Technology Review

Paper: **GloVe: Global Vectors for Word Representation.**

Authors: **Jeffrey Pennington, Richard Socher, ChristopherD. Manning**
Computer Science Department, Stanford University, Stanford, CA

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

Introduction:

In this paper a group of researchers from the Stanford university tries to show how a model efficiently leverages the statistical information by training only on the nonzero elements in the word-word co-occurrence matrix, rather than on entire spare matrix or on the large corpus of individual context windows and showed evidence of its performance better than other related models.

Global Vectors (**Glo Ve**) is a new global log-bilinear regression modal for the unsupervised learning of word representations that outperforms other models on word analogy, word similarity and name entity recognition tasks, which combines the advantages of the two major model families in the literature **global matrix factorization** and **local context window methods**. And the resulting representations showcase interesting linear substructures of word vector space.

Before this paper is proposed a primary method for evaluating word representations by a vector space models which are used in the information retrieval, document classification applications rely on the distance or angle between the pair of word vectors. The popular main model families for learning word vectors are global matrix factorization like latent semantic analysis (LSA) and local content window method suffers with major drawback, Though LSA efficiently leverages statistical information but perform poorly on word analogy. On other side skip-gram model handles word analogy in a better way but fails with utilizing statistics of corpus. In this paper authors tries to combine two positive side of earlier models to overcome the drawbacks. By considering the global log-bilinear regression models and global word-word co-occurrence counts, the new model produces word vector space with meaningful substructure to a state of art performance up to 75% accuracy on word analogy data set.

Lets shed some light on the Glo Ve model vector, Authors first took a co-occurrence matrix (denoted by X) with word to word counts which results a symmetric matrix. Then a Probability is calculated from the matrix.

$\mathbf{P_{ij} = P(j|i) = X_{ij}/X_i}$, where

P_{ij} or P(j|i) is the probability of word j,i may appear together.

X_{ij} is the number of times the words j,i appeared together in corpus, comes from co-occurrence matrix.

X_i is total number of times word i appeared in corpus.

Now, to achieve the goal using global statistics, compute the word vectors or we can say to calculate the prediction of co-occurrence ratio can be represented as P_{ik} / P_{jk}, we can represent it with some function F. for this we need to convert the multi-dimensional vectors to some scalar.

But here we got three element i,j,k. To create a function F with 3 elements is hard, so it is required to reduce to 2 elements. So Authors tries to solve these problems through this model.

$$F(w_i - w_j, \tilde{w_k}) \approx P_{ik} / P_{jk}$$

to create a linear relation between $w_i - w_j$ and $\tilde{w_k}$ authors used dot product which becomes like

$$F((w_i - w_j)^T \tilde{w_k}) \approx P_{ik} / P_{jk}$$

Now, authors used couple of tricks, by considering the log of right-hand side probability ratio, converted into subtraction between probabilities. $P_{ik} / P_{jk} = \log(P_{ik}) - \log(P_{jk})$

$$F((w_i - w_j)^T \tilde{w_k}) \approx \log(P_{ik}) - \log(P_{jk})$$

Added a bias term for each word, to the left to balance the log conversion loss in some words which appear more often.

$$(w_i - w_j)^T \tilde{w_k} + b_i - b_j = \log(P_{ik}) - \log(P_{jk})$$

For single entry in co-occurrence matrix equation become like

$$(w_i^T \tilde{w_k}) + b_i = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$$

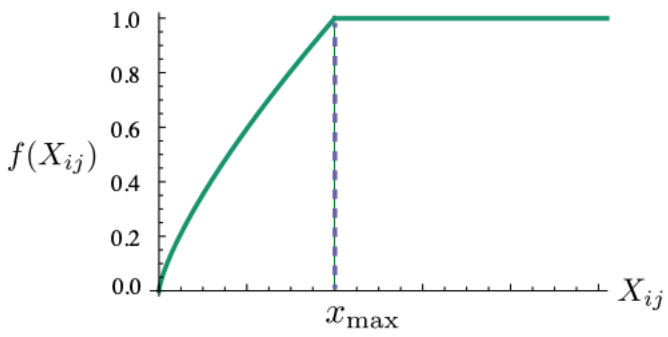
By absorbing the final term on the right-hand side into the bias term, and adding an output bias $\tilde{b_k}$ for $\tilde{w_k}$ restores the symmetry

$$(w_i^T \tilde{w_k}) + b_i + \tilde{b_k} = \log(X_{ik})$$

As the equation changed a lot from initially mentioned, still it has a problem with zero X_{ik} values. As a solution authors included addictive shift $\log(X_{ik})$ becomes $\log(1 + X_{ik})$.

Still it has a problem that it still weighs all co-occurrences equally even for rare ones. Which are noisy and less helpful. And authors proposed a new cost function

$f(x) = (x/x_{max})^{\alpha}$ if $x < x_{max}$, works well $\alpha = ¾$



Adding the cost function to the derived model function becomes.

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

In this paper author also compares the derived Glo Ve model with word2Vec reveals some interesting points that both derived into similar mathematical formulae with only differs on loss measurement. Glo Ve uses log mean squared errors between and actual and predicted probabilities where as word2vec uses cross-entropy loss over normalized probabilities. Both requires co-occurrence matrix pre-computed. Size of computed matrix can proportionate to non-zero matrix which will be very small when compared to large corpus sized matrix of word2Vec, so computation in this model should be simpler than the Word2Vec model.

In this paper, authors provided many comparable results by conducting experiments on the word analogy task and other tasks comparing results on various datasets and benchmarks from previous models. Especially on word analogy task the GloVe model perform better than other models with smaller size of corpora but shares the similar results with the Word2Vec model. Though the model shows similar results and mathematically looks like a Word2Vec model, but it derived in a better systematic approach in handling word embeddings.

References:

https://arxiv.org/pdf/1904.08067.pdf
https://nlp.stanford.edu/pubs/glove.pdf
https://www.youtube.com/watch?v=ASn7ExxLZws&t=4234s