MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis

Nan Xu[12] Wenji Mao[12]
[1] The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing 10090, China
[2] School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
{xunan2015,wenji.mao}@ia.ac.cn

ABSTRACT

With the prevalence of more diverse and multiform user-generated content in social networking sites, multimodal sentiment analysis has become an increasingly important research topic in recent years. Previous work on multimodal sentiment analysis directly extracts feature representation of each modality and fuse these features for classification. Consequently, some detailed semantic information for sentiment analysis and the correlation between image and text have been ignored. In this paper, we propose a deep semantic network, namely MultiSentiNet, for multimodal sentiment analysis. We first identify object and scene as salient detectors to extract deep semantic features of images. We then propose a visual feature guided attention LSTM model to extract words that are important to understand the sentiment of whole tweet and aggregate the representation of those informative words with visual semantic features, object and scene. The experiments on two public available sentiment datasets verify the effectiveness of our MultiSentiNet model and show that our extracted semantic features demonstrate high correlations with human sentiments.

KEYWORDS

Multimodal Sentiment Analysis; Visual Semantic Features; Deep Neural Network; Attentional Mechanism

1 INTRODUCTION

With the rapid development of social media, social networking sites have become the most important platform to post user-generated content. The modality of micro-blog has become more multiform and people are inclined to post an image in their tweets. Consequently, more and more tweets have both textual and visual content, which continuously brings new challenges to social media analytics and sentiment analysis in particular.

Traditional sentiment analysis methods aim to analyze the textual sentiment polarity, these methods can be divided into two groups: lexicon-based methods [1] and Machine learning based methods [8]. Recently, with the development of the deep neural network, researchers adopt deep learning methods for sentiment analysis, such as CNN [5], RNN [4] model, which get great improvement compared with traditional methods.

According to the kind of visual feature, image sentiment analysis can be divided into three categories: low-level, middle-level and high-level. Borth et al. [2] employed the traditional machine learning based method to detect 1200 adjective-noun pairs (ANP) as the middle-level features, generating a visual sentiment ontology. Motivated by the powerful performance of CNN model in image classification, Xu et al. [10] introduced the CNN to extract the high-level feature in image. They then transferred the weight of VGG on ImageNet into sentiment analysis task and fine-tuned the model in sentiment datasets.

In contrast to traditional single modality sentiment analysis, multimodal sentiment analysis has gained increasing attention in recent years. Most early works use feature selection models, and some recent works are based on deep neural network. In feature-based models, Wang et al. [9] presented a cross-media bag-of-words model to represent the text and image of a Weibo tweet as unified bag-of-words representation. Another representative work based on traditional feature selection model was done by Borth et al. [2]. They used SentiBank to extract 1200 adjective-noun pairs (ANP) as the middle-level features of image and SentiStrength to directly compute the sentiment scores for the tweet text, and then combined both results. In neural network based models, Cai et al. [3] pre-trained text CNN and image CNN to get the representation of text and image. They concatenated these feature vectors for classification with four fully connected layers. Yu et al. [12] also pre-trained CNNs to represent text and image, while its classifier is Logistics Regression. You et al. [11] utilized a cross-modality consistent regression (CCR) model, which used PCNN to extract image feature and added title information to represent the textual modality information. In general, the results show thatneural network based models have better performance then feature section models.

Generally speaking, users' sentiment in social media is usually triggered by the interaction of multiple objects under particular scene with some textual words or sentences. There is no doubt that visual content also contains useful semantic information, such as object and scene. As human sentiments have high corrections with these visual information, they are helpful to understand user's sentiment in multimodal tweets. However, previous work in multimodal sentiment analysis only extracts feature representations of single modality and fuse these features for classification. The detailed semantic information for sentiment analysis and the correlation between image and text have been ignored. To improve

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133142

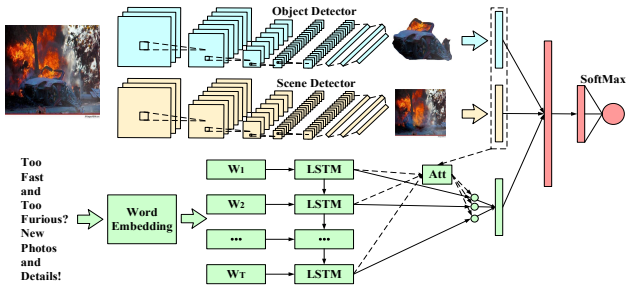

Figure 1: The framework of MultiSentiNet for multimodal sentiment classification.

the performance of deep neural network in sentiment analysis, in this paper, we proposed a deep semantic network, namely MultiSentiNet, which extracts deep semantic features including object and scene from image as the additional information for sentiment classification.

