International Journal of Grid and Distributed Computing
Vol. 9, No. 11 (2016), pp. 71-80
http://dx.doi.org/10.14257/ijgdc.2016.9.11.06

A Sense Embedding of Deep Convolutional Neural Networks for Sentiment Classification

Zhijian Cui[1], Xiaodong Shi[1*], Yidong Chen[1] and Yinmei Guo[2]

[1]*Department of Cognitive Science, Xiamen University, Xiamen, P.R. China*
[2]*Intellectual property Publishing House*
[1]*sgcuizhijian@163.com*, [1]*mandel@xmu.edu.cn*, [1]*ydchen@xmu.edu.cn*,
[2]*softsnow823@gmail.com*

Abstract

Sentiment classification task has attracted considerable interest as sentiment information is crucial for many natural language processing (NLP) applications.

The goal of sentiment classification is to predict the overall emotional polarity of a given text. Previous work has demonstrate the remarkable performance of Convolutional Neural Network (CNN). However, nearly all this work assumes a single word embedding for each word type, ignoring polysemy and thus inevitably casting negative impact on the downstream tasks. We extend the Skip-gram model to learn multiple sense embeddings for the word types, catering to introduce sense-based embeddings for CNN during sentiment classification. Instead of using the pipeline method to learn multiple sense embeddings of a word type, the sense discrimination and sense embedding learning for each word type are performed jointly based upon the semantics of its contextual words. We validate the effectiveness of the method on the commonly used datasets. Experiment results show that our method are able to improve the quality of sentiment classification when comparing with several competitive baselines.

Keywords: *Natural Language Processing; Text Classification; Sense Sensitive; Word Embedding*

1. Introduction

The task of sentiment classification aims to identify the opinions, attitudes and emotion with regard to the objects and entities. In recent years, with the continued growth of social media such as blogs and reviews, people are increasingly using public opinions in these media to seek the advice of decision making. This tendency brings about an extensive research in the areas of opinion mining and sentiment classification, hoping to develop systems that can automatically analyze and extract useful information important to the users.

A key problem in sentiment classification is how to learn effective feature representation. Conventional methods often take advantage of co-occurrence word-based model (e.g., unigram, bigram and trigram) or syntax clues to represent the classification features, which suffers from data sparsity problem, to some extent. To address this problem, feature selection (e.g., frequency, mutual information [1], topic model [2, 3]) have been received extensive discussions and comparisons. Nevertheless, the features selection applied in these methods are often time-consuming and computationally prohibit. What's more, most of the selected features are shallow features unsatisfied to capture the deeper semantics of the contexts.

Recently, word embedding has gained great success in many NLP tasks such as paraphrase detection [4] and parsing [5]. Word embedding is a dense, low-dimensional

* Corresponding Author

ISSN: 2005-4262 IJGDC

and real-valued vector. Word embedding uses a distributed representation to learn meaningful semantic regularities of words and greatly alleviates the data sparsity problem. Normally, the words occurred in similar contexts will have similar word embedding. With the help of word embedding as pre-trained vectors, many deep neural network based methods [4, 6, 7, 8, 9] have been proposed for the semantic composition of the texts with variable-sized. However, nearly all this work assumes a single word embedding for each word type, ignoring polysemy phenomenon which is pervasive in linguistics expressions. Hence, in this paper, we present a sense-based of deep CNN for sentiment classification.

Concretely, we consider multiple sense embeddings of the word types instead of single word embedding as input of the CNN. For learning sense embedding, our method is greatly inspired by the work [10] where each word type contains a global vector and multiple sense vectors. Normally, the global vector can be learned by toolkit Word2Vec [11]. Similar to the work [10], we use the global vector of the contexts to estimate the specific sense embedding of the center words. We use the hierarchical soft-max method [11] to compute the output layer effectively where each word type in the vocabulary is represented by a binary tree arranged by their frequency in the training dataset. Here, the specific sense of the center word is decided according to the average vector of the global vectors of its contexts, and the sense number is the hyper-parameter in our model. Using the already learned sense embedding, we use them as the input to feed forward the CNN for the task of sentiment classification. The framework of our proposed method is given in Figure 1.

