World Academy of Science, Engineering and Technology
International Journal of Biotechnology and Bioengineering
Vol:16, No:07, 2022

Genomic Sequence Representation Learning: An Analysis of K-Mer Vector Embedding Dimensionality

Authors : James Jr. Mashiyane, Risuna Nkolele, Stephanie J. Müller, Gciniwe S. Dlamini, Rebone L. Meraba, Darlington S. Mapiye

Abstract : When performing language tasks in natural language processing (NLP), the dimensionality of word embeddings is chosen either ad-hoc or is calculated by optimizing the Pairwise Inner Product (PIP) loss. The PIP loss is a metric that measures the dissimilarity between word embeddings, and it is obtained through matrix perturbation theory by utilizing the unitary invariance of word embeddings. Unlike in natural language, in genomics, especially in genome sequence processing, unlike in natural language processing, there is no notion of a “word,” but rather, there are sequence substrings of length k called k-mers. K-mers sizes matter, and they vary depending on the goal of the task at hand. The dimensionality of word embeddings in NLP has been studied using the matrix perturbation theory and the PIP loss. In this paper, the sufficiency and reliability of applying word-embedding algorithms to various genomic sequence datasets are investigated to understand the relationship between the k-mer size and their embedding dimension. This is completed by studying the scaling capability of three embedding algorithms, namely Latent Semantic analysis (LSA), Word2Vec, and Global Vectors (GloVe), with respect to the k-mer size. Utilising the PIP loss as a metric to train embeddings on different datasets, we also show that Word2Vec outperforms LSA and GloVe in accurate computing embeddings as both the k-mer size and vocabulary increase. Finally, the shortcomings of natural language processing embedding algorithms in performing genomic tasks are discussed.

Keywords : word embeddings, k-mer embedding, dimensionality reduction
Conference Title : ICNNMLB 2022 : International Conference on Neural Networks and Machine Learning in Bioinformatics
Conference Location : Helsinki, Finland
Conference Dates : July 19-20, 2022