Knowledge-Based Systems
Manuscript Draft

Manuscript Number: KNOSYS-D-20-01665

Title: Cross-modal image sentiment analysis via deep correlation of textual semantic

Article Type: Full Length Article

Keywords: sentiment analysis; cross-modal; social image; correlation

Abstract: Social media has become indispensable to people's lives, where they can share their views and emotion with images and texts. Analyzing social images for sentiment prediction can help understand human social behavior and provide better recommendation results. Most current researches on image sentiment analysis have achieved quite good progress, which ignores the semantic correlation between an image and its corresponding descriptive sentences. To capture the complementary multimodal information for joint sentiment classification, in this paper, we propose a novel cross-modal Semantic Content Correlation(SCC) method, which bridges the correlation between images and captions. Specifically, pre-trained convolutional neural networks (CNNs) are leveraged to encode the visual sub-regions contents, and a GloVe is employed to embed the textual semantic. Relying on visual contents and textual semantic, a joint attention network is proposed to learn the content correlation of the image and its caption, which is then exported as an image-text pair. To exploit the dependence of visual contents on textual semantic in caption effectively, the caption is processed by a Class-Aware Sentence Representation (CASR) network with a class dictionary, and a fully connected layer concatenates the outputs of CASR into a class-aware vector. Finally, the class-aware distributed vector is fed into an Inner-class Dependency Long Short-Term Memory network (IDLSTM) with the image-text pair as a query to further capture the cross-modal non-linear correlations for sentiment prediction. The performance of extensive experiments conducted on three datasets verifies the effectiveness of the model SCC.

Cross-modal image sentiment analysis via deep correlation of textual semantic

KE ZHANG[1], YUNWEN ZHU[1], WENJUN ZHANG [1,2], WEILIN ZHANG[3], YONGHUA ZHU[1]

[1] Shanghai Film Academy, Shanghai University, Shanghai, 210000 China
[2] School of Communication and Information Engineering, Shanghai University, Shanghai, 210000 China
[3] School of Computer Engineering and Science, Shanghai University, Shanghai, 210000 China

Corresponding author: YONGHUA ZHU (e-mail: zyh@shu.edu.cn).

This work is supported by National Key Research and Development Plan of China, under Grant2017YFD0400101,and CERNET Innovation Project, under Grant NGII20160325.

Abstract: Social media has become indispensable to people's lives, where they can share their views and emotion with images and texts. Analyzing social images for sentiment prediction can help understand human social behavior and provide better recommendation results. Most current researches on image sentiment analysis have achieved quite good progress, which ignores the semantic correlation between an image and its corresponding descriptive sentences. To capture the complementary multimodal information for joint sentiment classification, in this paper, we propose a novel cross-modal Semantic Content Correlation(SCC) method, which bridges the correlation between images and captions. Specifically, pre-trained convolutional neural networks (CNNs) are leveraged to encode the visual sub-regions contents, and a GloVe is employed to embed the textual semantic. Relying on visual contents and textual semantic, a joint attention network is proposed to learn the content correlation of the image and its caption, which is then exported as an image-text pair. To exploit the dependence of visual contents on textual semantic in caption effectively, the caption is processed by a Class-Aware Sentence Representation (CASR) network with a class dictionary, and a fully connected layer concatenates the outputs of CASR into a class-aware vector. Finally, the class-aware distributed vector is fed into an Inner-class Dependency Long Short-Term Memory network (IDLSTM) with the image–text pair as a query to further capture the cross-modal non-linear correlations for sentiment prediction. The performance of extensive experiments conducted on three datasets verifies the effectiveness of the model SCC.

Keyword: sentiment analysis; cross-modal; social image; correlation

1. Introduction

Sentiment analysis, also known as opinion mining, is a fundamental research domain about the opinions, emotions, affections, and attitudes expressed by individuals about many topics. People like to share their daily lives and feelings with others on social media platforms. With mass images and texts people uploaded online, social media platforms have become one of the most essential information sources. Sentiment analysis on social media platforms has been widely applied in many applications fields, such as the stock market [1], political events [2], and recommendation systems [3]. Meanwhile, the advancement of social media gives impetus to the emergence of diversified message posted forms. For example, users on Twitter usually upload images with texts to make tweets more clearly and easy to understand. What's more, users share their photos on the Flickr with descriptive sentences as captions and semantic supplement of their photos. All above suggest that as the application requirements, the sentiment is among the most critical factors for social media, and sentiment analysis has great practical significance.

To understand the connections between user emotion and their behavior, researchers have made efforts on sentiment analysis. Most previous researches [4-7] analyze the sentiment of single modalities, such as images or texts. However, these methods ignore the complementary knowledge between visual content and textual semantic, which plays a vital role in sentiment analysis. To accommodate different modalities, multimodal sentiment analysis methods have attracted researchers gradually, which can be divided into the following two categories. One category of works [8,9] handles features extracted from different modalities separately. Diversified classifiers are designed for different modality, and the outputs of these classifiers are arranged by some rules predefined to generate the final classification results. Although these works consider both images and textual descriptions, they ignore the intrinsic relationship between two modalities. The other category [10-12] is to conduct features extracted from different modalities jointly. The works [11,12] fuse individual features into a joint distributed vector, which is then inputted into a unified sentimental classifier for prediction. However, these methods are still challenging to capture the complex correlation of visual contents and textual semantic.

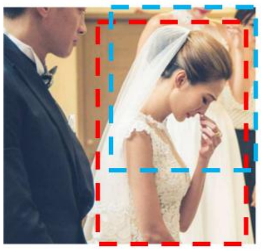

After the vows from the bridegroom, the bride cries with joy.

Figure.1 An illustration of the correlation of an image and its caption. There are many complex correlations between the image and its cation. Taking Twitter as an example, a tweet usually consists of an image and a corresponding descriptive sentence (caption). Therefore, a multimodal sentimental classifier must take both modalities into consideration. However, this task of multimodal sentiment classification remains challenging for the following three reasons.

First, there is a content correlation between an image and its caption. Unlike explicit textual expression, image sentiment is implicit but vivid, so a picture is worth a thousand words. As seen in Figure.1, the terms “bride” and “groom” in the caption are related to the image regions, while the word “joy” is more in line with the image sentiment. Hence, if the correlation between visual contents and textual semantic can be exploited, the result of sentiment classification will be more accurate.

Second, the semantic correlation between an image and its caption is complicated. The visual contents in an image may correspond to multiple words in the description, but the final sentiment of the image may be dependent on a particular word. In Figure 1, the image region corresponds to the word “bride,” which may be predicted as a negative sample according to the word “cry” in the caption. However, the word “joy” indicates that the final image sentiment should be positive. Therefore, the hierarchical correlation between images and captions needs to be analyzed deeply.

Third, on social media platforms, image captions added by users are usually informal and have individual characteristics. For tweets, the caption of an image can be expressed in another way, which may have great impacts on prediction results. For example, the “bride” in the caption can be replaced with “princess” in Figure 1. To capture the complementary information between the images and texts for joint sentiment classification, in this paper, we propose a novel cross-modal Semantic Content Correlation(SCC) method, which bridges the correlation between textual semantic and image contents. Specifically, re-trained convolutional neural networks (CNN) are leveraged to encode the visual sub-regions contents, and a GloVe [13] is employed to embed the textual semantic. Relying on the visual contents and textual implications, a joint attention network is proposed to learn the correlations of an image and its caption, which is then exported as an image-text pair. To exploit the dependence of visual contents on textual semantic in caption effectively, the caption is processed by a Class-Aware Sentence Representation (CASR) network with a class dictionary, and a fully connected layer concatenates the outputs of CASR into a class-aware distributed vector. Finally, the class-aware distributed vector is fed into an Inner-class Dependency Long Short-Term Memory network (IDLSTM) with the image–text pair as a query to further capture the non-linear cross-modal correlations for sentiment prediction. The results of extensive experiments conducted on three datasets verify the feasibility and effectiveness of the model SCC.

