Convolutional Neural Networks for Multimedia Sentiment Analysis

Guoyong Cai(✉) and Binbin Xia

Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China
{ccgycai,beybinxia}@gmail.com

Abstract. Recently, user generated multimedia contents (e.g. text, image, speech and video) on social media are increasingly used to share their experiences and emotions, for example, a tweet usually contains both texts and images. Compared to sentiment analysis of texts and images separately, the combination of text and image may reveal tweet sentiment more adequately. Motivated by this rationale, we propose a method based on convolutional neural networks (CNN) for multimedia (tweets consist of text and image) sentiment analysis. Two individual CNN architectures are used for learning textual features and visual features, which can be combined as input of another CNN architecture for exploiting the internal relation between text and image. Experimental results on two real-world datasets demonstrate that the proposed method achieves effective performance on multimedia sentiment analysis by capturing the combined information of texts and images.

Keywords: Multimedia · Sentiment analysis · Convolutional Neural Networks · Deep learning

1 Introduction

Online social networks are providing multiple forms of access to their users, for instance, people can post a tweet attached with images or videos. Social networks sites play an important role in people's live for requiring information and sharing experiences. Meanwhile, online users love to express their opinions on subjects they interested in, because of the free expression of speech on social networks. Sentiment analysis of online user generated data on social networks can be helpful to understand user behavior and improve applications aimed at online users. Among the large amount of data, we are particularly interested in analyzing sentiment of tweets containing both texts and images towards specific events and topics.

Deep neural networks have achieved remarkable performance in many fields, especially in compute vision [1, 2, 3, 4] and speech recognition [5] in recent years. In the field of natural language processing (NLP), works with deep learning methods were also widely used. As a challenging task NLP, sentiment analysis has been studied in various ways. Inspired by the enormous successes of deep learning, much research work on sentiment analysis has applied deep learning algorithms. However, most of

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 159–167, 2015.
DOI: 10.1007/978-3-319-25207-0_14

them are mainly focused on one single form of user content (such as text, image or video) separately instead of the combined representation. In fact, a great number of images posted do not contain any sentiment words in text at all, or the text sentiment is obvious but the image sentiment is unconspicuous. Figure 1 shows two tweets consisting of both text and image, their emotions can only be classified obviously through the combined text and image.

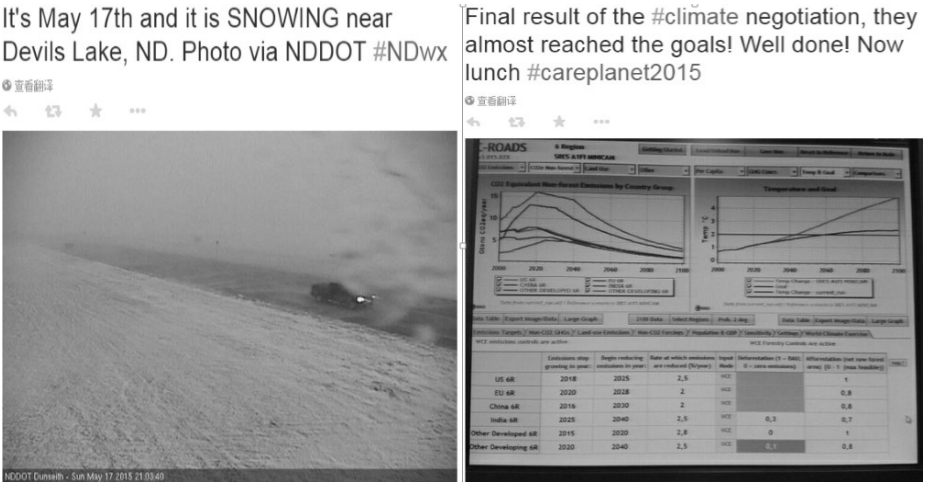

Fig. 1. Examples of two image tweets, text sentiment of the left one and image sentiment of the right one are not obvious. But image sentiment of the left one and text sentiment of the right one are relatively obvious.

In this paper, we focus on the problem of sentiment prediction based on the joint textual and visual information within an image post. Convolutional Neural Networks employed in prior works [6, 7, 8, 9] have been proved very powerful in solving image or text sentiment analysis tasks. Thus, to solving the challenging problem mentioned above, a novel deep learning architecture based on CNN was proposed. We intend to find out whether applying CNN to the joint information of text and image provides better performance than classifiers using only single form of information (either text or image).

The rest of the paper is organized as follows. In Section 2, we review research work focusing on sentiment analysis. Next we describe the proposed sentiment prediction based on CNN architecture in Section 3. Then we present datasets and experimental results in Section 4. Finally, some conclusions and future work are given in Section 5.