The main contributions of this paper are as follows. We define the triple features of text, object and scene as the representation of a multimodal tweet and regard these deep visual semantic features as the additional information in multimodal sentiment analysis task. We propose a visual feature guided attention LSTM model to extract words that are important to the sentiment of whole tweet and aggregate the representation of these words with visual semantic features, object and scene. The experiments on two public available sentiment datasets verify the effectiveness of our MultiSentiNet model and show that our extracted semantic features demonstrate high correlations with human sentiments.

2 PROPOSED METHOD

The framework of our deep semantic network, MultiSentiNet, for multimodal sentiment analysis is shown in Fig.1. We first use the Object-VGG and Scene-VGG Model to detect visual semantic features: object and scene. Then we propose a visual feature guided attention LSTM model to extract textual feature. Detailed introductions are as follows.

2.1 Visual Feature Detectors

2.1.1 Object Feature. Intuitively, objects are highly associated with sentiments. For example, a bunch of roses can be the symbol of love and brings about positive sentiment while a gun might lead to negative sentiment. Thus, we choose the VGG Model [7] as object detector for visual object extraction, which is composed of five convolutional blocks and three fully connected layers and has been demonstrated the powerful performance in image classification task on ImageNet. We also adopt the transfer learning strategy to overcome the different categories between ImageNet and our multimodal sentiment datasets. The Object-VGG has been pre-trained on ImageNet datasets first. Then the learned parameters are transferred into our sentiment analysis task. We extract the output of last fully connected layer as the object score I_o, which indicates the probability of 1000 object categories.

2.1.2 Scene Feature. The scene in image also contains some useful clues that can be used to assist with understanding user's sentiment potentially. For instance, a wedding chapel brings happiness and positive sentiment while a fire brings sadness and negative sentiment. Thus, scene could be complementary with the sentiment. We adopt the state-of-the-art Scene-VGG model [13] as the scene detector for scene features extraction. The model has been pre-trained on large scale dataset-Place365, which has millions of images for 365 scene classes classification. After transfer learning on our own datasets, we obtain a 365-dimension scene vector I_s to represent the probability of scene classes.

For the different dimensionalities of I_o and I_s, we use d neurons to transfer these visual semantic features into a high-level space to get the same dimensionality.

$$V_o = ReLU(W_o(I_o) + B_o) \in \mathbb{R}^d \tag{1}$$

$$V_s = ReLU(W_s(I_s) + B_s) \in \mathbb{R}^d \tag{2}$$

2.2 Textual Feature Extraction

2.2.1 Tweet Encoder. To considering the context information for understanding the text, we employ the LSTM model for textual feature extraction. Each word w_i of a given tweet is first pre-trained by the state-of-the-art word representation method Glove to generate the word vector embedding x_t. Then we sum up these vectors as the original tweet representation $[x_1, x_2, \ldots, x_T\ ]$, where T is the max length of the tweets. At each time step, the LSTM unit takes an input word vector x_t and output a hidden state $h_t \in \mathbb{R}^d$.

$$x_t = W_{glove} w_t, t \in [1, T] \tag{3}$$

$$h_t = LSTM(x_t), t \in [1, T] \tag{4}$$

2.2.2 Visual Feature Guided Attention. In standard LSTM, the average of the hidden states of all words is used as the feature of tweet. In fact, not all words contribute equally to the sentiment of whole tweet. So, we propose the visual feature guided attention mechanism to extract words that are important to sentiment and aggregate the representation of those informative words with visual semantic features, object and scene.

$$u_t = tanh(W_w h_t + W_o V_o + W s V_s + b) \tag{5}$$

$$\alpha_t = \frac{e^{u_t^T u_w}}{\sum e^{u_t^T u_w}} \tag{6}$$

$$V_\tau = \sum \alpha_t h_t \in \mathbb{R}^d, t \in [1, T] \tag{7}$$

We first feed each word's hidden state h_t with visual object feature V_o and scene feature V_s through a single layer perceptron to generate the deep hidden represention u_t. Then a softmax function is used to ouput a normalized attentional weight α_t. The context vector u_w indicates the informative word over whole tweet, which is randomly initialized and jointly learned during the training process. Last, we compute final textual feature vector V_τ by weighted average of the word hidden vectors based on attentional weight.

2.3 Sentiment Classification

Now we have the high-level triple features: object V_o , scene V_s and text V_τ to represent a multimodal tweet. We first use a fusion

layer for aggregating these triple features to obtain a final multimodal representation. Then a softmax classifier is added in top for sentiment classification.

$$V_{mul} = ReLU(W([V_\tau, V_o, V_s]) + b) \quad (8)$$

$$Pred = SoftMax(W_{mul}(V_{mul}) + b_{mul}) \quad (9)$$

We regard the cross entropy loss as the objective function of softmax. The optimal parameters are trained by back-propagation with RMSProp update rule. In order to avoid overfitting, dropout and early-stopping tricks are also employed.