The main contributions of this paper are summarized as follows:

- A new cross-modal model SCC is proposed to analyze the correlation between an image and its caption. Our proposed model SCC can explore the image-text pairs to highlight the semantic dependence of image contents in captions for cross-modal sentiment analysis.
- A joint attention network is proposed to learn correlations between visual contents and textual semantic, which is then represented as image-text pairs. Meanwhile, this paper proposes a sentimental inner-class dependency enhancement model IDLSTM. With an image-text pair as the query, the inner dependent relationships between the query and words in the caption are learned by the IDLSTM and the final sentimental prediction is achieved.
- Experiments are conducted on three open-access datasets. A class dictionary is imported to improve the performance of the model SCC. Extensive experimental results demonstrate the effectiveness and superiority of our method against the baselines on three datasets.

The rest of this paper is introduced as follows. In Section 2, we give a brief review of related work. Section 3 describes the methodology of the model SCC in detail. Experimental results are presented and discussed in Section 4. We conclude our work in Section 5.

2. Related Work

2.1 Textual sentiment analysis

On social media platforms, words are the most direct way to convey users' emotions. The text-based sentiment analysis has drawn research's attention, and various textual models are proposed [14]. The text-based sentiment analysis methods fulfill the task by encoding the probability of sentimental words appearing in captions. Sentimental word detection [15], statistical models [16], and semantic web methods [4] are typical approaches. Sentimental word detection is the most widely used method. The statistical model is a classifier trained by a large scale labeled corpus to identified the sentimental intensity of a word. Based on a knowledge graph constructed with expert knowledge, Semantic Web methods mine hidden information between semantic concepts of the corpus. Wordnet-Affect [17], Senticnet[18] are representative large sentiment knowledge graph.

With the development of social media, it is a trend to develop appropriate models for specific tasks. Task-based textual sentiment analysis methods explore the relationship between grammatical features and sentiment, which include aspect-based sentiment analysis (ABSA) [19] and targeted sentiment analysis (TSA). Wang et al. [20] merged attention into a multilayer neural network for textual sentiment analysis. For words in the same caption, the attention parameters show the pivotal word and is a way to measure the association degree of a given aspect, which is represented as maximum probability predicted by a fusion layer. Experiments validate the reduction of training loss in a recurrent neural network (RNN) used so that performance of this multilayer network outweighs that of single-layer classifiers. Combining the regional CNNs with a Long Short-Term Memory Network (LSTM), Liu et al. [21] proposed a model that retains the contents and time-series parameters in image comments without additional dependency analysis. The task of TSA is to exploit the associations of the target word with other sentimental words by deep networks. TDLSTM networks [22] match the target word with the outputs of Bi-LSTM encoders to obtain the sentimental polarity. Tang et al. [23] used an RNN fused a multiple attention layer to get classification results. The weight of the vital word is improved by multi-hop training. Liang et al. [24] proposed a new sentiment analysis method for the target word detection based on co-attention CNNs, which can optimize the deficiency of single attention networks adequately and considerably shorten the loss of training time by taking the parallel text as input.

2.2 Visual sentiment analysis.

Image-based sentimental methods are performed by constructing multi-level image classifiers for sentiment prediction [25,26]. Previous studies mainly covered the methods based on low-level feature extraction [27,28], semantic rules [29,30], and deep learning frameworks [31,32]. The key step of visual sentiment analysis is low-level image feature extraction, such as EEG signals [33], image texture, and color histograms. The most representative methods are based on human facial emotion [34], which can be easily recognized as the most obvious emotional symbols. Still, these methods are inapplicable to many fields due to the semantic gap between visual features and advanced sentiment. With the popularity of deep neural networks, You et al. [35] proposed pre-trained domain knowledge mapping frameworks for emotional analysis. Ahsan [36] proposed a framework to learn the visual attributes of images on social News for image sentiment classification. Song et al. [37] constructed a modified multilayer attention CNN which employs the saliency maps of image regions as a priori knowledge and regularize for sentiment prediction. Dong et al. [38] designed architecture with multiple parallel networks to learn shared parametersThis model is optimized by a joint contrastive loss function to explore the semantic connections of instances. The performance of this architecture on object tracking is excellent.

2.3 Multimodal sentiment analysis

Owing to the lack of direct sentimental correlations of visual contents, and simultaneously multimodal information is sufficient online. Hence studies [39-41] begin to use multimodal data for sentiment analysis. Wollmer et al. [42] proposed to add audio features in speech-based emotion recognition. Poria et al. [43,44] proposed a CNN model with Multi-Kernel Learning to share state parameters. Byrne et al. [45] employed simultaneous derivation on facial emotion classification. Zadeh et al. [46] proposed a multimodal fusion network to learn both intra-modality and inter-modality dynamically. Li et al. [47] established an evaluation model with six priorities to evaluate objects-sentiment dependency. This model realizes the automatic sentiment elements extraction through a merging algorithm of evaluation objects. Adjective Nouns Pairs (ANPs) are sentiment-related phrases extracted from the corpus. By means of ANPs, Borth et al. [48] constructed a visual sentiment ontologies (VSO) Sentibank [51] for sentiment analysis. Maurya et al. [49] summarized the sentiment classification methods based on the VSO. Li et al. [50] adopt a logistic regression model to predict the sentiment orientation based on the ANPs. Similarly, Jiang et al. [52] extracted object-sentiment pairs according to the visual hierarchical features and textual grammar analysis.

2.4 Cross-modal sentiment analysis

Unlike multimodal methods, the information of one modality may be supplementary for the other modality. Consequently, cross-modal sentiment analysis is trying to extract supplementary information from texts for image sentiment classification. In order to align the multimodal features with inconsistent dimensions, cross-modal methods transfer the knowledge from textual modality to visual modality by finding the mapping rules of two modalities. Tsai et al. [53] proposed a learning algorithm of cross-domain landmark selection for heterogeneous domain adaptation. Huang et al. [54] performed two parallel knowledge learning networks. The parameters learned on one modality are transferred to the other network for parameter fusion. Schmitter et al. [55] proposed a mapping method for sentimental annotation with a two-step sentimental transformation. Ji [56] Proposed a bi-layer hypergraph learning approach toward robust sentiment prediction of tweets. Xu et al. [57][61] fused social links, visual contents and textual semantic fragments with a relation network for sentimental prediction. The disadvantage of this method is that the social links are unreliable and unsuitable to be a basis for sentiment analysis. Van et al. [58] proposed an image weighting method with kernel learning to minimizes maximum mean discrepancy (MMD) between visual features and textual features. Wu et al. [59] proposed a compound CNN with the weakly weighted pairwise ranking loss for image sentimental annotation. Huang et al. [60] proposed a hybrid transfer network to learn cross-modal semantic correlation with two subnetworks as a bridge.

3. Proposed framework

In this section, we introduce the architecture of the model SCC in detail. Figure.2 shows the architecture of our method SCC. This model mainly consists of two phases, which are image-text pairs detection and sentiment classification.