2 Related Work

In recent years, compared to traditional sentiment analysis on text, sentiment analysis of visual content has also attracted much attention, especially prominent performance has been witnessed on image classification based on deep learning algorithms [3, 23, 25]. In this section, we review research work closely related to our study focusing on textual and visual sentiment analysis.

2.1 Textual Sentiment Analysis

Sentiment analysis of text has been a challenging and fascinating task since it is proposed, and researchers have developed different approaches to solve this problem. Generally, two main approaches can be distinguished: dictionary based method and machine learning method.

Dictionary based method for sentiment analysis usually depends on the pre-defined sentiment dictionaries. Turney [10] presented a simple unsupervised learning algorithm for classifying users' reviews by leveraging the average semantic orientation score of the phrases, which is calculated by mutual information measures.

For machine learning approaches, Pang et al. [11] took n-gram and POS as features to classify movie reviews with Naive Bayes, maximum entropy classification, and support vector machines. As a sub-field of machine learning, deep learning methods achieved tremendous success, which motivate researchers to employ different kinds of deep learning methods for textual sentiment analysis. Socher et al. [12] proposed a semi-supervised approach based on recursive autoencoders for predicting sentiment distributions. Kim [6] and dos Santos et al. [7] developed deep convolutional neural network built on top of word2vec for performing textual sentiment analysis.

2.2 Visual Sentiment Analysis

In contrast to textual sentiment analysis, research work focusing on sentiment prediction of visual content falls far behind. Previous researches on visual sentiment analysis have mostly been conducted by utilizing low-level image features [13, 14, 26] or mid-level image attributes [15, 16]. Jia et al. [13] developed a semi-supervised framework based on factor graph model, which takes advantage of color features and social correlation among images. Yang et al. [14] proposed a novel emotion learning method to exploit social effect correlate with the emotion of images. The method jointly modeled images posted by social users and comments added by friends. Yuan et al. [15] proposed an image sentiment prediction framework based on mid-level attributes which were generated from four general scene descriptors. Borth et al. [16] constructed a large-scale Visual Sentiment Ontology and a novel visual concept detector library to visual sentiment prediction.

In addition, several researches employed deep learning methods [6, 7] for visual sentiment analysis. Xu et al. [6] proposed a novel visual sentiment prediction framework with CNN. The framework performs transfer learning from a CNN [4] with millions of parameters, which is pre-trained on large-scale data for object recognition. In order to solve challenging problem of image sentiment, You et al. [7] proposed a progressive CNN, which a probabilistic sampling algorithm was employed to select the new training subset, namely removing instances with similar sentiment scores for both classes with a high probability.

2.3 Multimedia Sentiment Analysis

To our best knowledge, there are few works focusing on sentiment analysis of combined textual and visual content. Borth et al. [16] compare the performance on twitter dataset using text only (SentiStrength), visual only (SentiBank), and their combination. The experimental results show that visual content predicted by the SentiBank-based classifier plays a much more important role in predicting the overall sentiment of the tweet, and the combined classifier achieve best performance. Wang et al. [17] propose a novel Cross-media bag-of-words Model (CBM) for Microblog sentiment analysis. The model represent the text and image of a tweet as a unified bag-of-words features, which are taken as input of machine learning methods (i.e., NB, SVM and Logistic Regression).

3 Textual and Visual Sentiment Analysis with CNN

Previous works have proven the powerful performance of CNN for textual [6, 7] and visual [8, 9] sentiment analysis. In this section, we introduce a comprehensive framework for joint sentiment analysis with CNN. As shown in Figure 2, the overall architecture of the proposed framework consists of three components: text CNN, image CNN and multi CNN, and each of the three components is a CNN architecture. Multi CNN takes joint text-level and image-level representation as input, and two kinds of representation are respectively extracted by vectorizing the features in the penultimate layer of text CNN and image CNN.

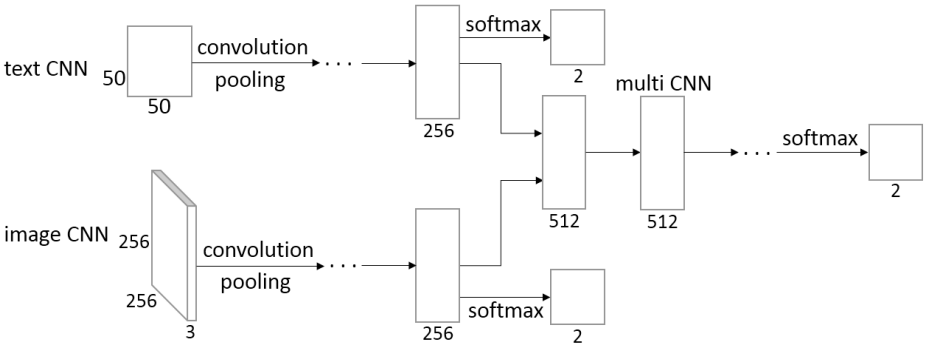

Fig. 2. The overall architecture of the proposed multimedia sentiment prediction framework

