Abstract

Multi-label Document Categorization, the task of automatically assigning a text document into one or more categories has various real-world applications such as categorizing news articles, tagging Web pages, maintaining medical patient records and organizing digital libraries among many others. Statistical Machine Learning approaches to document categorization have focused on multi-label learning algorithms such as Support Vector Machines, k-Nearest Neighbors, Logistic Regression, Neural Networks, Naive Bayes, Generative Probabilistic Models etc. while the input to such algorithms i.e. the vector representation for documents has traditionally been used as the bag-of-words model. Though the usage of simple bag-of-words representation gives surprisingly accurate results, it suffers from sparsity, high-dimensionality, lack of similarity measures along with other drawbacks such as the inability to encode word ordering and contextual information in which the words occur. Encoding contextual information about words in documents is crucial to capture the correct semantic content of the highly complex and ambiguous human language.

Our work is focused on learning continuous distributed vector representations for documents by embedding all the documents in the same low-dimensional space such that documents that are similar in their semantic content have similar vector representations. To tackle the issues in bag-of-words representation model, we present an unsupervised neural network model that uses the document vector to predict words in the document along with using the contextual information in which the word occurs and jointly learns distributed document and word representations. We use a modified version of the logistic regression algorithm to learn similar distributed representations for categories to perform the document categorization task. We show that our model gives state-of-the-art results on the standard *Reuters-21578* dataset, improving the bag-of-words model by 9% and previous state-of-the-art by 3.26% in terms of the F1 Score. We also show the effectiveness of our model in imputing missing categories on the Wikipedia articles against the bag-of-words representations. As we embed documents, categories and words in the same low-dimensional space our model can also estimate semantic similarities between them. We qualitatively demonstrate that the learned representations capture the semantic dependencies between categories and words which is not directly observed in the data.

Abstract

Multi-label Document Categorization, the task of automatically assigning a text document into one or more categories has various real-world applications such as categorizing news articles, tagging Web pages, maintaining medical patient records and organizing digital libraries among many others. Statistical Machine Learning approaches to document categorization have focused on multi-label learning algorithms such as Support Vector Machines, k-Nearest Neighbors, Logistic Regression, Neural Networks, Naive Bayes, Generative Probabilistic Models etc. while the input to such algorithms i.e. the vector representation for documents has traditionally been used as the bag-of-words model. Though the usage of simple bag-of-words representation gives surprisingly accurate results, it suffers from sparsity, high-dimensionality, lack of similarity measures along with other drawbacks such as the inability to encode word ordering and contextual information in which the words occur. Encoding contextual information about words in documents is crucial to capture the correct semantic content of the highly complex and ambiguous human language.

Our work is focused on learning continuous distributed vector representations for documents by embedding all the documents in the same low-dimensional space such that documents that are similar in their semantic content have similar vector representations. To tackle the issues in bag-of-words representation model, we present an unsupervised neural network model that uses the document vector to predict words in the document along with using the contextual information in which the word occurs and jointly learns distributed document and word representations. We use a modified version of the logistic regression algorithm to learn similar distributed representations for categories to perform the document categorization task. We show that our model gives state-of-the-art results on the standard *Reuters-21578* dataset, improving the bag-of-words model by 9% and previous state-of-the-art by 3.26% in terms of the F1 Score. We also show the effectiveness of our model in imputing missing categories on the Wikipedia articles against the bag-of-words representations. As we embed documents, categories and words in the same low-dimensional space our model can also estimate semantic similarities between them. We qualitatively demonstrate that the learned representations capture the semantic dependencies between categories and words which is not directly observed in the data.

Abstract

Multi-label Document Categorization, the task of automatically assigning a text document into one or more categories has various real-world applications such as categorizing news articles, tagging Web pages, maintaining medical patient records and organizing digital libraries among many others. Statistical Machine Learning approaches to document categorization have focused on multi-label learning algorithms such as Support Vector Machines, k-Nearest Neighbors, Logistic Regression, Neural Networks, Naive Bayes, Generative Probabilistic Models etc. while the input to such algorithms i.e. the vector representation for documents has traditionally been used as the bag-of-words model. Though the usage of simple bag-of-words representation gives surprisingly accurate results, it suffers from sparsity, high-dimensionality, lack of similarity measures along with other drawbacks such as the inability to encode word ordering and contextual information in which the words occur. Encoding contextual information about words in documents is crucial to capture the correct semantic content of the highly complex and ambiguous human language.

Our work is focused on learning continuous distributed vector representations for documents by embedding all the documents in the same low-dimensional space such that documents that are similar in their semantic content have similar vector representations. To tackle the issues in bag-of-words representation model, we present an unsupervised neural network model that uses the document vector to predict words in the document along with using the contextual information in which the word occurs and jointly learns distributed document and word representations. We use a modified version of the logistic regression algorithm to learn similar distributed representations for categories to perform the document categorization task. We show that our model gives state-of-the-art results on the standard *Reuters-21578* dataset, improving the bag-of-words model by 9% and previous state-of-the-art by 3.26% in terms of the F1 Score. We also show the effectiveness of our model in imputing missing categories on the Wikipedia articles against the bag-of-words representations. As we embed documents, categories and words in the same low-dimensional space our model can also estimate semantic similarities between them. We qualitatively demonstrate that the learned representations capture the semantic dependencies between categories and words which is not directly observed in the data.

Abstract

Multi-label Document Categorization, the task of automatically assigning a text document into one or more categories has various real-world applications such as categorizing news articles, tagging Web pages, maintaining medical patient records and organizing digital libraries among many others. Statistical Machine Learning approaches to document categorization have focused on multi-label learning algorithms such as Support Vector Machines, k-Nearest Neighbors, Logistic Regression, Neural Networks, Naive Bayes, Generative Probabilistic Models etc. while the input to such algorithms i.e. the vector representation for documents has traditionally been used as the bag-of-words model. Though the usage of simple bag-of-words representation gives surprisingly accurate results, it suffers from sparsity, high-dimensionality, lack of similarity measures along with other drawbacks such as the inability to encode word ordering and contextual information in which the words occur. Encoding contextual information about words in documents is crucial to capture the correct semantic content of the highly complex and ambiguous human language.

Our work is focused on learning continuous distributed vector representations for documents by embedding all the documents in the same low-dimensional space such that documents that are similar in their semantic content have similar vector representations. To tackle the issues in bag-of-words representation model, we present an unsupervised neural network model that uses the document vector to predict words in the document along with using the contextual information in which the word occurs and jointly learns distributed document and word representations. We use a modified version of the logistic regression algorithm to learn similar distributed representations for categories to perform the document categorization task. We show that our model gives state-of-the-art results on the standard *Reuters-21578* dataset, improving the bag-of-words model by 9% and previous state-of-the-art by 3.26% in terms of the F1 Score. We also show the effectiveness of our model in imputing missing categories on the Wikipedia articles against the bag-of-words representations. As we embed documents, categories and words in the same low-dimensional space our model can also estimate semantic similarities between them. We qualitatively demonstrate that the learned representations capture the semantic dependencies between categories and words which is not directly observed in the data.