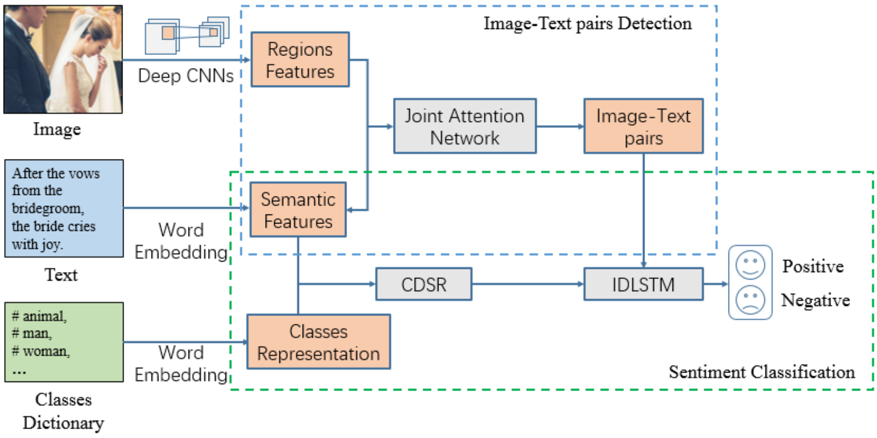


Figure.2 The architecture of our proposed method SCC

To learn the complementary information, an image, its caption, and a class dictionary are inputted to our model for joint sentiment prediction. Accurately, the image and its caption are processed separately at first. Pre-trained Deep CNNs are leveraged to encode the visual sub-regions contents, and a GloVe is employed to embedded each word in the caption to semantic features. Then, in the image-text pair detection phase, to learn the correlation between the image regions and the corresponding textual description, a joint attention network is proposed to interact image sub-regions with each word in the caption, which the correlation is denoted as attention scores to obtains the image-text pairs. In the sentiment classification phase, a class dictionary is used by the Classes-Aware Sentence Representation (CASR) model to fuse the caption with the class information. The outputs of CASR are input into an Inner-Class Dependency Long Short-Term Memory network (IDLSTM) with the image-text pair used as a query to further learn the non-linear correlations for the sentiment polarity classification. Section 3.1 will illustrate the visual and textual features extraction. Section 3.2 will explain in detail the joint attention network for image-text pairs detection. Section 3.3 will provide more information about the CASR and IDLSTM of sentiment classification procedure.

3.1. Feature extraction

We first define the task in this paper. Let I denote an image, S denotes the descriptions of the image. The task is to predict the image sentiment polarity (positive, negative) with the correlation between the two modalities. The image and its caption are preprocessed separately to obtain multimodal distributed vectors. The image is analyzed by pre-trained Deep CNNs, and the caption is transformed through word embedding. For an image, V is denoted as the set of sub-regional maps, and each sub-region map v_i related to one local region in the image. The image region maps V extracted through Deep CNNs are represented as follows:

$$V = DeepCNNs(I) \tag{1}$$

$$V = \{v_1, \dots, v_i, \dots, v_N\} \in \mathbb{R}^{D_I \times N} \tag{2}$$

where $v_i \in \mathbb{R}^{D_I}$ is the D_I-dimensional region vector, N is the maximum region number of the raw image. The caption of the given image is denoted as $S = [W_1, \dots, W_i, \dots, W_L]$, where W_i is the word in the caption, and L is the maximum word number of the sentence in the caption. The GloVe embeds each word W_i as a word vector $w_i \in \mathbb{R}^{300}$, and the sentence is represented as $s = \{w_1, \dots, w_i, \dots, w_L\} \in \mathbb{R}^{300 \times L}$.

3.2. Image-text Detection

The image sentiment is related to image contents [14], and the most concerned content in an image has an impact on the final image sentiment polarity classification. The visual attention mechanism can exploit the interested contents by attention scores. Besides, the attention mechanism can relate the words in image caption to meaningful image regions, instead of considering all regions in the image equally. Different image regions may be associated with different words. These cross-modal semantic associations between image regions and words are beneficial. However, most existing cross-modal sentiment analysis approaches have ignored.

We consider that both the fine-tuned deep CNNs and word embedding networks in the same task already contain the same semantic information. We propose a joint attention network for image-text pairs detection. With a multiplicative embedding method to exploit varying correlation degrees between the image regions and words, the region with the maximum attention score will be selected by the joint attention network as the sentimental image region, and the image-text pair will be obtained at the same time. Compared with embedding the whole sentence of the image description, embedding each word separately for joint attention will be more helpful in improving the accuracy of image-text pairs detection.

After the last feature extraction phase, region maps of an image obtained are denoted as $V = \{v_1, \dots, v_i, \dots, v_N\} \in \mathbb{R}^{D_I \times N}$, and corresponding image descriptive words are denoted as $s = \{w_1, \dots, w_i, \dots, w_L\} \in \mathbb{R}^{300 \times L}$. A score α_{ij} is assigned as the attention weight to each image region v_i with same word w_j through a softmax function:

$$\alpha_{ij} = \frac{exp\ (e_{ij})}{\sum_{i=1}^{N} exp\ (e_{ij})} \tag{3}$$

$$exp(e_{ij}) = \varphi((w_j)^T W_l v_i + b_l) \tag{4}$$

To maintain the input length of the neural network, the embedding s is 0-padded for joint learning. Where $exp(e_{ij})$ calculates how closely the region v_i associated to the word w_j. $1 \le i \le N$, $1 \le j \le L$. $\varphi(\cdot)$ is the smooth function. W_l is the weight matrix, and b_l is the bias term. $Tanh$ is adopted as an activation function. All parameters are learned in the training stage. The correlations between different regions and words are learned when attention scores obtained. Then, word w_j related to the region v_i with maximum attention score can be heightened and selected as an image-text pair :

$$w_j = \mathrm{argmax}_{v_i}\, \alpha_{ij} \tag{5}$$

Through the above method, we can obtain the image-text pair corresponding to the image region with maximum attention. Each of the regions interacts respectively with words in caption to mine the correlation between image contents and textual semantic. And the most optimal semantic matching text for a concerned region is prepared by attention scores for sentiment classification.

3.3. Sentiment classification

In the image-text pairs detection phase, maximum attention weights highlight the most concerning image content and the word, and then the corresponding semantic image-text pair is obtained. We construct a classifier for sentiment classification in this section. The sentiment classifier consists of two parts: CASR and IDLSTM. CASR concatenates all words represented with a given class in the descriptive sentences. IDLSTM is an LSTM proposed for modeling the dependency of one word with the other words in the same caption. Figure.3 displays the architecture of the classifier proposed.

