GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

Presenters:

Sohan Patnaik and Yatindra Indoria

Reading Session XII

Kharagpur Data Analytics Group

IIT Kharagpur

September 19, 2021

Important Notes

- We expect you all to have certain queries regarding the presentation, certain intricate doubts maybe... **Please put them in the chat of this meeting (if you feel shy) else unmute and speak**.
- Finally, a **Google Form** [Link] will be released for feedback but most importantly, we ask you to put up name of any AI-related paper to be presented in the upcoming sessions!

Link to GitHub Repo : Click Here
Link to join Slack Workspace : Click Here
Link to KDAG YouTube Channel : Click Here
Link to doc on good AI Blogs/Resources/Topics : Click Here

Outline

- Introduction
- The GloVe Model
  - Notations
  - Complexity of the model
- Experiments
  - Evaluation Methods
  - Corpora and Training Details
  - Results
  - Model Analysis
- Conclusion

Introduction

1) Previous models succeeded in capturing fine-grained semantic and syntactic regularities, but the origin had remained opaque.

2) LSA efficiently leverage statistical information, while perform poorly on the word analogy task, indicating a sub-optimal vector space structure.

3) Skip-gram does better on the analogy task, but it poorly utilizes the statistics of the corpus.

4) **GloVe Model:** A log bilinear model that combines global matrix factorization and local context window methods.

The GloVe Model: Notations

- X - Matrix of word-word co-occurrence where Xij is the number of times word j occurs in the context of word i.
- Xi - Number of times any word appears in the context of word i.
- Pij = P(j | i) = Xij / Xi - Probability that word j appeach in context of i

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

- The appropriate starting point for word vector learning should be with ratios of co-occurrence probabilities rather than the probabilities themselves.

The GloVe Model: Model Description

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

- wi, wj are the d-dimensional word vectors and ˜w are separate context word vectors and F is some function.
- The most natural way to encode function F is to take vector differences since vector spaces are inherently linear structures.

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

- The right hand side is a scalar, so intuitively taking dot product makes sense!

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

The GloVe Model: Model Description

- For word-word co-occurrence matrices, the distinction between a word and a context word is arbitrary and we are free to exchange the two roles. We must not only exchange w ↔ ˜w but also X ↔ XT.

- For this, F should be homomorphism between two groups (R, +) and (R>0, x).

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$$

So we get

$$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$$

The GloVe Model

The solution to F from the previous equation is intuitively exponential.

$$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$$

Consider log Xi as bias bi and to keep symmetry with respect to i and k, introduce another bias bj.

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

A main drawback to this model is that it weighs all co-occurrences equally, even those that happen rarely or never.

The GloVe Model: Model Description

Solution to the previous problem is a weighted least squares regression.

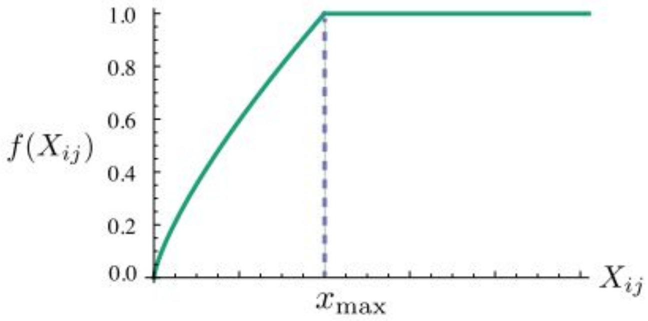
$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

Some properties of f(x) are

f(0) = 0.

f(x) should be non-decreasing.

f(x) should be relatively small for large values of x.



The GloVe Model: Complexity of the model

- The computational complexity of the model depends on the number of nonzero elements in the matrix X.
- This number is always less than the total number of entries of the matrix, so it scales it has quadratic complexity to the size of vocabulary.
- Assumptions to estimate count of non-zero elements in X.
- Xij can be modelled using the power law.

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

Where, rij is the frequency rank of word pair

The GloVe Model: Complexity of the model

- The total number of words in the corpus is proportional to the sum over all elements of the co-occurrence matrix X.

$$|C| \sim \sum_{ij} X_{ij} = \sum_{r=1}^{|X|} \frac{k}{r^{\alpha}} = kH_{|X|,\alpha}$$

Where the last sum is written in terms of the generalized harmonic number Hn,m.

- The upper limit of the sum, is the maximum frequency rank, which coincides with the number of nonzero elements in the matrix X, equal to the maximum value of r such that Xij ≥ 1, i.e., |X| = k^(1/α).

$$|C| \sim |X|^{\alpha} H_{|X|,\alpha}$$

- Using harmonic function approximation and reimann zeta function, |X| is approximated as follows:

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

Experiments: Evaluation Methods

- 1st experiment is conducted on the Mikolov word-analogy task.
- Secondly, there are word similarity tasks like WorldSim-353, MC, RG, SCWS, RW
- Model is also Evaluated on NER (Named Entity Recognition).