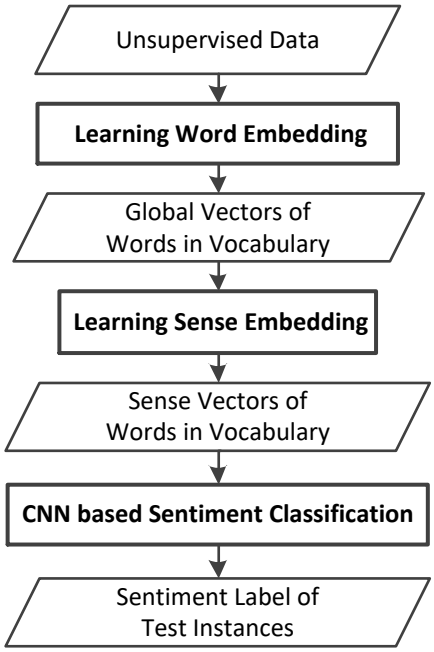

Figure 1. An Illustration of our Proposed Framework for Sentiment Classification

Our method is different from the work [10] in two aspects. First, the output layer of our method is hierarchical soft-max instead of introducing noisy sampling. Second, we stabilize the global vector of the contexts and only update the sense embedding in terms of its contexts during model training.

Overly, the main contributions of our work can be summarized in the following two aspects:

- We extend the Skip-gram model to learn multiple sense embeddings for the word types.

- In sentiment classification task, we use a specific sense embedding to feed forward the CNN rather than use the word embedding.

The remainder of this paper is organized as follows: Section 2 summarizes and compares related work. Section 3 presents our method on how to learn multiple senses embedding for each word type in the vocabulary. Also, the detailed discussion of the sentiment classification of the CNN will be given. Section 4 describes our experiments and shows results with discussions. Finally, Section 5 concludes and outlines future directions.

2. Related Work

In this section, we present a brief review with respect to the related work into two aspects: (1) sentiment analysis, and (2) learning distributed representations.

2.1. Sentiment Analysis

Sentiment analysis, broadly speaking, is the technique that allows to identify the emotional polarity within the text. Generally, the methods employed in sentiment analysis can be divided into two categories: (1) sentiment knowledge based methods, and (2) classified features based methods.

The sentiment knowledge based methods [12, 13, 14, 15] mainly resort to pre-defined sentiment lexicons where each sentiment word is manually annotated with a emotional polarity. With these sentiment words and a set of sentiment computation paradigms, then the sentiment of a sentence or document is induced directly. Despite their wider applications to sentiment analysis, sentiment lexicons is domain-dependent and often require excessive human investment.

The classified features based methods [12, 16, 17, 18] consider the sentiment analysis as a task of text classification based upon their emotional polarity. The main idea is to learn the representative features for a sentiment classifier during training stage. At the test stage, the learned classifier is applied directly to predict the emotional polarity of the test sentence or document. Overly, most of existing methods involve excessive feature engineering. These features are regarded as the gold standard to discriminate different emotional polarities. Hence, if they are not selected or defined properly, then the accuracy of the system will degrade sharply. A representative work is found in [19], which develops a sentiment classifier via a great number of hand-crafted features, and this classifier is able to achieve a promising performance in SemEval 2013 Twitter Sentiment Classification Task.

More recently, [20, 21] learns the distributed representation for sentiment analysis, which jointly incorporates the syntactic context and sentiment polarity. The distributed representations of sentiment-specific words are proved to be informative in sentiment analysis. In fact, many researchers attempt to model words, phrases and even the whole sentences with distributed semantic representations. Early works on using neural networks to learn phrase representations can be found in [22], which uses a recurrent neural network to learn dense real-value representations of the phrases. More recently, RAE based methods have been developed in many NLP tasks [4, 6, 7, 8, 23].

In addition, the cross-domain sentiment classification [24, 25] and cross-lingual sentiment classification [26, 27, 28] are also received tremendous attentions.

2.2. Learning Distributed Representation

Conventional methods such as [12] use a one-hot vector to represent each word. The length of such vector is the size of the vocabulary, and only one dimension in the vector are set to 1 while the rest are all set to be 0. However, the one-hot vector contains severe data sparsity issue and somehow unable to capture complex linguistics knowledge.

According to the deep learning framework [29], distributed semantic representations of words are learned in an unsupervised fashion, capable of providing a simple and generalized feature to enhance existing supervised NLP tasks. [25] takes advantage of Stacked Denoising Autoencoders for sentiment classification. Socher et al. develops Recursive Neural Network [4, 5], Recursive Vector-Matrix [7] and Recursive Neural Tensor Network [8] to learn representations of the phrases or even the whole sentence recursively. Recursive deep learning can jointly learn distributed semantic space representations of words and hierarchical syntactic structures within the contexts in an unsupervised fashion, allowing to capturing complex linguistic phenomenon during context modeling.