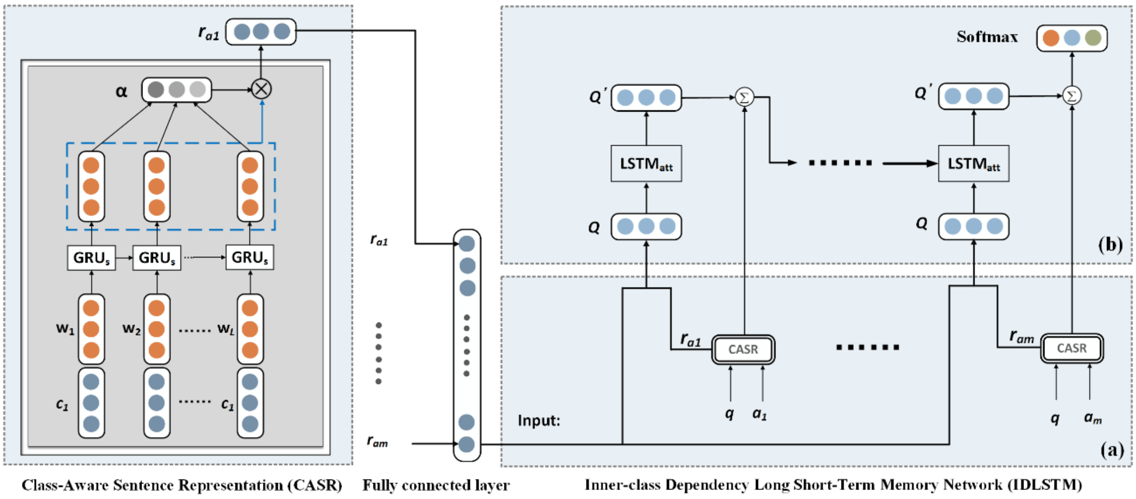


Figure 3. The architecture of proposed sentimental classifier

A. Class-Aware Sentence Representation

Considering that different people may attach personalized characters to describe the same image content online, a class dictionary is utilized to reduce the deviations from personal expression. The class dictionary is denoted as $C = [C_1, \ldots, C_i, \ldots, C_M]$, where C_i represents each class in the dictionary, and M is the maximum class number of the dictionary. The caption of a given image is represented as sentence $S = [W_1, \ldots, W_i, \ldots, W_L]$, where W_i is the word in the caption, and L denotes the maximum word number of the caption. The word W_i and class C_i are embedded by the GloVe. And the embedded dimensions of image caption are $s \in \mathbb{R}^{L\times D}$, and the class word is represented as $c_i \in \mathbb{R}^{D}$. Inspired by [62], each word vector w_i in the sentence is concatenated to a given class c_i:

$$S_{c_i} = \{w_1c_i, w_2c_i, \ldots, w_Lc_i\} \in \mathbb{R}^{L\times 2D} \tag{6}$$

In order to propagate the context information, the distributed representation S_{c_i} is input to a gated recurrent unit (GRU), an attention layer followed by the GRU to weight the word in sentence. The formulations are:

$$z = \sigma(x_tU^z + s_{t-1}W^z) \tag{7}$$

$$r = \sigma(x_tU^r + s_{t-1}W^r) \tag{8}$$

$$h_t = tanh(x_tU^h + (s_{t-1} * r)W^h) \tag{9}$$

$$s_t = (1-z) * h_t + z * s_{t-1} \tag{10}$$

where h_t and s_t are hidden outputs and the cell states at time t. z is the update gate, and r is the reset gate. σ and $tanh$ are activation functions. This step is represented as follows: $R_{c_i} = GRU_s(S_{c_i})$, where $R_{c_i} \in \mathbb{R}^{L\times D_s}$, $U_s^z \in \mathbb{R}^{2D\times D_s}$, $W_s^z \in \mathbb{R}^{D_s\times D_s}$, $U_s^r \in \mathbb{R}^{2D\times D_s}$, $W_s^z \in \mathbb{R}^{D_s\times D_s}$, $U_s^h \in \mathbb{R}^{2D\times D_s}$, $W_s^h \in \mathbb{R}^{D_s\times D_s}$.

To amplify the sentimental relevance of the words to class c_i, we add an attention layer to capture the class-aware representation:

$$z = R_{c_i}W_s + b_s \tag{11}$$

$$\alpha = softmax(z) \tag{12}$$

$$r_{c_i} = \alpha^T R_{c_i} \tag{13}$$

where $z = [z_1, z_2, \ldots, z_L] \in \mathbb{R}^{L\times 1}$, $softmax(x) = \left[\frac{e^{x_1}}{\sum_j e^{x_j}}, \frac{e^{x_2}}{\sum_j e^{x_j}}, \ldots, \frac{e^{x_j}}{\sum_j e^{x_j}}\right]$, attention weight $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_L] \in \mathbb{R}^{D_s\times 1}$, the class-aware representation is $r_{c_i} \in R^{D_s}$, $W_s \in \mathbb{R}^{D_s\times 1}$, and b_s is a scalar.

In the above method, classes in the dictionary are iterated by model CASR with words in the caption. A fully connected layer combines all attended class-aware sentence representations r_c as the input of the model IDLSTM.

B. Inner-class Dependency Long Short-Term Memory Network

The image-text pair of the given image and the class-aware representation r_c are prepared. To reinforce the semantic context of the caption related to the visual content. We employ a modified LSTM to explore the dependency of one word related to visual content on the other words in the same caption by updating the memory of the word.

Figure 3 shows the framework of the IDLSTM, where the image-text pair obtained by the joint attention network is supplied as the query q. To maintain the consistency of the image-text pair and the caption, the query q is also concatenated to class c_i as $q' = qc_i \in \mathbb{R}^{2D}$. In Figure 3 part (a), the joint representation Q is input to LSTM as a supplement.

$$Q = q' {r_C}^T \tag{14}$$

$$\beta = \mathrm{softmax}(Q) \tag{15}$$

where $Q = [Q_1, Q_2, \ldots, Q_M] \in \mathbb{R}^{M\times 1}$and attention weight $\beta = [\beta_1, \beta_2, \ldots, \beta_M] \in \mathbb{R}^{M\times 1}$. Each β_i is a refined measurement for the correlation of the query with the sentence. To avoid forgetting the memory easily, an advanced LSTM with attention $LSTM_{att}$ with size D_o learn the dependency of query q on different words. As shown in Figure 3 part (b).

$$Q' = LSTM_{att}(Q) \tag{16}$$

where the parameters of $LSTM_{at}$ are $U_a^z \in \mathbb{R}^{D_s\times D_o}$, $W_a^z \in \mathbb{R}^{D_o\times D_o}$, $U_a^r \in \mathbb{R}^{D_s\times D_o}$, $W_a^r \in \mathbb{R}^{D_o\times D_o}$, $U_a^h \in \mathbb{R}^{D_s\times D_o}$, and $W_a^h \in \mathbb{R}^{D_o\times D_o}$.The vector o represents the memory of dependencies weighted by β according to the relatedness:

$$o = \beta^T Q' \tag{17}$$

where $o \in R^{D_o}$. At last, the query q reified the visual content is accumulated with the output o of each iteration. The sum is passed to a softmax function for sentiment classification.

$$q'_{(h+1)} = q'_{(h)} + o \tag{18}$$

$$\mathrm{P} = \mathrm{softmax}[(q'_{(h+1)})W_{smax} + b_{smax}] \tag{19}$$

$$\hat{y} = \mathrm{argmax}_i(P[i]) \tag{20}$$

where $W_{Smax} \in \mathbb{R}^{D_o\times 2}$, $b_{smax} \in R^2$, and $\hat{y}$ is the prediction result.

3.4 Optimization

To optimize the model SCC, we use the algorithm ADAM [63] based on stochastic gradient descent (SGD). We cross-validate the model with the loss function as follows:

$$Loss = -\frac{1}{n}\sum_{i=1}^{N}\sum_{k=0}^{C-1}\log p[k] + \lambda\|\theta\|_2^2 \tag{21}$$

where i is the index of samples, n is the number of samples, k is the number of class, λ is the weight of L2 regularization.

4. Experiments

In this section, we will conduct extensive experiments to verify the performance of our model SCC in terms of three datasets. We introduce three datasets and the class dictionary used for experiments firstly. Experimental settings are presented in Section 4.2. Then, section 4.3 offers four evaluation protocols and baselines. Next, experiments in section 4.5-4.7 will validate the effectiveness of the SCC on three datasets. Finally, an ablation study will be conducted on the three models with different components removed in section 4.8. And visualization of two qualitative case studies will illustrate the performance of the model SCC for cross-modal image sentiment analysis.

4.1. Dataset

Experiments are carried out on three social image datasets collected from Flickr, Getty Images, and Twitter.