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**

Word Analogy

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

Word Similarity

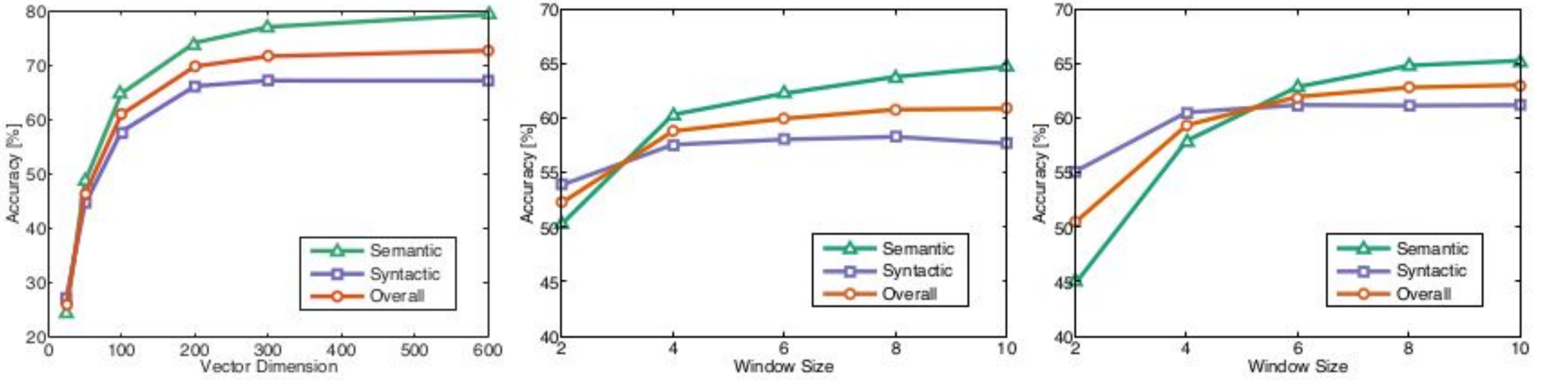
Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**

Named Entity Recognition

Experiments: Corpora and Training Details

- The model is trained on 5 different corpora of varying sizes.
- The corpus is tokenized using the Stanford tokenizer, to build a vocabulary of the 400,000 most frequent words.
- A weighting function is used in all cases such that words that are 'd' apart, contribute 1/d to the count.
- The model generates two sets of word vectors, W and $\tilde{W}$. When X is symmetric, W and $\tilde{W}$ are equivalent and differ only as a result of their random initializations; the two sets of vectors should perform equivalently. Thus we use $W + \tilde{W}$ in the model.

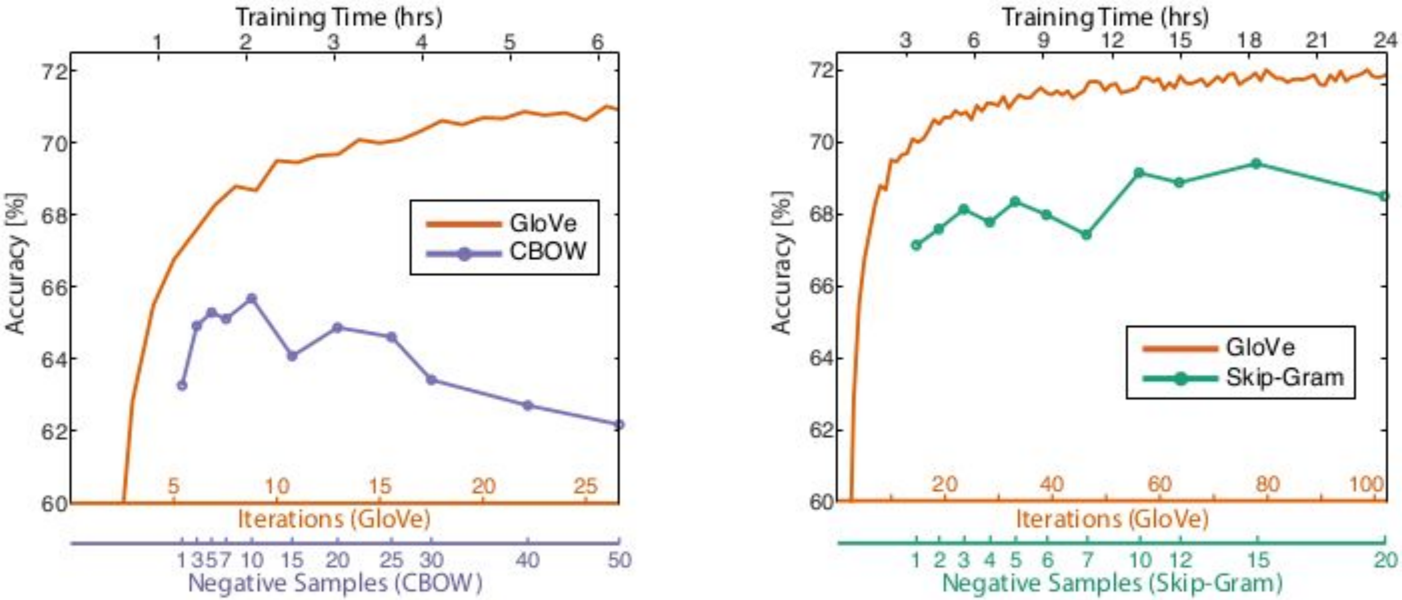
Experiments: Model Analysis



(a) Symmetric context (b) Symmetric context (c) Asymmetric context

Figure 2: Accuracy on the analogy task as function of vector size and window size/type. All models are trained on the 6 billion token corpus. In (a), the window size is 10. In (b) and (c), the vector size is 100.

Experiments: Model Analysis



(a) GloVe vs CBOW

(b) GloVe vs Skip-Gram

Conclusion

- The two classes of methods (contextual and statistical) are not completely different at fundamental level as they both probe the underlying co-occurrence statistics of the corpus.
- Combining the two approaches, they proposed a global log bilinear model which leverages the advantages of both the approaches.

Thank You