Distributed Representations of Sentences and Documents

Authors: Quoc Le, Tomas Mikolov
Publisher: International Conference on Machine Learning
Year: 2014

Keywords: Sentiment Analysis, Information Retrieval, Paragraph Vector, NLP

1. Abstract:

"Many machine learning algorithms require the input to be represented as a fixed-length feature vector. When it comes to texts, one of the most common fixed-length features is bag-of-words. Despite their popularity, bag-of-words features have two major weaknesses: they lose the ordering of the words and they also ignore semantics of the words. For example, "powerful," "strong" and "Paris" are equally distant. In this paper, we propose Paragraph Vector, an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. Our algorithm represents each document by a dense vector which is trained to predict words in the document. Its construction gives our algorithm the potential to overcome the weaknesses of bag-of-words models. Empirical results show that Paragraph Vectors outperform bag-of-words models as well as other techniques for text representations. Finally, we achieve new state-of-the-art results on several text classification and sentiment analysis tasks."

2. Introduction:

- The bag of words had been used for the sentence classification and the information retrieval applications, despite of it's disadvantages.
- To overcome those advantages, the new approach with the introduction of paragraph vector had been proposed by the authors.
- The process consists of concatenation of the paragraph vectors along with the word vectors in which paragraph vectors had been unique throughout the procedure and the word vectors had been common in different paragraphs.
- This process overcomes the cons of bag-of-words such as word order representation and the semantic orientation.

3. Related Works:

- The article's approach had been relevant to the word vector representation through neural networks accomplished by Bengio et al., 2006.
- The extension of that models resulted in producing the sentence and phrase level representations.

- The simple approaches such as averaging the weights of the vectors and the complex approaches such as creating parse trees had also been adopted in the applications of NLP. But those even had their limitations in word order.
- Those constraints of word order and the initial semantic orientation which had not been attained through the procedure had carried out by this process.

4. Content and Approach:

- Initially, the words are represented as vectors and the word matrix is created consisting all it's vectors indexed by the word number.
- This makes to predict the next word by averaging and taking the probability based on the previous words and the average.
- This prediction task is carried out by the softmax function. It is generally preferred to use hierarchical softmax than the normal one for the fast training.
- The hierarchical function used by the authors is the Huffman tree, in which the most frequent words were attached with short codes.
- These word vectors were trained by using the stochastic gradient descent acquired from the backpropogation algorithm.
- Later the training, the words which are alike were converged towards same position in vector space.
- As said earlier, the authors introduced a paragraph vector which is similar to the word vectors, but the paragraph vectors will be unique throughout the document, where as word vectors may be similar.
- The paragraph is represented by a vector and D represents the paragraph matrix consists of all paragraph vectors.
- These paragraph vectors can be used as the indices for the word vectors or memory to remember what does it misses in the context.
- Thus, we call this paragraph vectors as distributed memories(PV-DM). The paragraph vector is shared only through the same paragraph but not through all different paragraphs.
- While predicting, the paragraph vector is computed through inference which obtained by the gradient descent.
- The paragraph vectors can be represented as the features after the training. These features can be used for predicting labels through SVM, K-means and logistic regression algorithms.
- The other process to initialize the paragraph vector is to randomize through stochastic gradient descent and assigning the vectors. These doesn't have any ordering.
- This technique is called as the Paragraph vector's Distributed Bag-of-Words (PV-DBOW).

- The authors taken the combination of the two vectors (PV-DM) and (PV-DBOW) for the experimentation.
- The experiments were done on two applications: sentiment analysis and the information retrieval.
- For the sentiment analysis the SST and IMDB were taken. These refers the classification of movie reviews from negative to positive on a scale of 0-1, which in fact needed the fixed length vectors.
- For information retrieval they have taken triplets of paragraphs, in which two were similar orientation and the other differs from the other two.The error representation was based on the distance between the similar paragraphs and the distance between different ones.

5. Results:

- The bag of words with NB, SVM and RNNs had shown very poor results for the Stanford dataset with around 20% error rate, whereas the Paragraph vector had outperformed with 12.2% error rate.
- For the lengthy reports, Bag of Words performed very well with 12.2% error, however the paragraph vector had achieved best with 7.8%.
- For the third experiment of paragraph triplets also, the paragraph vector method had shown a significant results of 3.8% error rate compared to bag-of-words model with 8.10% error rate.
- Furthermore, to be noted PV-DM had performed very well compared to PV-DBOW. The former alone can achieve many results without the combination of PV-DBOW.
- The other problem which had to be noted is that the paragraph vector is expensive, which can be improved by training it, parallel to test-time.

6. Conclusion and Future Scope:

- The paragraph vectors which can be used for unsupervised learning with variable length inputs had been introduced and shown exceptional results compared to former approaches.
- The paragraph vector also overcome the problem of capturing semantics of the sentences and the paragraphs. In addition to that the vectors also protected the word order, which had been a problem with the bag-of-words models.
- The future works can be concentrated on the application of paragraph vectors in the non text domains such as computer vision etc., In addition to that the works to illustrate the texts can be focused with this approach.