Flickr: 1200 ANPs contained in the SentiBank of VSO [32] have strong sentimental relevance. Inspired by this, we use these 1200 ANPs as keywords to download images on the Flickr website by Flickr API. For the task of cross-modal image sentimental classification, we only crawl images with English captions and labels. The final image sentiment of each image is weakly labeled in the light of the ANPs in the descriptive sentences. To make the images and captions more appropriate and serviceable for our experiments, only those images with captions have less than 50 words are kept. Hence the Flickr dataset obtained contains approximately 30,000 images with weakly labeled ANPs, of which 15,593 images are labeled as positive images. Flickr images were then assigned to 5 annotators for manual labeling to get more accurate sentimental labels.

Getty Images: The main reason for using Getty Images as the dataset is that the images are weakly labeled and attached with relatively formal visual descriptions as captions. What's more, Getty images can be easily queried with keywords conveniently. Similarly, 1200 ANPs are used as keywords to retrieve images to get an experimental Getty Images dataset consisting of images, labels, and relevant textual descriptions. By this method, the final weakly labeled Getty Images dataset is obtained in line with the ANPs. The Getty Images dataset contains 10,496 images with weak labels, and in which 6,564 images are labeled as positive images.

Twitter: The Twitter dataset contains 1,269 real-world images collected from the Twitter website. Each sample of the Twitter dataset comprises an image, a textual description, and image labels. The sentimental labels (positive, negative) of samples in this popular benchmark are labeled manually by 5 Amazon Mechanical Turk staffs. According to the number of the 5 AMT staff agree to add the same sentimental label to a sample, the samples in the Twitter dataset can be split into three confidence-level batches. The details of the Twitter dataset are shown in Table 1.

Table 1. The detail of Twitter dataset

Twitter dataset	Positive	Negative
5 agree (high confidence)	581	301
At least 4 agree (medium confidence)	689	427
At least 3 agree (low confidence)	769	500

Class dictionary: To maintain textual semantic consistency in captions, inspired by [64], a class dictionary is used to train the model SCC with datasets jointly. The dictionary consists of 368 classes, which are manually summarized from ANPs and image captions.

4.2. Evaluation Metrics and Baselines

We used four evaluation protocols to evaluate our model: precision (pre), recall (rec), F1, and accuracy(acc). The compared methods are introduced below:

Single textual model: only textual features are fed to a classifier to predict the sentiment. Tan [65] proposed an algorithm using multi-kernel learning (MKL) to learn textual vectors as the input of SVM for sentiment classification. Mikolov [66] proposed an unsupervised model that learns fixed-length distributed representations (DR) from variable-length pieces of texts to predict sentimental polarity.

Single visual model: Single visual model is a regression model using only visual features for sentiment prediction, such as colors and textures. Siersdorfer [67] studied the connection between the sentiment of images expressed in the global color histogram (GCH) and their visual content. You [35] proposed a progressive CNN (PCNN) method to recognize visual contents. Both single textual models and single visual models are unimodal models.

Multimodal model: multimodal models concatenate both textual and visual features extracted from multimodal data for joint sentiment prediction. Borth [48] presented a method Sentibank to automatically detect sentimental ANPs in datasets with a pre-constructed VSO. Yuan [68] proposed a classifier Sentribute to vote middle-level multi-attributes for image sentiment prediction. TFN [69] is the tensor fusion networks modeling intra-modality and inter-modality dynamics.

Cross-modal model: The cross-mode models are developed for sentiment analysis on one modality, and the information of the other modality is consulted as a supplement simultaneously. You [64] proposed a cross-modality consistent regression (CCR), which imposes consistent constraints across different modalities. HDF [57] is a three-level hierarchical LSTMs to learn the inter-modal correlations between image and text at different levels for accurate analysis. Our proposed SCC model is a cross-modal model that uses a joint attention network and an advanced LSTM for image sentiment prediction.

4.3 Experimental settings

In the experimental stage, the image and its caption will be processed by CNN and GloVe networks, respectively. To obtained the image content feature v_i, pre-trained CNN networks are used for image contents detection. Concretely, images resized to 225×225 are fed to a Faster R-CNN network [70], and the top-5 main visual contents are detected. Each content is represented as an image local region feature from a fully connected layer "fc8" of pre-trained CNN VGG-19[71] with the dimensionality of 1000. From a semantic perspective, the last fully connected layer "fc8" of VGG-19 can be seen as the representation of these top-5 important image objects. To captions, users on social media sites usually provide image descriptions. Image captions are saved as a document, and the GloVe is employed to obtain the word embedding after several preprocessing steps. First, we remove the numbers and special characters in the document. Then, we tokenize the document with a tokenizer model NTLK. We also remove the words that appear less than 5 times. In addition, spelling mistakes in the description are corrected. The experimental parameters of the SCC model are as follows: $d_v = 1000,\ d_c = 300,\ d_c = 300,\ \lambda = 10^{-4}$.

In the implementation, the effectiveness of our model SCC is verified with cross-verification. Experiments are performed with 2×NVIDIA GTX1060 GPUs. Samples in each dataset are divided into three partitions randomly by 60%, 20%, and 20%. 60% partition of the dataset is used for training, 20% samples of the dataset are selected for verification, and the rest 20% partition is viewed as the test set. During the training stage, the parameters of the model are optimized by SGD. To prevent overfitting, the benchmark of dropout is set to 0.5, and the learning rate is set to 0.001. Experientially, adaptive learning is used in the training stage, and its parameters are then obtained. Training iterates 10 hops, and the settings are adjusted on the validation. We evaluate the model with the best verification performance on the test set to achieve quantitative results, and the samples are shuffled before the next iteration.

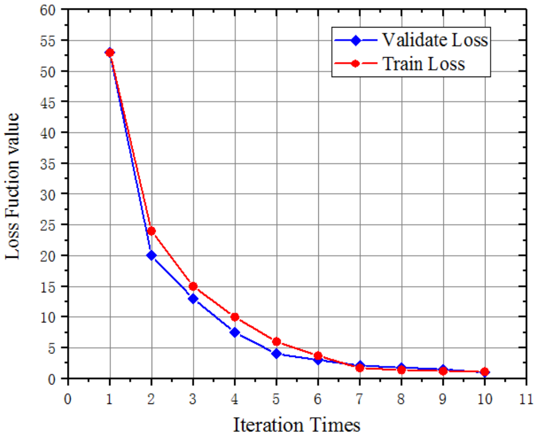


Figure.4 Performance of Loss Function on Train datasets and Validate datasets

Fig.4 shows the change of the objective loss function with the increase of batch iterations. The preliminary analysis result shows that the loss function is inversely proportional to the number of iterations on the randomly chosen batches and converges after 10 iterations. And some randomly selected validation datasets and training datasets are comparable changes on the loss function values.

4.4 Performance on Flickr

Table 2: Performance on Flickr datasets. (%)

Methods	Precisi	Recall	F1	Accura
	on			cy
MKL	0.738	0.742	0.740	0.742
DR	0.765	0.762	0.763	0.761
GCH	0.787	0.801	0.794	0.798
PCNN	0.796	0.803	0.799	0.801
Sentibank	0.755	0.749	0.752	0.766
Sentribute	0.773	0.782	0.777	0.762
TFN	0.836	0.818	0.827	0.831
CCR	0.809	0.815	0.812	0.811
HDF	0.835	0.817	0.826	0.833
SCC	**0.841**	**0.836**	**0.834**	**0.842**

The performance on the Flickr dataset is shown in Table 2. The unimodal methods based on the single modality consider the unimodal features instead of multimodal correlations, so unimodal models perform worse. Comparing the multimodal methods Sentibank and Sentribute with the unimodal methods, the result of multimodal approaches shows that the multimodal contents play a more critical role than unimodal models in sentiment classification. A multimodal fusion network Sentribute proposed for joint learning, which effectively utilizes the relationship between multimodal data. Therefore, the performance of multimodal methods is almost similar to the result of the cross-modal method CCR. It states clearly that the relationship between multimodal data has significant improvement for sentiment classification from the semantic perspective. TFN models focus on semantic alignment so that the values of evaluation protocols are lower than other models for sentiment analysis tasks. Compared with Sentribute, the model CCR adds consistency constraints across the two modalities. What’s more, an effective cross-modal strategy is designed for exploiting in-depth semantic features between visual-text content, which increases its performance by 3.6% relative to Sentribute. The values of the four protocols show that our proposed model SCC is better than other baselines. It indicates that our model achieves the state of the art performance.

