GloVe

Global Vectors for Word Representation

Presented by Sanjana Sahayaraj

Introduction

- Recent methods of word vectorization have captured the grammatical and lexical function structure
- But the origin of this structure has remained opaque
- In this paper they analyze and present the model properties for such syntactic and semantic regularities to appear in word vectors
- Literature
    - Global matrix factorization
    - Local context window methods
- Log bilinear regression model combines the benefits of both while trying to get rid of the shortcomings

Other models

- Word embedding - real-value vectors
- Mostly - distance or angle between pairs of word vectors
- 2013 paper by Mikolov et al. *Linguistic regularities in continuous space word representation* - analogies
- Example: king-queen = man-woman. Dimensions of meaning
- **Matrix factorization:**
    - Leverages statistical information of the corpus
    - Does poorly on the word analogy task
- **Local content window:**
    - Does better on analogy task
    - Do not utilize statistics of the corpus

How they work

- Matrix factorization:
  - Large matrix capturing statistical information about the corpus
- LSA (Latent Semantic Analysis):
  - Term document - rows are words and columns are documents
- HAL (Hyperspace Analogue to Language):
  - Rows are words & columns are no. of times a words occurs in content of the other
- HAL - disproportionate amount of similarity measure. Eg. and, the
- COALS (Correlated Occurrence Analogue to Lexical Semantics) method - correlation or entropy based normalization
- Newer model - Positive pointwise mutual information (PPMI) - based on co-occurrence counts

- Shallow window based methods
- Local context regions or windows
    - Simple neural network architecture
- 2003 - context with respect to previous word
- 2008 to 11 - full content for a word rather than just the preceding context
- 2013 - *Efficient estimation of word representations in vector space* - Single layer neural net based on inner product of two word vectors and *Vector log bilinear models*
- Skip-gram - predict word’s context given word itself
- CBOW - predict word given it’s context
- Learns linguistic patterns - but not global and fails to see repetition in data

How GloVe works

- Notation
  - *X :* word co-occurence matrix
  - X_{ij} : number of times word j occurs in the context of word i
  - X_i : number of times any word occurs in context of word i and is the summation of X_{ik} for all k
  - $P_{ij} = P(j \mid i) = X_{ij} / X_i$, is the probability that word j would occur in the context of word i
- Example:

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

- From the example, argument is that ratios of co-occurrence probabilities should be appropriate for word vectors
- To generalize the model in terms of a function:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}, \qquad (1)$$

Where w is word vector for the word under consideration and ŵ is context word vector - 10 words in front and 10 words at the back

- Some transformations to the equation need to be done
  - Make it linear - subtraction
  - Vector vs scalar - dot product
  - Be able to exchange w and ŵ and require F be a homomorphism between (R,+) and ($R_{>0}$, x)
  - Log
  - Add additional bias

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}) \, . \qquad (7)$$

Eqn. (7) is a drastic simplification over Eqn. (1),

- Logarithm diverges when argument is zero
- Additive shift in the logarithm $\log(X_{ik}) = \log(1+X_{ik})$
- Still weighs co-occurrences equally
  - Introduce a weighting function based on least squares

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

Where V is the size of the vocabulary

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

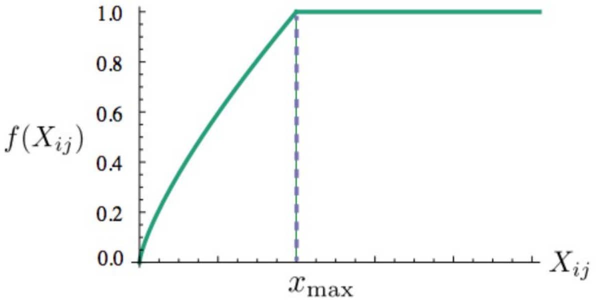


Figure 1: Weighting function f with $\alpha = 3/4$.

Relationship to skip gram

- Occurrence statistics - commonalities
- Skip-gram
  - Probability of j occurring in context of i
  - Q_{ij} is the probability that word j appears in context of word i - softmax
  - Since softmax is expensive - approximation is used which is similar to the weighted function