3.1 Textual Sentiment Analysis with CNN

We develop the text CNN for textual sentiment analysis to generate text-level representation. Pre-trained word vectors are used to initialize the word representations, which are taken as input of the text CNN. Detailed process of learning pre-trained word vectors will be discussed in Section 4. The overall architecture of the text CNN consists of three convolutional layers, two full connected layers and one softmax layer. Each convolutional layer is connected to a max pooling layer. Detailed information of text CNN is described as follows.

The first convolutional layer filters the word representations with 16 kernels of size 5*5, the second convolutional layer takes the pooled output of the first convolutional layer as input with 32 kernels of size 4*4. Pooled output of the second convolutional layer is connected to the third convolutional layer with 64 kernels of size 3*3. The last max pooling layer is followed by two full connected layers, and each of them has the same amount of neurons. The last softmax layer is used to classify the output of last full connected layer over two class labels. The text-level representation v_{text} is computed as follows:

$$v_{text} = f(\mathbf{w} \cdot (CNN_{text}(T)) + b) \tag{1}$$

Where f denotes the activation function, CNN_{text} is the text CNN, $\mathbf{w}$ is the weight matrix and b is a bias term. Thus, each text T can be represented as a fixed dimension vector v_{text}.

3.2 Visual Sentiment Analysis with CNN

Similar to the text CNN, the image CNN is developed for visual sentiment analysis and generating image-level representation. The image CNN is composed of five convolutional layers, three full connected layers and one softmax layer. The input images for the first convolutional layer is resized to the same size (256*256*3). Details of image CNN is discussed as follows.

The number of kernels of each convolutional layer is same with [2], and the corresponding size of kernels is respectively 17*17, 13*13, 7*7, 5*5 and 3*3. The output of the last full connected layer is taken as input for softmax layer. The formula of computing image-level representation v_{image} is similar with Equation (1).

3.3 Multimedia Sentiment Analysis with CNN

Aiming at solving the problem of multimedia sentiment analysis, we develop the multi CNN to take the joint text-level and image-level representation as input. The multi CNN does not contain any convolutional layer and max pooling layer at all, it just consists of four full connected layer and one softmax layer. The multi CNN is described in detail as follows.

The input features are mapped by four full connected layer, and the output features are passed to a softmax layer, which produces a distribution over two class (positive or negative) labels.

3.4 Classification

As described above, text-level and image-level representation are both in vector form, which can be taken as features for linear classifiers, such as Naïve Bayes, SVM and Logistic Regression. The experiment results of Borth et al. [16] show that Logistic Regression achieves better performance than SVM for visual sentiment prediction, and Logistic Regression is also employed in Xu et al. [8]. In Section 4, we employ Logistic

Regression as classifier with the vectorized features in the penultimate layer of text CNN, image CNN and multi CNN.

4 Experimental Setup and Results

4.1 Datasets

In our work, training dataset is constructed with randomly chosen 20K image posts (one image post consists of one image and corresponding description) from SentiBank [16], which consists of collected image posts on Flickr. The SentiBank are weakly labeled by 1200 adjective noun pairs (ANPs), which are based on psychological theory, Plutchik's Wheel of Emotions [24]. Similar to the work [9], we employ a probabilistic sampling algorithm to generate the new training dataset.

We evaluate the performance of proposed CNN architecture on two real-world twitter datasets, which have respectively been used in prior work [16, 9]. Both of two datasets are collected from image tweets, each of which contains text and corresponding image. The first twitter datasets (TD1) includes 470 positive tweets and 133 negative tweets, and the second one (TD2) includes 769 positive tweets and 500 negative tweets.

4.2 Pre-trained Word Vectors

In this work, word vectors initialized by skip-gram model [19], which has shown powerful performance in previous works. Word vectors are trained with word2vec tool on the latest English Wikipedia corpus, which is processed by removing paragraphs are not in English and sentences are less than 20 characters. The dimension of word vectors is set to 50 with a context window of size 5.

4.3 CNN Training

In our experiments, training is processed by stochastic gradient descent (SGD) with mini-batch size of 128 for optimization. Early-stopping [20] and dropout [21, 22] (with probability of 0.5) are employed for avoiding over-fitting. ReLU [21, 23] is adopted as activation function for text CNN, image CNN and multi CNN. Words are not in pre-trained word vectors are initialized randomly and the randomly initialized vectors are taken as parameters of networks, which will be fine-tuned in training process. In order to handle sentences of variable length, the maximum length of sentence is fixed to 50 for text CNN, zero vectors are padded if length is less than 50. The dimension of text-level and image-level representation are both set to 256. We implement our experiments for the proposed CNN architecture on Keras, which is an effective deep learning framework implementation.