4.5 Performance on Getty Images

In Table 3, the precision of multimodal methods is generally higher than the precision of other methods, and recall of unimodal approach GCH proposed is the highest (84.0%). It is because that colors and textures are the intuitive sentimental expressions, but this kind of method cannot classify the sentimental polarity accurately from the semantic perspective. The multimodal models (Sentibank, Sentribute) have higher precision, F1, and accuracy than the unimodal models, but the recall is generally lower than others. Cross-modal model CCR achieves the highest precision of 84.6%. The method GCH achieves the highest recall value at the expense of precision. Even the recall of our model SCC is 5% lower than GCH recall, but the results of the other three metrics are all above that of GCH. Our model SCC beats other baselines with the highest F1 and accuracy. The cross-modal model CCR carries out sentimental image annotation according to parameters learned from the labeled dataset. The accuracy of the labeled images determines the performance of this method. The key to our approach is the correlations of multimodal data to sentiment classification. Hence, even though the precision of CCR is about 1.4% higher than that of the SCC, our model SCC predicts image sentiment from an interpretable perspective by parsing the correlations between images and its complex sentences, which can be confirmed on the basis of F1 and accuracy. Besides, the performance of our model SCC on Flickr is better than on Getty Images, and this verifies the effectiveness of the manual labeling. It means that our adaptive method has better generalization performance in applications.

Table 3: Performance on Getty Images datasets. (%)

Methods	Precisi on	Recall	F1	Accura cy
MKL	0.718	0.685	0.701	0.659
DR	0.820	0.788	0.804	0.790
GCH	0.687	**0.840**	0.756	0.697
PCNN	0.730	0.744	0.737	0.717
Sentibank	0.742	0.727	0.734	0.675
Sentribute	0.769	0.698	0.731	0.727
TFN	0.819	0.803	0.801	0.802
CCR	**0.846**	0.759	0.780	0.800
HDF	0.807	0.801	0.804	0.799
SCC	0.832	0.791	**0.810**	**0.806**

4.6 Performance on Twitter

Table 4: Performance on Twitter dataset. (%)

Methods	5 agree				At least 4 agree				At least 3 agree			
	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc
MKL	-	-	-	-	-	-	-	-	-	-	-	-
DR	0.746	0.693	0.727	0.722	-	-	-	-	-	-	-	-
GCH	0.708	**0.888**	0.787	0.684	0.687	0.840	0.756	0.665	0.687	**0.836**	0.749	0.660
PCNN	0.797	0.811	0.836	0.733	0.786	**0.842**	0.811	0.759	0.755	0.805	0.778	0.723
Sentibank	0.785	0.768	0.776	0.709	0.742	0.727	0.734	0.675	0.720	0.723	0.721	0.662
Sentribute	0.789	0.823	0.805	0.738	0.750	0.792	0.771	0.709	0.733	0.783	0.757	0.696
TFN	0.809	0.803	0.811	0.812	-	-	-	-	-	-	-	-
CCR	0.831	0.805	0.818	0.809	-	-	-	-	-	-	-	-
HDF	0.842	0.849	0.845	0.852	-	-	-	-	-	-	-	-
SCC	**0.880**	0.866	**0.862**	**0.863**	**0.831**	0.812	**0.821**	**0.771**	**0.793**	0.776	**0.788**	**0.742**

As shown in Table 4, Our SCC is validated on three confidence-level batches of the Twitter dataset. Four evaluation metrics in Table 4 illustrates that the performance of the model SCC decreases with falling image confidence accordingly. Apparently, the final sentiment of an image is affected by people's subjectivity. However, even considering subjectivity, our model SCC can maintain great advantages in four metrics. Compared with the experimental result of the SCC on three confidence-level batches, the recall of visual methods (GCH, PCNN) is higher than the SCC. Still, the overall performance of single visual models on high confidence-level is not outstanding. The performance suggests that the model SCC learns multi-modalities jointly for sentiment classification has undeniable advantages.

A noticeable phenomenon of the result in Table 4 is that the recall of our model SCC is roughly equal to the multimodal model (Sentribute) in three confidence levels batches, and the precision of the model SCC is about 9.1%, 8.1%, and 12.5% higher than that of the multimodal model, respectively. Meanwhile, the accuracy of the SCC is about 8.5%, 6.2%, and 4.6% higher, respectively. On average, our model SCC has an advantage over multimodal models in the precision and accuracy of sentimental prediction. The precision and accuracy of the model SCC are approximately similar to the cross-modal model CCR, but the recall and F1 are increased by 6% and 4.4%, respectively. As can be seen, the general performance of our model SCC is better than that of the CCR model. These results provide valuable insights into the correlation between visual contents and textual semantics, which plays a heuristic role in image sentiment analysis. In summary, all the above results show that the correlations between images and captions are useful for joint learning, and our proposed model SCC has achieved the best performance on cross-modal image sentiment prediction.

4.7 Ablation study

In section 4.7, an ablation study is performed to evaluate the necessity of three components in the architecture of the model SCC. Particularly, the model SCC is re-trained and evaluated without one following component respectively on Flickr dataset:

(1) Single textual model(SCC-V): To quantify the effect of the visual content, the joint attention network is ablated. Prediction of the model SCC-V is the polarity of image captions, which finally decided by sentimental words.

(2) Image sentiment classifier(SCC-C): which consists of the joint attention network and IDLSTM. To study the effect of component CDSR on precision, the captions are input directly to IDLSTM without class fusion after GloVe embedding.

(3) Image sentiment classifier (SCC-I), where the component IDLSTM ignored. The image-text pair, which has been detected by a joint attention network, is input to CDSR as the keyword for word-level fusion. The joint and the component CDSR are both responsible for sentiment prediction.

Table 5: Performance of ablation experiments (%)

Model	Precision	Recall	F1	Accuracy
SCC-V	0.796	**0.803**	0.771	0.765
SCC-C	0.780	0.789	0.790	0.793
SCC-I	0.691	0.697	0.694	0.698
SCC	**0.832**	0.791	**0.810**	**0.806**

In the reason of the uncertainty of the Twitter dataset, four models have been tested on the Flickr dataset. Table 5 shows the performance of the different models. According to the result of SCC-V, even the recall of SCC is not as good as SCC-V by 1.2%, it is evident that the joint attention network can improve the performance of the SCC. Compared with SCC-C and SCC-I, the result indicates that the inner dependency of the words in the caption is sufficient for image sentiment classification. In this study, the effectiveness of each component is verified. It can be concluded that the direct correlations and deep semantic dependence between an image and its caption exist, which affect the final image sentiment. Ablation study shows that visual information has less effect on sentiment than textual semantic. Therefore, our proposed IDLSTM is the key novelty of the method.

4.9 Case Studies

In this section, two case studies will illustrate the effectiveness of the model SCC.

