GloVe

Global Vectors for Word Representation [1]

Machine Learning Assignment

Name: Mahim Mahbub

Student ID: 1505022

What is it ?

- Unsupervised Learning Algorithm
  - A log-bilinear model with a weighted least-squares objective
- Obtains word embeddings
  - Numerical/vectorized representations of words
- Maintaining word similarities in the corpus
  - “King” - “Man” = “Queen” - “Woman”
  - “Bangladesh” - “Dhaka” = “Saudi Arabia” - “Riyadh”

How does it work ?

- Formulates a co-occurrence matrix of word pairs in the corpus.

- Intuition: Ratios of word-word co-occurrence probabilities may have some encoded meaning.
  - **Context (eg. a 5 word window to left & a 5 word window to the right):**
    - Cristiano Ronaldo is the best **footballer** in the history of mankind
    - 5 words left & right of **footballer** represents its context

  - Let P_{ij} be the probability that word “j” is in the context of word “i”
  - The ratio P_{ik} / P_{jk} has meaning
  - $P_{water, liquid} / P_{football, liquid}$ => high; water and liquid are extremely related, football & liquid unrelated.
  - $P_{water, goal} / P_{football, goal}$ => low; water & goal unrelated, football and goal are extremely related.
  - $P_{water,human} / P_{football, human}$ => 1; human is similarly related to both football and water (~ Probably !!)

How does it work - Cont^{d}

- $F(w_i , w_j , \tilde{w_k}) = P_{ik} / P_{jk}$ Objective Function “F” is to be obtained.
- Formulated to a least squares regression problem

$$J = \sum_{i,j=1}^{V} (w_i^T \bar{w}_j + b_i + \bar{b_j} - \log X_{ij})^2$$

  - A weighted function f(x) included where $f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases}$
  - Reduces weights on rare and frequent pairs

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

- Learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence

Training

- AdaGrad Optimizer
- $x_{max} = 100$
- $\alpha = ¾$
- Initial Learning Rate = 0.05
- 50 iterations for $|v| < 300$
- 100 iterations for $|v| > 300$
- Context: 10 words to left and 10 words to right.

References

1. Jeffrey Pennington, Richard Socher, and Christopher D. Manning 2014. GloVe: Global Vectors for Word Representation. In Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).