GloVe : A Review

Barnik Pal & Sucheta Ghosh

INTRODUCTION

- Semantic vector space models of language represent each word with a real-valued vector. These vectors can be used as features in a variety of applications, such as information retrieval, document classification, question answering, and so on.
- The models should capture the multi-clustering idea of distributed representations and various dimensions of difference between word vectors. For example, the analogy "king is to queen as man is to woman" should be encoded in the vector space by the vector equation "*king - queen ≈ man - woman*".
- Methods like skip-gram may do better on the analogy task, but they poorly utilize the statistics of the corpus since they train on separate local context windows instead of on global co-occurrence counts.

THE GLOVE MODEL

- The following model for word representation is called GloVe, for Global Vectors, because the global corpus statistics are captured directly by the model.
- First we establish some notation. Let the matrix of word-word co-occurrence counts be denoted by $\boldsymbol{X}$. Let $X_i = \sum_k X_{ik}$ be the number of times any word appears in the context of word 'i'. Finally, let $P_{ij} = X_{ij}/X_i$ be the probability that word 'j' appear in the context of word 'i'.
- We begin with a simple example that showcases how certain aspects of meaning can be extracted directly from co-occurrence probabilities.

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 above example suggests that the appropriate starting point for word vector learning should be with ratios of co-occurrence probabilities rather than the probabilities themselves. The most general model takes the form:

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

- We would like F to encode the information present the ratio P_{ij}/P_{jk} in the word vector space. Since vector spaces are inherently linear structures, the most natural way to do this is with vector differences. Next, we note that the arguments of F are vectors while the right-hand side is a scalar. To avoid this issue, we can first take the dot product of the arguments,

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

- Note that for **X**, the distinction between a word and a context word is arbitrary and that we are free to exchange the two roles. To do so consistently, we must not only exchange $w \leftrightarrow \tilde{w}$ but also X $\leftrightarrow X^T$. Our final model should be invariant under this relabeling,

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

- The solution to Eqn. (4) is F = exp, or

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

❖However, $\log(X_i)$ is independent of k so it can be absorbed into a bias b_i for w_i . Also we include an additive shift in the logarithm which maintains the sparsity of X while avoiding the divergences,

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(1 + X_{ik})$$

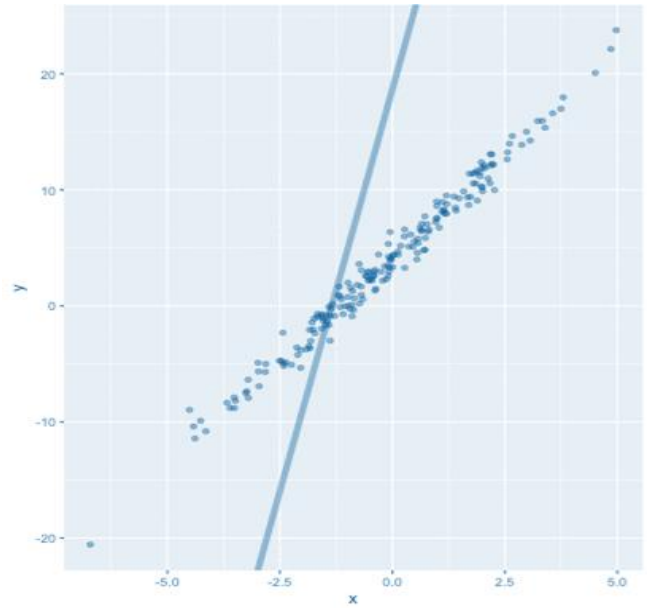
❖A main drawback to this model is that it weighs all co-occurrences equally. The GloVe model proposes a new weighted least squares regression model that addresses these problems:

