Paper: *GloVe: Global Vectors for Word Representation*

Razwan Ahmed Tanvir

Date: 01/23/2022

Quote *Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus*

Overview Various methods of representing word vectors were successful in terms of semantic and syntactic meanings. However, the origin of these regularities in the meanings are not yet clear and how the models show these regularities is yet to unleash. The paper in discussion, from the authors of Stanford University, proposed a model that generalizes the model parameters to have these regularities in word embedding models. Primary goal of this paper was to figure out the necessary parameters for a model to show semantic and syntactic meanings for word vectors. In this paper, two approaches were taken to take advantages, one is, Global Matrix Factorization and the other one is Local Context Window. This model generates a sub-structure which is meaningful and this result is validated by the output of analogy tasks.

This paper produced some mentionable output. GloVe uses a global log-bilinear regression model that generates meaningful vector representations and also performs well on analogy tasks. Moreover, GloVe trains on nonzero elements in a word-word co-occurrence. In addition this model makes use of the statistical information of the whole corpus. GloVe outperforms state-of-the-art word similarity models.

Intellectual Merit The previous models for representing word vectors did not delve into why these semantic meanings exist in their model and how these regularities learned by the model. This paper aims to answer this and outputs the parameters necessary for a word embedding model. In terms of originality, this research stands out with its novel approach of using Global Matrix Factorization as an instance, Latent Space Analysis (LSA) and Local Context Window model namely, Skip-Gram model. In this paper, author claimed an accuracy of 75% in word analogy task. Moreover, the model performs better with respect to other recent models on Named Entity Recognition and similarity task. The authors are from Stanford and they are qualified individual to carry this research. Furthermore, they had access to a much larger corpus to train their model, which qualifies for the robustness of the outcome of the result. The authors used data from Wiki dump, Gigaword 5 and web crawl data.

Broader Impact This research unleashes the parameters necessary for a word embedding model to have semantic and syntactic meaning in the represented vector space. This research is widely used for exploring the model and it gets referenced frequently in the research community. This research were carried out by the academics from the Stanford which include a post-doc researcher. The authors of this paper released their source code on Github for open access and the datasets used by this research are also available online. However, all the authors of this research were male.

Keywords Natural Language Processing, Word Embedding, word-word-co-occurrence matrix

Discussion Questions

- Although the authors delve into various aspect of this research, and the paper shows prominent output, but authors did not discuss about potential bias in their model. By bias, I mean there could be some word semantics that do not comply with the moral values of the society that we want to see.
- The paper mentions about a specific weighted least squares model which is not clear to me and I want to discuss about this model and its underlying methods.

Table 1: Grade deductions by section

Overview	Intellectual M.	B. Impact	Keywords	Questions	Is Online?