GloVe: Global Vectors (Pennington, Socher and Manning 2014)

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

Bibliographical Reference

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

The word2vec Algorithm

- Imposes a sliding window over the whole corpus and goes through whole corpus one focus word at a time.
- Skipgram with negative sampling predicts surrounding words of each word focus word one at a time.
- Skipgram with negative sampling is trained by a logistic regression model with cross-entropy loss function.
    - Captures cooccurrences of words only locally with a given window and does not capture global cooccurrences over the entire corpus.
    - is inefficient in that repeated cooccurrences are re-computed ”from scratch”.
- By contrast: GloVe computes cooccurrences counts over the entire corpus.
- The GloVe model is trained by a weighted linear regression model with a least squared loss function.

Previous Approaches for Learning Word Vectors

Global matrix factorization methods such as Latent Semantic Analysis (LSA; Deerwester et al. 1990)

- widely used in Information Retrieval (IR) and based term-document matrices
- uses Singular Value Decomposition (SVD) to rerank the dimensions of a matrix from most to least informative
- LSA, practicioners assume that only the top 300 or so dimensions (out of tens or even hundreds of thousands) are useful for capturing the meaning of texts.
- downsides of LSA:
  - not suitable for very large corpora
  - does not adequately capture the substructure of the vector space and thus does poorly on analogy tasks.

Previous Approaches for Learning Word Vectors

Local context window methods such as the Skipgram Model

- uses a sliding window of local contexts over a large corpus.
- does not directly capture global information of the corpus.

The GloVe Approach

GloVe is a global log-bilinear regression model. More specifically:

- a weighted least squares model trained on global word-word co-occurrence counts obtained from a large corpus, rather than on
    - sparse term-term matrices
    - a sliding window of local contexts over a large corpus
- uses a term-term co-occurrence matrix
- supports fast training
- good performance even with small corpora and small vectors
- scalable to huge corpora

Distinguishing ratios of target words with discriminative and non-discriminative context words

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

with target words:	*ice, steam*
with discriminative words:	*gas, solid*
with ”noise” words:	*water, fashion*

Loss Function for a Weighted Linear (aka: Least Squares) Regression Model

$$J = \sum_{i,j=1}^{V} = f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b_j} - logX_{ij})^2, where \qquad (1)$$

(i) f is a weighting function:

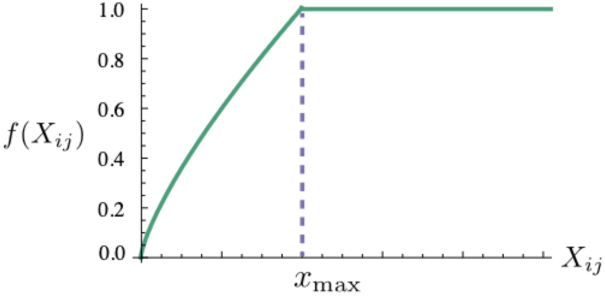
$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max} \\ 1 & otherwise \end{cases} \qquad (2)$$

(ii) $X_{i,j}$ tabulates the number of times word j occurs in the context of word i: $X_{i,j}/X_i$

Minimizing the Loss Function J for the GloVe Model by Gradient Descent

- amounts to updating the word vectors in such a way that the values for ${w_i}^T \tilde{w}_j + b_i + \tilde{b}_j$ and for $logX_{ij}$ is successively minimized.

Weighting Function f with $\alpha = 3/4$ and $x_{max} = 1$



For more detailed discussion

watch the youtube video by **Richard Socher**:

GloVe: Global Vectors for Word Representation.

`https://www.youtube.com/watch?v=ASn7ExxLZws&t=2376s`