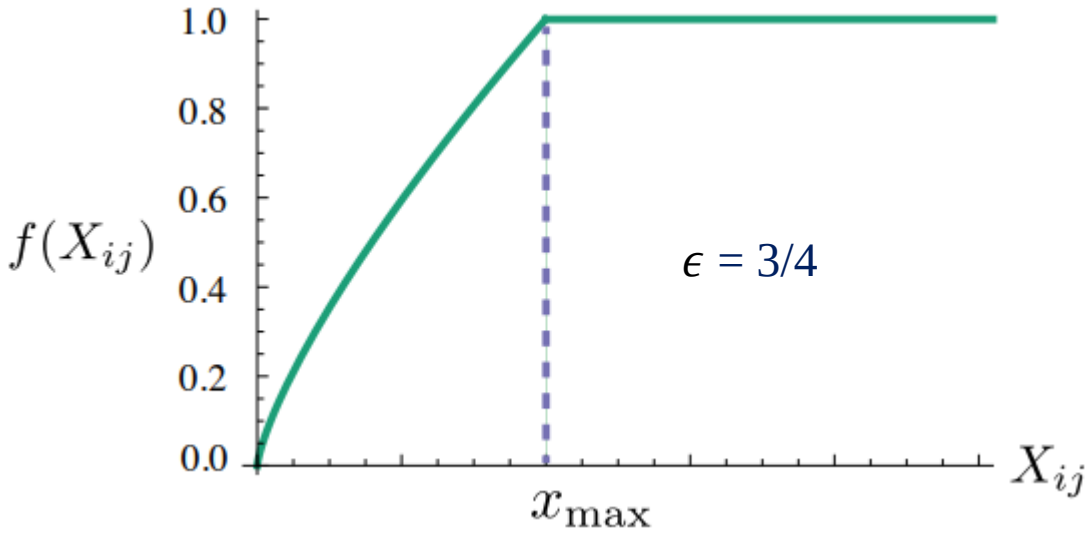
$$J = \sum_{i,j=1}^{V} f(X_{ij})\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log(X_{ik})\right)^2$$

❖Here f (x) should be non-decreasing so that rare co-occurrences are not overweighted and should be relatively small for large values of x, so that frequent co-occurrences are not overweighted. The weighting function can be written as:

$$f(x) = \begin{cases} (x/x_{max})^{\epsilon}, & x < x_{max} \\ 1, & otherwise \end{cases}$$

❖The performance of the model depends weakly on the cutoff, with $\boldsymbol{x_{max} = 100}$ for all experiments. Also, $\boldsymbol{\epsilon = 3/4}$ gives a modest improvement over a linear version.





Complexity of the Model:

- The computational complexity of the model depends on the number of nonzero elements in the matrix X. As this number is always less than the total number of entries of the matrix, the model scales no worse than $\mathcal{O}(|V|^2)$.
- But $|V|^2$ can be in the hundreds of billions, which is actually much larger than most corpora. For this reason it is important to determine whether a tighter bound can be placed.
- It is assumed that X_{ij}, can be modeled as a power-law function of the frequency rank of that word pair, r_{ij}:

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

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

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

- We are interested in how |X| is related to |C| when both numbers are large; therefore we expand the RHS of the equation for large |X|:

$$H_{x,s} = \frac{x^{1-s}}{1-s} + \zeta(s) + \mathcal{O}(x^{-s}) \;\; if \; s > 0, s \neq 1$$

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

- In the limit that X is large, we have:

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

- For the corpora studied in this article, we observe that α = 1.25. In this case we have that:

$$|X| = \mathcal{O}(|C|^{0.8})$$

- Therefore we conclude that the complexity of the model is much better than $\mathcal{O}(|V|^2)$.

EXPERIMENTS AND ANALYSIS

Word analogies

- The word analogy task consists of questions like, "a is to b as c is to ?" The dataset contains 19,544 such questions, divided into a semantic subset and a syntactic subset. The semantic questions are typically analogies about people or places. The syntactic questions are typically analogies about verb tenses or forms of adjectives. The model answers the question by finding the word 'd' whose representation w_d is closest to $w_b - w_a + w_c$ according to the cosine similarity.

- The GloVe model performs significantly better than the other baselines, often with smaller vector sizes and smaller corpora.