Unlike previous method which uses one single word embedding and thus inevitably casting negative impact on the sentiment classification, our method allows to obtain multiple sense embedding for CNN during sentiment classification, such that the polysemy phenomenon can be discriminated more effectively.

3. Our Method

In this section, we will describe how to learn multiple senses embedding for each word type in the vocabulary. Also, the detailed discussion of the sentiment classification of the CNN will be given.

Learning Sense Embedding As previous section points out, the sense embedding of the word types are obtained by its surrounding contexts. In practice, we use the Skip-gram model and hierarchical soft-max in toolkit Word2Vec [11] to learn the global vectors. Figure 2 gives the basic idea of our method.

In Figure 2, $V_g(context_{1\sim5})$ are the global embedding of contexts surrounding the center words. As for the center word, we predefine the context size N to be 5 and its sense number K to be 3, each of which is attached with a cluster vector ($V(c_i)$) and a sense embedding ($V(s_i)$). For example, the $V(c_1)$ and $V(s_1)$ are the cluster vector and sense embedding of sense 1, respectively.

In our method, for the first step, based upon the average of the global vector of the contexts, the most possible cluster vector will be selected in accordance with the similarity between the average vector and the cluster vectors (the $V(c_3)$ in our case labeled with green line). Here, the label of the most possible cluster will be decided as follow:

$$c_i = \arg\max_{i=1\sim K}\{sim(V(c_i), average)\} \tag{1}$$

$$sim(x, y) = \cos(x, y) = \frac{\langle x, y\rangle}{\|x\|\|y\|} \tag{2}$$

$$average = \frac{1}{N}\sum_{i=1}^{N}\left(V_g\left(context_i\right)\right) \tag{3}$$

For the second step, given the cluster label, we update its corresponding sense embedding (the $V(s_3)$ in our case labeled with green line) by the hierarchical soft-max method. Intuitively, the more a word occurs in the context of the specific sense, the more effect to the specific sense this word will be. Here, each word type in the hierarchical binary tree is arranged in terms their corresponding frequency. In our method, the update strategy of the specific sense embedding based upon the hierarchical soft-max is similar to the one in Word2Vec [11].

CNN based Sentiment Classification We uses the CNN based method to conduct the sentiment classification. Our method is similar to the work [9]. The biggest difference

is that the input of our CNN framework is sense sensitive. Concretely, for a word in the sentence, we use its specific sense embedding based upon the co-occurring words, regarding as the projection from a word into a distributed semantic representation. For example, if the sense number is 3, then we will have 3 senses embeddings for each word type. And the specific sense embedding will be selected according to its contexts to initialize the input of CNN. In contrast, [9] only projects a word to a single word embedding (sense insensitive).

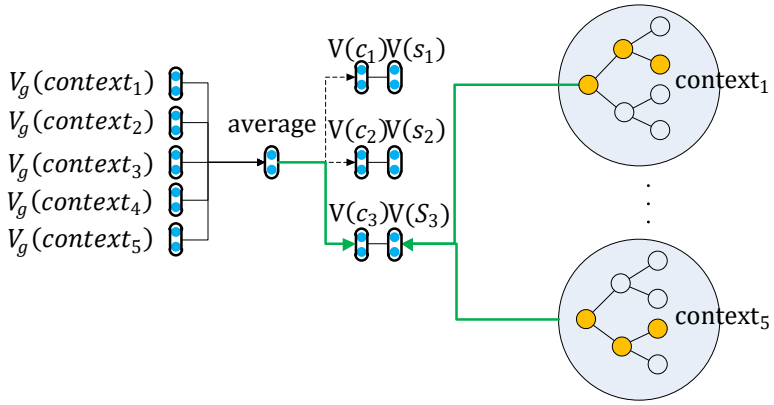

Figure 2. An Illustration of our Proposed Framework for Learning Sense Embedding

By doing so, we can have an input matrix where each row represents one word in the sentence and its corresponding column is the distributed representation. Followed by the method [9], we totally use 100 filter maps for each region size (2, 3 and 4). After this, we use 1-max pooling strategy to select the most important features from the output of the filters.

