A Quick Summary: GloVe: Global Vectors for Word Representation

Original Paper: https://nlp.stanford.edu/pubs/glove.pdf

2 March 2019

1 Ideas:

(a) An alternative formulation for continuous representations of word embeddings based on word counts.

2 Explanations:

(a) The formulation for this is surprisingly elegant. Let X be the matrix of word-word-co-occurrence counts, whose entries X_{ij} are the number of times word j occurs in the context of word i. Let $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ be the probability that word j appears in the context of word i. For a word k related to word i but not to j, the ratio $\frac{P_{ik}}{P_{jk}}$ would be expected to be large.

We then see how the formulation occurs, from a most general approach that converges on the proposed model:

i The most general model takes the form

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

ii Now, if we want to take into account the difference of the two target words, we have:

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

iii To explicitly preserve the linear structure, we write:

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

iv It would be elegant if F were to be a homomorphism between $(\mathbb{R}, +)$ and $(\mathbb{R}_{>0}, \times)$:

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

I This implies that

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

II And that a possible solution for F is

$$F(y) = \exp(y)$$

v Thus we have

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

vi To preserve symmetry, we absorb $log(X_i)$ into a bias b_i, and we introduce another bias $\tilde{b}_k$ to restore the symmetry:

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

vii Thus, we obtain the objective function

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

where f is a weighing function.

3 Results:

(a) Outperforms other models on word analogy, word similarity, and named entity recognition tasks.

4 Notes:

(a) This model seems quite elegant, and it's one of the reasons why it does seem quite attractive.

(b) Is there a way to build on this so as to take into account the ordering instead of just the context counts? I do think that the symmetry of the formulation would not be able to be preserved in this case.