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 Similarity

❖While the analogy task is our primary focus since it tests for interesting vector space substructures, we also evaluate our model on a variety of word similarity tasks.

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

❖The table above shows results on five different word similarity datasets. A similarity score is obtained from the word vectors by first normalizing each feature across the vocabulary and then calculating the cosine similarity. We compute Spearman's rank correlation coefficient between this score and the human judgments. GloVe outperforms CBOW* while using a corpus less than half the size.

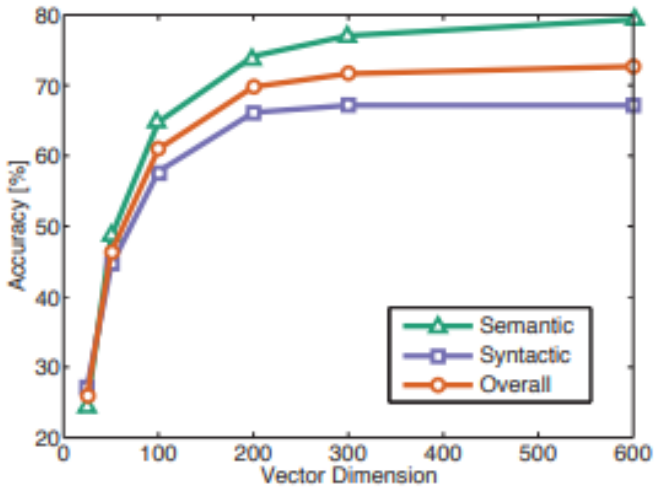
Named Entry Recognition

❖The model is tested on the CoNLL-2003, which is a collection of documents from Reuters newswire articles, annotated with four entity types: person, location, organization, and miscellaneous.

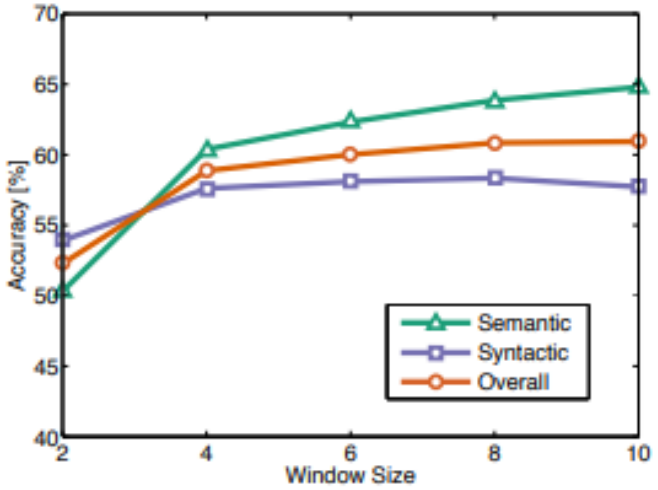
Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**

❖The above table shows results on the NER task with the CRF-based model. The model labeled Discrete is the baseline using a comprehensive set of discrete features that comes with the standard distribution of the Stanford NER model, but with no word vector features. The GloVe model outperforms all other methods on all evaluation metrics, except for the CoNLL test set, on which the HPCA method does slightly better. It is evident that the GloVe vectors are useful in downstream NLP tasks.

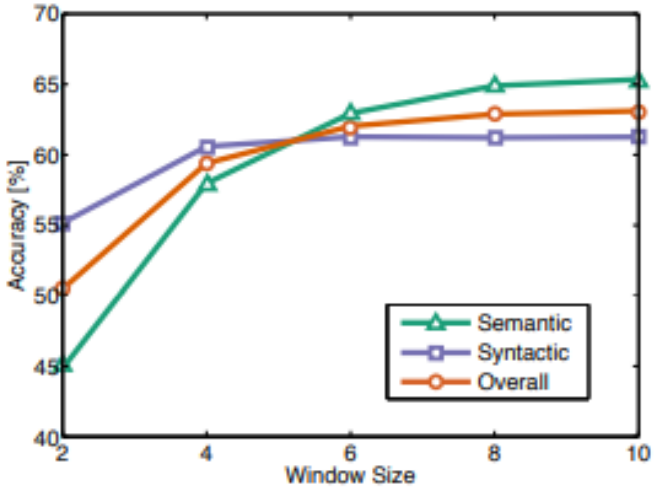
Model Analysis: Vector Length and Context Size