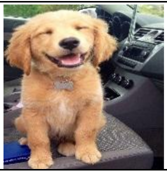
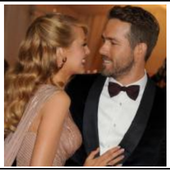
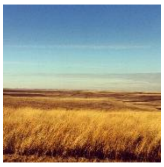
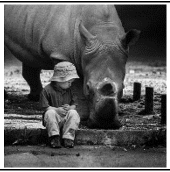
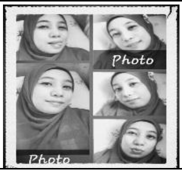
Case 1: Performance on random samples

To compare the performance of different open-source models on different samples, 6 samples with different confidence-level are specially picked from the Twitter dataset and marked with index 1-6. Each image of the six samples has been manually annotated with a sentimental label “positive” or “negative,” which is regarded as the ground truth.

In Table 6, sample 1,2,5 are all positive with high confidence, image 3, 6 are negative samples with low confidence, and sample 4 is negative with high confidence. Samples 1 and 4 with the same high confidence are selected to compare the performance of different models. Samples 2 are chosen to explore the performance of each method on an image with an elaborate caption. Samples 1 and 5 with the same high confidence are selected to compare the performance of different models on different image textures. Sample 3 is used to evaluate the different models on a real-world image in which the image and caption are not matched. Sample 6 is selected to evaluate the different models on a confusing image.

Due to the sentiment of the low confidence samples is debatable and subjective, we consider that the final sentiment of the image is easier to be influenced by image caption. The prediction is the probability of an image being a positive or negative sentiment. Table 6 shows the sentimental labels of different methods forecasted to the 6 samples. Apparently, the single textual model MKL only considers captions. The captions of sample 2,3,4,6 predicted by MKL are negative, but image 2 is actually positive. The single visual models identify images1,2,3, 6 with bright colors as positive samples, and classify sample 4,5 with dull tones as negative images. It is easy to see that only visual features are taken into account make the predict of image5 wrong. Multimodal method Sentibank predicted image2 as a negative image, whereas image2 is positive. We guess the possible reason for this performance might be that the prediction of the Sentibank relies intensely on ANPs detection. If there are no ANPs in image captions, the performance will be inaccurate. The multimodal method Sentribute gives positive prediction to image3, whereas the image3 is a negative sample. What stands out of Sentribute is that the Sentribute is trained to detect particular image sentimental objects, and the final prediction of the image is constrained by the training set. With the model SCC, results of image1,2 are positive and of image3,4,6 are negative. All predicted results consistent with ground truth. The study of case 1 states that our model SCC is adaptive for cross-modal image sentiment analysis.

Table 6: The detail and performance of 6 samples on different models (%)

Sample						
Caption	Hello there sweetie. My best friend :)	This kinda makes me more emotion than sadness.	Autumn is full of desolation:(	Animals are our friend. No trading, no killing.	Self-timer will make me more perfect.	I'm still upset over the game!
Index	1	2	3	4	5	6
Confidence level	High	High	low	High	High	low
Ground truth	positive	positive	negative	negative	positive	negative
MKL	positive	negative	negative	negative	positive	negative
GCH	positive	positive	positive	negative	negative	positive
PCNN	positive	positive	positive	negative	negative	positive
Sentibank	positive	negative	positive	negative	positive	negative
Sentribute	positive	positive	negative	negative	positive	negative
SCC	positive	positive	negative	negative	positive	negative

Case 2: Visualization of the proposed method

In case 2, we mainly visualize the procedure of our model SCC by capturing the memory of the IDLSTM after each iteration. Owing to the same class of words in the image caption might belong to, attention is appropriated to find the words related to image contents. In Figure.5, we pick an image with a caption "The coffee is well prepared, and the service impeccable." The "coffee" and "service" in the sentence are both sentimental entities. When "coffee" and "service" coexist in a sentence, the joint attention network of the model SCC detects the image content as a priori knowledge, and then output the "food" as the image-text pair. The attention-based component IDLSTM analyzed the relation of the word "coffee" with other words in sentence repeatedly and finally highlighted sentimental words that determines the image sentiment.

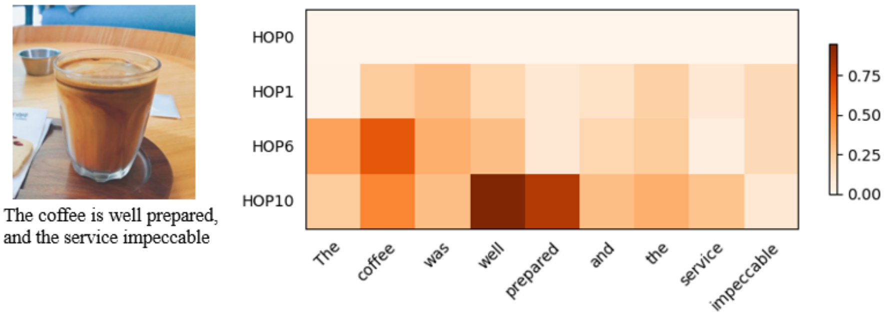


Figure.5 The visualization of the model SCC in image sentiment prediction

Figure.5 illustrates the visualization of the model SCC in image sentiment prediction. As for visualization of the model SCC, the attention weights for the sentence "The coffee is well prepared, and the service impeccable" is reflected by the word importance hotmap. The stronger sentimental correlation will make the deeper color in the hotmap. The weight of each word is initialized to 0 at the beginning. Then, the sentimental words related to the image content "coffee" are gradually highlighted with the incremental iterations. Such a complex correlation between the "coffee" and the corresponding sentiment-bearing word "well" is grasped by our model SCC. The model SCC performs considerably well when the correlation plays a vital role in understanding the sentence structure and meaning for image sentiment analysis.

5. Conclusion

In this paper, we propose a novel cross-modal Semantic Content Correlation method that bridges the correlation between images and captions. This model detects image-text pairs with a joint attention network and then use it as the query to get the dependency of words in the caption. Thus the prediction of image sentiment is obtained by the correlation of two modalities. In contrast to the previous works based on content recognition or multimodal feature extraction, the final cross-modal image sentiment prediction of the model SCC depends on the correlations of image content and textual semantic. In experiments, results validate the performance of our model SCC. However, investigations have exposed several unsatisfied problems of the model SCC, such as performance is limited by datasets and excessive memory overhead. In future work, we will improve the model from the following aspects：1) Model parameters optimization and structural reconstruction. 2) The model SCC will be enhanced to take audio as input.

ACKNOWLEDGMENT

This work is supported by National Key Research and Development Plan of China, Grant/Award Number: 2017YFD0400101; CERNET Innovation Project, Grant/Award Number: NGII20160325.

Reference:

[1] Haider Maqsood, Irfan Mehmood, Muazzam Maqsood, et al. “A local and global event sentiment based efficient stock exchange forecasting using deep learning”, International Journal of Information Management, vol.50, pp: 432-451, 2020.

[2] Liu D, Lei L. “The appeal to political sentiment: An analysis of Donald Trump’s and Hillary Clinton’s speech themes and discourse strategies in the 2016 US presidential election”. Discourse, Context and Media, vol.25, no.10, pp:143-152, 2018.

[3] Ming-Ti Chiang, Mei-Chen Lin. “Market sentiment and herding in analysts’ stock recommendations”. The North American Journal of Economics and Finance.nol.48, pp: 48-64, 2019.

[4] Giatsoglou M, Vozalis M G, Diamantaras K I, et al. “Sentiment analysis leveraging emotions and word embeddings”. Expert Systems with Applications., vol.60, no.1, pp: 214-224, 2017.