4.4 Results

We compare the text CNN with Naïve Bayes, SVM and Logistic Regression for textual sentiment analysis. As for visual sentiment analysis, the image CNN is compared with low-level features [26], SentiBank [16] and Sentribute [15]. Since little works focus on multimedia sentiment analysis, we just take text CNN, image CNN, the combination of SentiStrength and SentiBank, SVM and Logistic Regression as comparative methods against multi CNN. Results of the proposed method on two twitter datasets can be respectively seen in table 1, table 2 and table 3. The experimental results show that the proposed multi CNN lead to better performance than other methods for multimedia sentiment analysis.

Table 1. Accuracy of algorithms on twitter datasets of text

Algorithms	TD1	TD2
NB	0.70	0.72
SVM	0.72	0.74
LR	0.73	0.76
Text CNN	0.74	0.77

Table 2. Accuracy of algorithms on twitter datasets of image

Algorithms	TD1	TD2
Low-level	0.710	0.664
SentiBank	0.709	0.662
Sentribute	0.738	0.696
Image CNN	0.773	0.723

Table 3. Accuracy of algorithms on twitter datasets

Algorithms	TD1	TD2
SentiStrength +SentiBank	0.72	0.723
SVM	0.76	0.781
LR	0.77	0.783
Multi CNN	0.78	0.796

5 Conclusions

In this paper, we propose a new CNN architecture that fully uses joint text-level and image-level representation to perform multimedia sentiment analysis. Based on idea of the complementary effect of the two representations as sentiment features, the proposed method exploits the internal relation between text and image in image tweets and achieves better performance in sentiment prediction. In future work, we would like to explore multimedia sentiment analysis with much more combination among text, image and other type of social media.

References

1. LeCun, Y., Boser, B., Denker, J.S., et al.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
2. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation **18**(7), 1527–1554 (2006)
3. Ciresan, D.C., Meier, U., Masci, J., et al.: Flexible, high performance convolutional neural networks for image classification. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence 22(1), p. 1237 (2011)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
5. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649. IEEE (2013)
6. Kim, Y.: Convolutional neural networks for sentence classification (2014). arXiv preprint arXiv:1408.5882
7. dos Santos, C.N., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland (2014)
8. Xu, C., Cetintas, S., Lee, K.C., et al.: Visual Sentiment Prediction with Deep Convolutional Neural Networks (2014). arXiv preprint arXiv:1411.5731
9. You, Q., Luo, J., Jin, H., et al.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI) (2015)
10. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 417–424 (2002)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10. Association for Computational Linguistics, pp. 79–86 (2002)
12. Socher, R., Pennington, J., Huang, E.H., et al.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 151–161. Association for Computational Linguistics (2011)
13. Jia, J., Wu, S., Wang, X., et al.: Can we understand van gogh's mood?: learning to infer affects from images in social networks. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 857–860. ACM (2012)
14. Yang, Y., Jia, J., Zhang, S., et al.: How Do Your Friends on Social Media Disclose Your Emotions. In: Proc. AAAI, 14, pp. 1–7 (2014)
15. Yuan, J., Mcdonough, S., You, Q., et al.: Sentribute: image sentiment analysis from a mid-level perspective. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, p. 10. ACM (2013)
16. Borth, D., Ji, R., Chen, T., et al.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 223–232. ACM (2013)

17. Wang, M., Cao, D., Li, L., et al.: Microblog sentiment analysis based on cross-media bag-of-words model. In: Proceedings of International Conference on Internet Multimedia Computing and Service, p. 76 ACM (2014)
18. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pp. 1–12 (2009)
19. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space (2013). arXiv preprint arXiv:1301.3781
20. Caruana, R., Lawrence, S., Giles, C.L.: Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In: Advances in Neural Information Processing Systems 13, Proceedings of the 2000 Conference, p. 402. MIT Press (2001)
21. Hinton, G.E., Srivastava, N., Krizhevsky, A., et al.: Improving neural networks by preventing co-adaptation of feature detectors (2012). arXiv preprint arXiv:1207.0580
22. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8609–8613. IEEE (2013)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
24. Plutchik, R.: Emotion: A psychoevolutionary synthesis. Harpercollins College Division (1980)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
26. Siersdorfer, S., Minack, E., Deng, F., et al.: Analyzing and predicting sentiment of images on the social web. In: Proceedings of the International Conference on Multimedia, pp. 715–718. ACM (2010)