\text{pred}} = v_b - v_a + v_c$$

2. Find nearest word using cosine similarity:

$$w = \arg\max_v \ \text{sim}(v, v_{\text{pred}})$$

3. Count as correct if:

$$w = d$$

- Accuracy = percentage of correctly solved analogies
- Reported separately for **semantic**, **syntactic**, and **total**

Main Tasks: Word Similarity

Task

Do similar words have similar embeddings?

- Example:

$$\text{sim(car, automobile)} > \text{sim(car, banana)}$$

Evaluation

- Compute cosine similarity between word vectors
- Compare model scores with human similarity judgments
- Use **Spearman's rank correlation** as the metric

Main Tasks: Named Entity Recognition (NER)

Task

Identify and classify named entities:

- Person, Location, Organization, Miscellaneous

Evaluation Pipeline

Embeddings as features $\rightarrow$ CRF model $\rightarrow$ Entity prediction $\rightarrow$ F1 score

- Word embeddings augment traditional discrete features
- Performance measured using Precision, Recall, and F1
- CRF = Conditional Random Field. It's a probabilistic graphical model used for structured prediction.

Code Summary: GloVe Training

Algorithm 1: GloVe Optimization Procedure

Input: Co-occurrence counts X_{ij}
Output: Word vectors $w, \tilde{w}$
for *each training iteration* **do**
  total_cost $\leftarrow 0$;
  for *each chunk of data* **do**
    for *each co-occurrence* (i, j, X_{ij}) **do**
      pred $\leftarrow w_i^\top \tilde{w}_j + b_i + \tilde{b}_j$;
      error $\leftarrow$ pred $- \log X_{ij}$;
      $L \leftarrow L + \frac{1}{2} f(X_{ij}) \cdot \text{error}^2$;
      Compute gradients: $\frac{\partial L}{\partial w_i}$, $\frac{\partial L}{\partial \tilde{w}_j}, \frac{\partial L}{\partial b_i}, \frac{\partial L}{\partial \tilde{b}_j}$;
      Update parameters using AdaGrad (lr=f(total_cost));
    end
    total_cost $\leftarrow$ total_cost $+L$;
  end
end

Results: Word Analogy

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

Figure: Accuracy (%) on the word analogy task. Underlined: best within model size group; bold: best overall.

Dataset: Mikolov et al. (2013a) Word Analogy Dataset

Results: Word Similarity

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	72.7	75.1	56.5	37.0
CBOW†	6B	57.2	65.6	68.2	57.0	32.5
SG†	6B	62.8	65.2	69.7	58.1	37.2
GloVe	6B	65.8	72.7	77.8	53.9	38.1
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	**75.9**	**83.6**	**82.9**	**59.6**	**47.8**
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

Figure: Spearman rank correlation on word similarity tasks (300-dimensional vectors).

Datasets: WordSim-353 (Finkelstein et al., 2001), MC (Miller & Charles, 1991), RG (Rubenstein & Goodenough, 1965), SCWS (Huang et al., 2012), RW (Luong et al., 2013)

Results: Named Entity Recognition (NER)

Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**

Figure: F1 score on the NER task using 50-dimensional vectors. *Discrete* denotes the baseline without word embeddings.

Training: CoNLL-2003 training set
Evaluation: CoNLL-2003 test, ACE (2001–2003), MUC-7

Limitations and Future Directions

- **Outdated vocabulary & semantic drift:** New words are missing and meanings of existing words shift over time, leading to poor performance on modern datasets.
  - Retraining on recent corpora, including social media and web text
- **Temporal & domain bias (older, Western-centric data)** biased toward the West
  - More diverse, global, recent corpora

Limitations and Future Directions

- **Poor handling of rare words:** Rare words suffer from noisy co-occurrence statistics and data sparsity.
  - Introduced a Minimum Frequency Threshold (MFT) strategy inspired by GloVe-V
- **Overestimation of antonym similarity:** Distributional similarity places antonyms close together.
  - Acknowledged in analysis; addressing this would require lexical constraints or contextual models

Summary

- **GloVe** learns word embeddings by leveraging **global word–word co-occurrence statistics.**
- GloVe vectors are useful in downstream NLP tasks:
  - Word analogy
  - Word similarity
  - Named Entity Recognition (NER)
- Limitations of GloVe (static, word-level representations) motivated later advances:
  - Recurrent Neural Networks (RNNs) & Long Short-Term Memory networks (LSTMs)
  - Transformers

References

- **GloVe: Global Vectors for Word Representation**
  Jeffrey Pennington, Richard Socher, Christopher D. Manning
  Computer Science Department, Stanford University, Stanford, CA 94305
- **A New Pair of GloVes**
  Riley Carlson, John Bauer, and Christopher D. Manning
  Stanford NLP Group, Stanford University
  353 Jane Stanford Way, Stanford, CA 94305-9035, U.S.A.
- GloVe Official Repository:
  https://github.com/stanfordnlp/GloVe/tree/master

Questions?

Thank you :) !