3 EXPERIMENTS

3.1 Datasets

We conduct experiments based on the MVSA datasets [6], which consist of two separate datasets, MVSA-Single and MVSA-Multi. The former contains 5129 text-image pairs from Twitter. Each pair is shown to a single annotator, who assigns one of three sentiments (positive, negative and neutral) to the text and image respectively. The latter consists of 19600 text-image pairs. While each pair is shown to three annotators, and each annotator's judgment about the sentiments of text and image is independent.

For the MVSA-Multi dataset, we first get the real label for single modality by taking the majority vote out of the three sentiments. That is, an image or a text is considered valid only when at least two of three annotators agree on the exact label. It is natural to verify the sentiment of samples when the textual and visual labels are consistent with each other. However, there are many tweets, in which the textual label and visual label are inconsistent. To ensure high quality data, we first remove these tweets in which one label is positive and the other is negative. While in the case that one label is neutral, the other is positive (or negative), we regard the sentiment polarity of this multimodal tweet as positive (or negative). Thus we get the new MVSA-Single dataset with 4511 text-image pairs and MVSA-Multi dataset with 17024 text-image pairs for our experiments.

3.2 Baselines

We make comparison between our MultiSentiNet model and several baselines, which include several representative works for single modality sentiment analysis (i.e. SentiBank, VGGs for image sentiment and SentiStrength, CNN-Multichannel, LSTM-Avg for text sentiment) and three related works for multiple modalities data (i.e. SentiStrength+SentiBank, CNN-Multi and DNN-LR). 1) *SentiBank* [2] extracts 1200 adjective-noun pairs (ANP) as the middle-level features of image for classification. 2) *SentiStrength* [8] calculates the sentiment scores based on the English grammar and spelling style of text. 3) *CNN-Multichannel* [5] is a CNN model with multichannel and each filter is applied to multichannel, but gradients are back propagated only through one of the channels. 4) *LSTM-Avg* [4] takes the whole document as a single sequence and the average of the hidden states of all words is used as a feature for classification. 5) *SentiBank+SentiStrength* [2] combines the results of SentiBank and SentiStrength to handle multimodal tweet sentiment analysis. 6) *CNN-Multi* [3] uses pre-trained text CNN and image CNN to extract text and image features. And then it concatenates these feature vectors for classification with four fully connected layers. 7) *DNN-LR* [12] also uses pre-trained CNNs to extract the representations of text and image respectively. While the logistics regression is utilized for sentiment classification. 8) For better comparison, we first select two of three features for fusion and generate three variants of MultiSentiNet model: Scene+Object, Text+Scene, Text+Object. While the attention layer changes into object feature guided attention in Text+Object and scene feature guided attention in Text+Scene. We also define the MultiSentiNet-Avg model, which removes the attention layer from our proposed model.

3.3 Experimental Results

We randomly divide the MVSA datasets into training set (80%), validation set (10%) and test set (10%). The metrics used in our experiment are accuracy and F1-measure. We set the learning rate to 0.01, size of mini-batch to 128, the dimension of these triple features to 128 and the dimension of word embeddings to 200. We learn word embeddings by Glove and fine-tune them in training process to adapt the domain of our multimodal sentiment task. Note that all sentences are padded into maximum length T=140, because of the 140 words limit number on Twitter. Table 1 shows the comparison results of different methods on MVSA datasets.

Table 1: Comparison Results of Different Methods

Method	MVSA-Single		MVSA-Multi	
	Acc	***F1***	***Acc***	***F1***
SentiBank	45.22	43.8	55.02	51.15
Scene	63.64	60.4	67.69	65.24
Object	62.08	56.45	65.80	64.75
Scene+Object	64.08	62.33	67.98	66.23
SentiStrength	49.86	48.45	50.57	49.84
CNN-Multichannel	65.19	62.55	65.57	63.24
LSTM-Avg	65.85	64.11	65.69	65.63
SentiBank+SentiStrength	52.05	50.08	65.62	55.36
CNN-Multi	61.20	58.37	66.39	64.19
DNN-LR	61.42	61.03	67.86	66.33
Text+Scene	68.07	66.86	67.80	67.66
Text+Object	66.96	65.72	67.39	66.59
MultiSentiNet-Avg	66.74	66.59	67.86	66.49
MultiSentiNet-Att	**69.84**	**69.63**	**68.86**	**68.11**

Group one in Table 1 shows the prediction effects of different methods only using image data of MVSA. Using scene or object feature performs better than SentiBank with significant improvement, demonstrating the powerful performance of the deep neural network in image classification task. The results also show that using scene feature performs better than object feature for image sentiment analysis. We further combine the scene and object features to get a higher level visual feature for sentiment analysis and gain the best performance on image data.