(a) Symmetric context



(b) Symmetric context

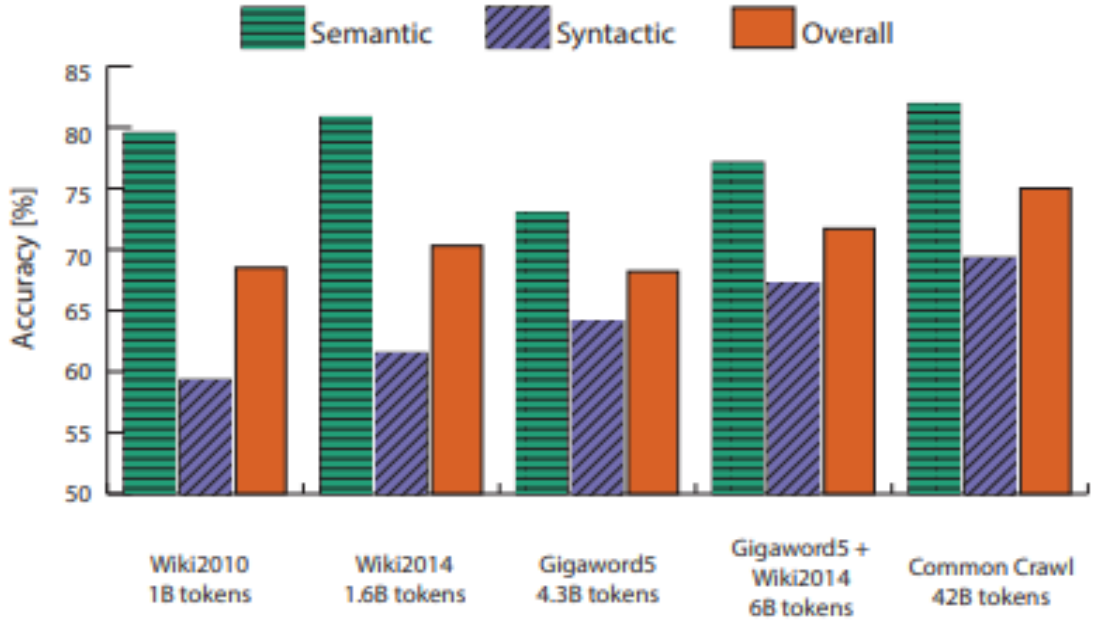


(c) Asymmetric context

- A context window that extends to the left and right of a target word will be called symmetric, and one which extends only to the left will be called asymmetric. 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.

- Performance is better on the syntactic subtask for small and asymmetric context windows, which aligns with the intuition that syntactic information is mostly drawn from the immediate context and can depend strongly on word order. Semantic information, on the other hand, is more frequently non-local, and more of it is captured with larger window sizes.