Finally, we are able to obtain the most representative sentiment features for each training instance. The object of our CNN framework is cross-entropy loss with imposing the regularization constraint. Using the gradient descent, we update the parameters iteratively and thus the optimal parameters can be learned automatically and effectively.

4. Experiment

In this section, we are going to compare our proposed method to the baseline systems.

4.1. Setup

Dataset We evaluated our method on the sentence polarity dataset [30] which will be denoted as the SPD in the following section. We tuned the regularization hyper-parameters via cross-fold validation, optimizing for accuracy. We show the average results from 10-folds over SPD. The detailed information of SPD is given in Table 1.

Table 1. Data Descriptions of our Sentiment Analysis

Sentiment	Instance	Average Length	Maximum Length
Positive	5331	21	59
Negative	5331	20	56

Baseline System As previous section stated, we extend the skip-gram model to learn the multiple sense embedding rather than use a single word embedding for each word type. Hence, we use the word embedding learned from toolkit Word2Vec [11] to initialize the CNN classifier as the baseline systems, denoted as Non-Static-Word2Vec-CNN and Static-Word2Vec-CNN. Here, "Non-Static" means the word embedding will be tuned for

sentiment classification. In contrast, "Static" means the word embedding will not be updated. The detailed parameter configurations are described in Table 2.

4.2. Result

In the sentiment evaluation, we use the multiple senses embedding to initialize our CNN sentiment classification. We would like to testify whether the multiple sense embedding is superior to the CNN model initialized with word embedding which is unable to discriminate different senses for polysemous words. Notice that the performance of our baseline (Static-Word2Vec-CNN and Non-Static-Word2Vec-CNN) is not as good as the results reported in the work [9] using available Google word embedding. The reason is that the size of our training dataset for word embedding is limited which may affect the coverage and learning accuracy of word embedding.

Table 2. Data Descriptions of our Sentiment Analysis

Parameter	Value
window	8
sample	1e-4
skip gram	True
hierarchical softmax	True
vector size	300
filter region size	(2,3,4)
pooling	1-max pooling
dropout rate	0.5
convolution active function	ReLU
sense number	2

Table 3. Experiment Results on the Test Set

Method	Accuracy(%)
Static-Word2Vec-CNN	75.7
Static-OurMethod-CNN	75.91
Non-Static-Word2Vec-CNN	78.09
Non-Static-OurMethod-CNN	78.32

In Table 3, the performance of our method ``Non-Static-OurMethod-CNN'' achieves the best performance which demonstrates the effectiveness of our proposed method. It's worth mentioning that the "Non-Static" tunes the parameters based on the supervised learning so that the embedding tends to be sentiment-dependent. This is the reason why it achieves better result than the "Static" one.

5. Conclusion

We have proposed a sense-based of deep convolutional neural networks for sentiment classification. Instead of using the word embedding to initialize the input of the CNN model, we learn multiple senses embedding for each word type so that different senses of the polysemous words can be discriminated directly. In our experiment, we have compared the ``Static'' and "Non-Static" methods and learn that the latter one is better than the former one. The experiment results demonstrate the effectiveness of our method.

Acknowledgments

We would like to thank all the referees for their constructive and helpful suggestions on this paper. This work is supported by the Natural Science Foundation of China (Grant

No. 61005052, Grant No. 61075058 and Grant No. 61303082), the Key Technologies R\&D Program of China (Grant No. 2012BAH14F03), the Fundamental Research Funds for the Central Universities (Grant No. 2010121068), the Natural Science Foundation of Fujian Province, China (Grant No. 2010J01351), the Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20120121120046) and the Ph.D. Programs Foundation of Ministry of Education of China (Grant No. 20130121110040).