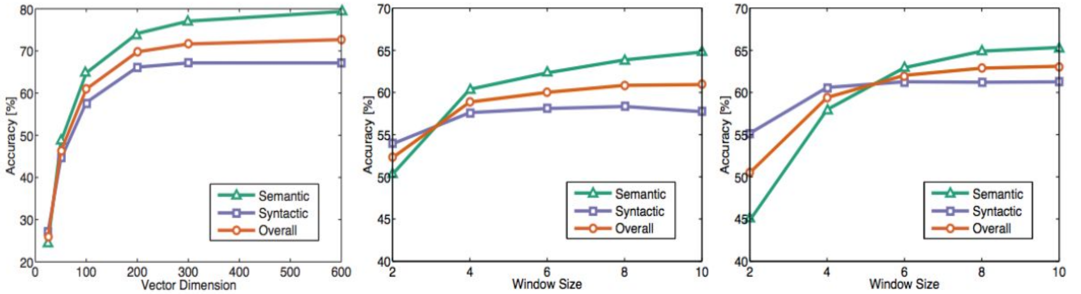
$$J = -\sum_{i=1}^{V}\sum_{j=1}^{V} X_{ij} \log Q_{ij}\,, \qquad (12)$$

Relationship - GloVe is global skip gram

Performance and evaluation metrics

Compared performance over word similarity/ analogy tasks - percent accuracy.

CoNLL 2003

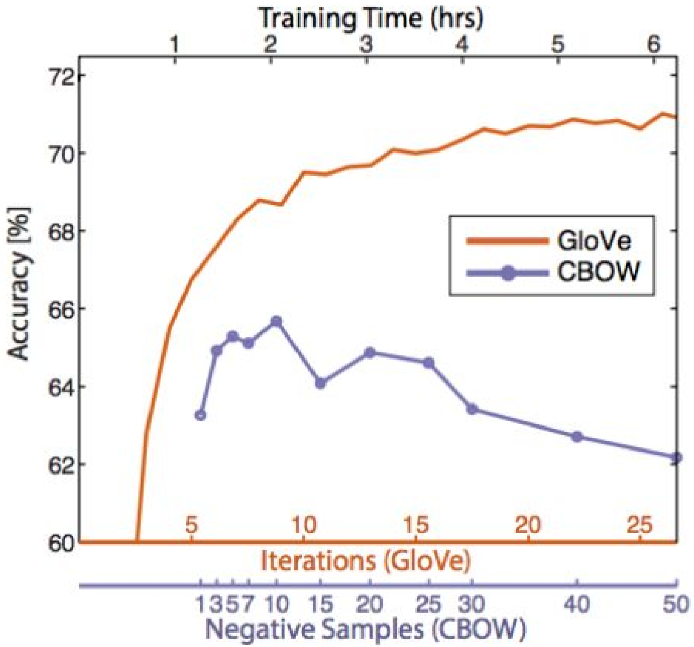


(a) Symmetric context (b) Symmetric context (c) Asymmetric context

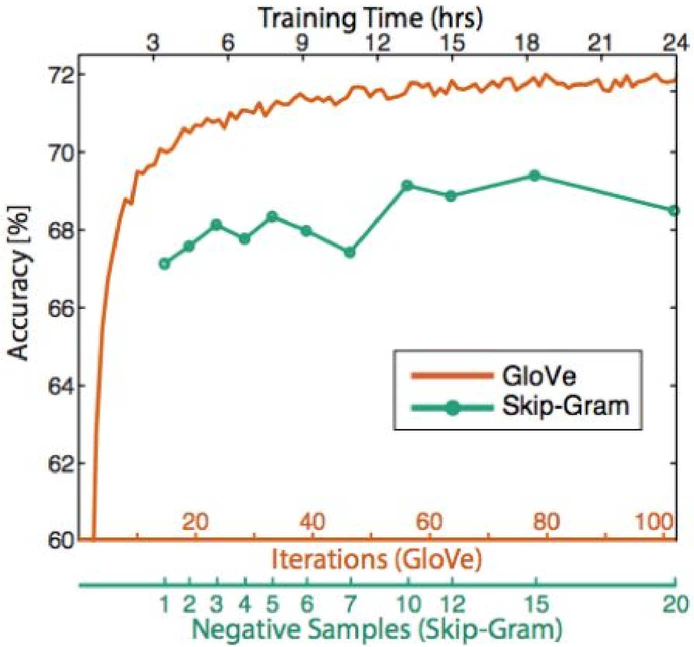
Figure 2: Accuracy on the analogy task as function of vector size and window size/type. All models are trained on the 6 billion token corpus. In (a), the window size is 10. In (b) and (c), the vector size is 100.

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**

Training time and accuracy



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

Conclusion

- Count based or prediction based
- While showing the two are not dramatically different
- Construct a model that utilizes main benefits of count based while capturing substructures from log - bilinear prediction based methods
- Result is GloVe - unsupervised learning of word representations
- Uses:
  - Document classification
  - NER
  - Question answering