Group two in Table 1 shows the performance of different methods only using text data. The SentiStrength gets the worst results compared with CNN and LSTM, which also demonstrates the powerful effect of the deep neural network. Besides, it is clear that

the LSTM model outperforms the CNN model for the effects of considering the contextual interactions in textual information by recurrent model.

Group three in Table 1 shows the comparison of our models with baselines using multimodal data. It is clear that our MultiSentiNet-Att model achieves the best accuracy and F1 values. Though the SentiBank+SentiStrength outperforms both SentiBank and SentiStrength, it is still less superior than those deep learning based methods. CNN-Multi and DNN-LR perform better than SentiBank+SentiStrength, but are still obviously worse than our model. Our MultiSentiNet-Att model considers the context information of the text and detects visual semantic features. It extracts key words and aggregates the representation of those informative words with object and scene features by our proposed visual feature guided attention. So it performs better than all baselines. When we remove one kind of visual semantic features, the decline in experimental results of Text+Scene and Text+Object suggest the importance of these visual semantic features in multimodal sentiments analysis. Further, when we remove the attention layer, the performance decreases significantly in MultiSentiNet-Avg, it also demonstrates the power of our visual feature guided attention mechanism in considering those informative words for sentiment.

3.4 Further Analysis

Both scene and object are highly associated with user' sentiment. In order to illustrate precisely this relationship, we calculate the information gain ratio for each category of scene and object with respect to different sentiment polarity and quantify the relevance of particular object and scene to certain sentiment.

Table 2: Top10 Related Negative Objects and Positive Scenes for Sentiments on MVSA-Single Dataset

Top10	Negative-Object	Postive-Scene
1	prison	beauty salon
2	hand-held computer	arena performance
3	police van	shoe shop
4	bison	amusement arcade
5	sawmill	dressing room
6	great dane	florist shop indoor
7	wreck	candy store
8	claw	veterinarians office
9	military uniform	bakery shop
10	trailer truck	child's room

We take negative sentiment related object (negative-object) and positive sentiment related scene (positive-scene) for further analysis. Table 2 shows the results of top10 related negative-object and positive-scene for sentiments on MVSA-Single dataset, which are sorted by information gain ratio. From the table, we can see the sentiment related objects and scenes people are familiar with. For example, the prison is usually associated with criminals. The police van is often accompanied by criminal activities and offenders. Wreck indicates that there was a disaster etc. These objects often bring about negative sentiment. While, when becoming beautiful in beauty salon, enjoying wonderful performance of arena, or getting relaxed in amusement arcade, users would like to publish tweets to share their happiness and express positive sentiment.

4 CONCLUSIONS

Generally speaking, users' sentiments in social media is usually triggered by the interaction of multiple objects under particular scenes with some textual words. Previous work on multimodal sentiment analysis directly extracts features of each modality and fuse these features for classification. The detailed semantic information, object and scene, and the correlation between image and text have been ignored. In this paper, we proposed a deep semantic network, MultiSentiNet, to extract object and scene from the image as the additional information in multimodal sentiment analysis. We also propose a visual feature guided attention model to extract words that are important to understand the sentiment of whole tweet and aggregate the representation of those informative words with these visual semantic features. The experiments on two public available datasets verify the effectiveness of our MultiSentiNet model and show that our extracted semantic features demonstrate high correlations with human sentiments.

ACKNOWLEDGMENTS

This work is supported by NSFC Grant 71621002, 61671450 and 71472175, the Ministry of Science and Technology of China Major Grant 2016QY02D0205, CAS Key Grant ZDRW-XH-2017-3 and Grant 2017A074.

REFERENCES

[1] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. In *LREC*, Vol. 10. 2200–2204.

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 223–232.

[3] Guoyong Cai and Binbin Xia. 2015. Convolutional neural networks for multimedia sentiment analysis. In *Natural Language Processing and Chinese Computing*. Springer, 159–167.

[4] Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).

[5] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[6] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*. Springer, 15–27.

[7] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[8] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology* 61, 12 (2010), 2544–2558.

[9] Min Wang, Donglin Cao, Lingxiao Li, Shaozi Li, and Rongrong Ji. 2014. Microblog sentiment analysis based on cross-media bag-of-words model. In *Proceedings of international conference on internet multimedia computing and service*. ACM, 76.

[10] Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. 2014. Visual sentiment prediction with deep convolutional neural networks. *arXiv preprint arXiv:1411.5731* (2014).

[11] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 13–22.

[12] Yuhai Yu, Hongfei Lin, Jiana Meng, and Zhehuan Zhao. 2016. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms* 9, 2 (2016), 41.

[13] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.