References

[1] Thomas M Cover and Joy A Thomas, "Elements of information theory", John Wiley & Sons, **(2012)**.
[2] Lijuan Cai and Thomas Hofmann, "Text categorization by boosting automatically extracted concepts", Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, **(2003)**, pp. 182-189.
[3] Swapnil Hingmire, Sandeep Chougule, Girish K Palshikar, and Sutanu Chakraborti, "Document classification by topic labeling", Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM, **(2013)**, pp. 877-880.
[4] Richard Socher, Eric H Huang, Jerey Pennin, Christopher D Manning, and Andrew Y Ng, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection, Advances in Neural Information Processing Systems, vol. 24, **(2011)**, pages 801-809.
[5] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng, "Parsing natural scenes and natural language with recursive neural networks", Proceedings of the 28th international conference on machine learning (ICML-11), **(2011)**, pp. 129-136.
[6] Richard Socher, Jerey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, **(2011)**, pp. 151-161.
[7] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng, "Semantic compositionality through recursive matrix-vector spaces", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, **(2012)**, pp. 1201-1211.
[8] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts, "Recursive deep models for semantic compositionality over a sentiment Treebank", Proceedings of the conference on empirical methods in natural language processing (EMNLP), Citeseer, **(2013)**, pp. 1631- 1642.
[9] Ye Zhang and Byron Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classication". arXiv preprint arXiv:1510.03820, **(2015)**.
[10] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum, "Efficient non-parametric estimation of multiple embeddings per word in vector space", arXiv preprint arXiv:1504.06654, **(2015)**.
[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Je Dean, "Distributed representations of words and phrases and their compositionality", Advances in Neural Information Processing Systems, **(2013)**, pp. 3111-3119.
[12] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics, **(2002)**, pp. 79-86.
[13] Xiaowen Ding, Bing Liu, and Philip S Yu, "A holistic lexicon-based approach to opinion mining", Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, **(2008)**, pp. 231-240.
[14] Maite Taboada, Julian Brooke, Milan To_loski, Kimberly Voll, and Manfred Stede, "Lexicon-based methods for sentiment analysis", Computational linguistics, vol. 37, **(2011)**, 267-307.
[15] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou, "Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology", vol. 63, **(2012)**, pp. 163-173.
[16] Alexander Pak and Patrick Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining", LREC, vol. 10, **(2010)**, pp. 1320-1326.
[17] Dmitry Davidov, Oren Tsur, and Ari Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys", Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, **(2010)**, pp. 241-249.
[18] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao, "Target-dependent twitter sentiment classification", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, **(2011)**, pp. 151-160.

[19] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets", Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR.13), **(2013)**.
[20] Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou, "Coooolll: A deep learning system for twitter sentiment classification", SemEval 2014, **(2014)**, pp. 208-212.
[21] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin, "Learning sentiment-specific word embedding for twitter sentiment classification", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, **(2014)**, pp. 1555-1565.
[22] Jerey L Elman, "Distributed representations, simple recurrent networks, and grammatical structure", Machine learning, vol. 7, **(1991)**, pp. 195-225.
[23] Peng Li, Yang Liu, and Maosong Sun, "Recursive autoencoders for itg-based translation", Proceedings of the Conference on Empirical Methods in Natural Language Processing, **(2013)**, pp. 567-577.
[24] Danushka Bollegala, David Weir, and John Carroll, "Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification", Proceedings of the 49th Annual Meeting of the Association, **(2011)**, pp. 132-141.
[25] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Domain adaptation for large-scale sentiment classi_cation: A deep learning approach", Proceedings of the 28th International Conference on Machine Learning (ICML-11), **(2011)**, pp. 513-520.
[26] Xiaojun Wan, "Co-training for cross-lingual sentiment classification", Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Association for Computational Linguistics, **(2009)**, pp. 235-243.
[27] Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K Tsou, "Joint bilingual sentiment classification with unlabeled parallel corpora", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, **(2011)**, pp. 320-330.
[28] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang, "Cross-lingual mixture model for sentiment classification", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, **(2012)**, pp. 572-581.
[29] Yoshua Bengio, "Deep learning of representations: Looking forward. In Statistical language and speech processing", Springer, **(2013)**, pp. 1-37.
[30] Bo Pang and Lillian Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, **(2005)**, pp. 115-124.

Authors

Zhijian Cui, he currently is a Ph.D. candidate at the School of Information Science and Engineering, Xiamen University. His research interests include natural language process and statistical machine translation.

Xiaodong Shi, he received the Ph.D. degree in computer software from National University of Defense Technology, Changsha, China, in 1994. He is now a professor in the Cognitive Science Department of Xiamen University. His research interests include natural language processing and artificial intelligence.

Yidong Chen, he received his Ph.D. degree in mathematics from Xiamen University, Xiamen, China, in 2008. He is now an associate professor in the Cognitive Science Department of Xiamen University. His research interests include statistical machine translation and semantic analysis.

Yinmei Guo, she has received the master degree at the School of Information Science and Engineering, Xiamen University. Her research interests include natural language process and statistical machine translation.