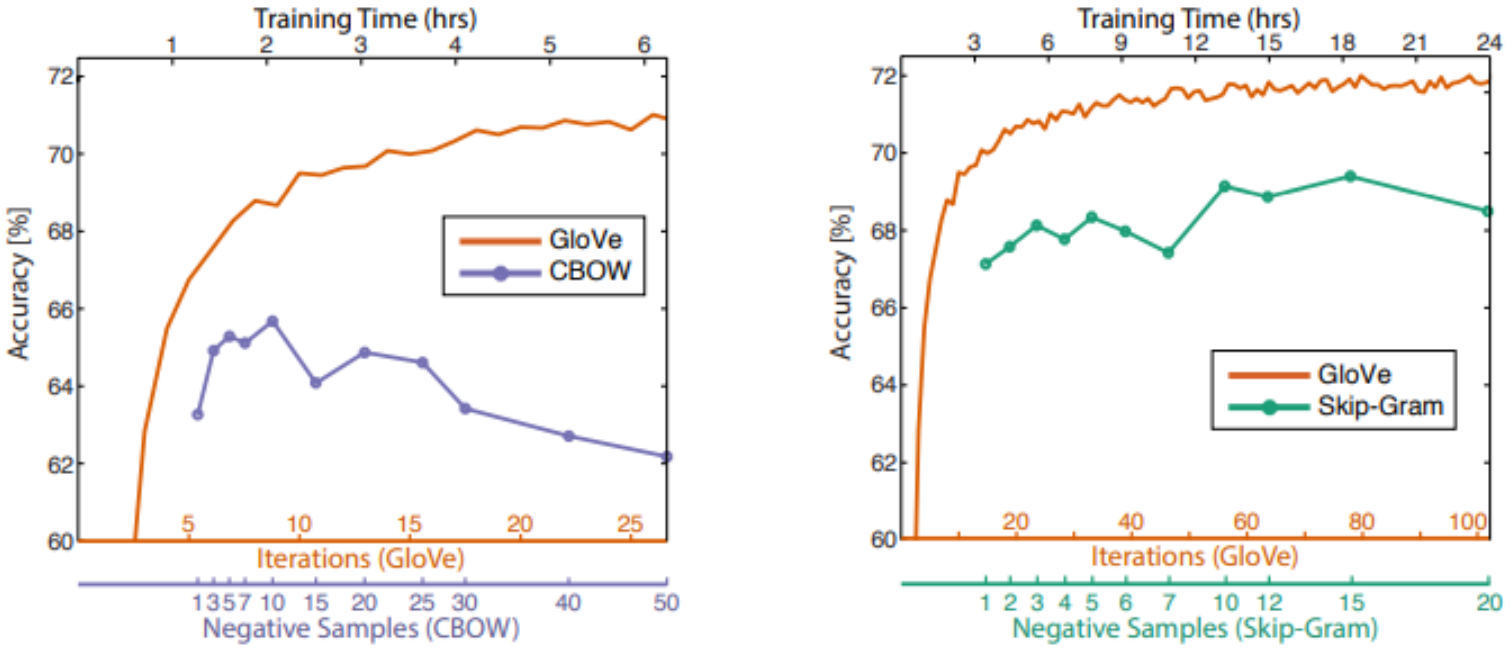
Model Analysis: Corpus Size



- The above figure shows performance on the word analogy task for 300-dimensional vectors trained on different corpora. On the syntactic subtask, there is a monotonic increase in performance as the corpus size increases. This is to be expected since larger corpora typically produce better statistics.

- Interestingly, the same trend is not true for the semantic subtask, where the models trained on the smaller Wikipedia corpora do better than those trained on the larger Gigaword corpus. This is likely due to the large number of city- and country based analogies in the analogy dataset and the fact that Wikipedia has fairly comprehensive articles for most such locations.

Model Analysis: Run-time



- The above figure plots the overall performance on the analogy task as a function of training time. We note that word2vec's performance decreases if the number of negative samples increases beyond about 10. Presumably this is because the negative sampling method does not approximate the target probability distribution well.

- For the same corpus, vocabulary, window size, and training time, GloVe consistently outperforms word2vec. It achieves better results faster, and also obtains the best results irrespective of speed.

CONCLUSION

- Recently, considerable attention has been focused on the question of whether distributional word representations are best learned from count-based methods or from prediction-based methods.
- In this work we argue that the two classes of methods are not dramatically different at a fundamental level since they both probe the underlying co-occurrence statistics of the corpus, but the efficiency with which the count-based methods capture global statistics can be advantageous.
- The result, GloVe, is a model that utilizes this main benefit of count data while simultaneously capturing the meaningful linear substructures.
- The source code for the model as well as trained word vectors can be found at:

https://nlp.stanford.edu/projects/glove/

THANK YOU

FOR YOUR ATTENTION !

ANY QUESTIONS ?