[5] Singh V, Ram M, Pant B. “Identification of Zonal-Wise Passenger’s Issues in Indian Railways Using Latent Dirichlet Allocation (LDA): A Sentiment Analysis Approach On Tweets.” Mathematics Applied in Information Systems., vol.2, no.1, pp: 265-276, 2018.

[6] Chaturvedi I, Ragusa E, Gastaldo P, et al. “Bayesian network based extreme learning machine for subjectivity detection.” Journal of The Franklin Institute, vol.355, no.4, pp: 1780-1797, 2018.

[7] Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, Deepak Padmanabhan. “Lexicon Generation for Emotion Detection from Text”. IEEE Intelligent Systems., vol.32, no.7, pp: 102-108, 2017.

[8] Rozgic V, Ananthakrishnan S, Saleem S, et al. “Ensemble of SVM trees for multimodal emotion recognition.” In Signal & Information Processing Association Annual Summit and Conference, Asia-Pacific. 2012, pp: 1-4.

[9] Rosa R L, Schwartz G M, Ruggiero W V, et al. “A Knowledge-Based Recommendation System That Includes Sentiment Analysis and Deep Learning.” IEEE Transactions on Industrial Informatics, vol.15, no.4, pp: 2124-2135, 2019.

[10] Li Y, Pan Q, Wang S, et al. “A Generative Model for category text generation.” Information Sciences, vol.450, no.1, pp: 301-315, 2018.

[11] Shunxiang Zhang, Zhongliang Wei, Yin Wang, Tao Liao. “Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network.” Procedia Computer Science, vol.147, no.1, pp: 361-368, 2019.

[12] Yazhou Zhang, Dawei Song, Xiang Li, et al. “A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis”. Information Fusion, vol.62, pp:14-31,2020.

[13] Pennington J, Socher R, Manning C D, et al. “Glove: Global Vectors for Word Representation.” in Preceeding of the 2014 empirical methods in natural language processing (EMNLP), 2014, pp: 1532-1543.

[14] Cambria E, Das D, Bandyopadhyay S, et al. “Affective Computing and Sentiment Analysis.” in A Practical Guide to Sentiment Analysis. Springer International Publishing., 2017, chapter 1, pp: 1-15.

[15] Xu G, Yu Z, Yao H, et al. “Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary”. IEEE Access., vol.7, no. 1, pp:

43749-43762, 2019.

[16] Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, Deepak Padmanabhan. “Lexicon Generation for Emotion Detection from Text”. IEEE Intelligent Systems., vol.32, no.7, pp: 102-108, 2017.

[17] Kulkarni P V, Nagori M B, Kshirsagar V P. “An In-Depth Survey of Techniques Employed in Construction of Emotional Lexicon”. In Information and Communication Technology for Intelligent Systems. Springer, Singapore, 2019, pp: 609-620.

[18] Cambria E, Poria S, Hazarika D, et al. “SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings.” In Thirty-Second AAAI Conference on Artificial Intelligence.2018, pp: 1795-1802.

[19] Zainuddin N, Selamat A, Ibrahim R. “Hybrid sentiment classification on twitter aspect-based sentiment analysis.” Applied Intelligence, vol.48, no.5. pp: 1218-1232, 2018.

[20] Wang Y, Huang M, Zhao L. “Attention-based LSTM for aspect-level sentiment classification”. in Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 606-615.

[21] Liu G, Huang X, Liu X, et al. “A Novel Aspect-based Sentiment Analysis Network Model Based on Multilingual Hierarchy in Online Social Network”. The Computer Journal, 2019. DOI: 10.1093/comjnl/bxz031.

[22] Ma Y, Peng H, Cambria E. “Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM.” In Thirty-Second AAAI Conference on Artificial Intelligence. April 2018, pp: 5876-5883.

[23] Tang D, Qin B, Feng X, et al. “Effective LSTMs for Target-Dependent Sentiment Classification.” In International conference on computational linguistics, 2016, pp: 3298-3307.

[24] Liang Bin, Liu Quan, Xu Jin, Zhou Qian, Zhang Peng. “Aspect-Based Sentiment Analysis Based on Multi-Attention CNN.” Journal of Computer Research and Development.vol.54, no.8, pp:1724-1735, 2017.

[25] Poria S, Cambria E, Bajpai R, et al. “A Review of Affective Computing: From unimodal analysis to multimodal fusion.” Information Fusion, vol.37, no.1, pp: 98-125, 2017.

[26] Chaturvedi I, Cambria E, Welsch R E, et al. “Distinguishing between facts and opinions for sentiment analysis: Survey and challenges.” Information Fusion, vol.44, no.1, pp: 65-77, 2018.

[27] Lu X, Suryanarayan P, Adams R B, et al. “On shape and the computability of emotions.” In ACM International Conference on Multimedia, 2012, pp: 229-238.

[28] Wang S, Wang J, Wang Z, et al. “Multiple Emotion Tagging for Multimedia Data by Exploiting High-Order Dependencies Among Emotions.” IEEE Transactions on Multimedia, vol.17, no.12, pp: 2185-2197, 2015.

[29] Srishti Vashishtha, Seba Susan. “Inferring Sentiments from Supervised Classification of Text and Speech cues using Fuzzy Rules”. Procedia Computer Science, vol.167, pp: 1370-1379, 2020.

[30] Bowen Zhang, Xiaofei Xu, Xutao Li et al. “Sentiment analysis through critic learning for optimizing convolutional neural networks with rules”. Neurocomputing, vol.356, pp:21-30, 2019.

[31] Zhao S, Yao H, Gao Y, et al. “Continuous probability distribution prediction of image emotions via multitask shared sparse regression.” IEEE Transactions on Multimedia, vol.19, no.3, pp: 632-645, 2017.

[32] Wang J, Fu J, Xu Y, et al. “Beyond object recognition: visual sentiment analysis with deep coupled adjective and noun neural networks.” In the International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp: 3484-3490.

[33] Sudhanshu Kumar, Mahendra Yadava, Partha Pratim Roy. “Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction”. Information Fusion, Vol.52, pp:41-52, 2019.

[34] Rout J K, Choo K K R, Dash A K, et al. “A model for sentiment and emotion analysis of unstructured social media text.” Electronic Commerce Research, vol.18, no.1, pp: 181-199, 2018.

[35] You Q, Luo J, Jin H, et al. “Robust image sentiment analysis using progressively trained and domain transferred deep networks.” in National conference on artificial intelligence(AAAI), 2015, pp: 381-388.

[36] Ahsan U, De Choudhury M, Essa I. “Towards using visual attributes to infer image sentiment of social events.” In 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017, pp: 1372-1379.

[37] Song K, Yao T, Ling Q, et al. “Boosting image sentiment analysis with visual attention”. Neurocomputing, vol.312, no.27, pp: 218-228, 2018.

[38] Dong X, Shen J, Wu D, et al. “Quadruplet Network with One-Shot Learning for Fast Visual Object Tracking.” IEEE Transactions on Image Processing, vol.28, no.7, pp: 3516-3527, 2019.

[39] Majumder N, Hazarika D, Gelbukh A, et al. “Multimodal sentiment analysis using hierarchical fusion with context modeling.” Knowledge-Based Systems, vol.161, no.12, pp: 124-133, 2018.

[40] Huang F, Zhang X, Zhao Z, et al. “Image–text sentiment analysis via deep multimodal attentive fusion.” Knowledge-Based Systems, vol.167, no.3, pp: 26-37, 2019.

[41] Chaturvedi I, Satapathy R, Cavallari S, et al. “Fuzzy commonsense reasoning for multimodal sentiment analysis.” Pattern Recognition Letters, vol.125, no.6, pp: 264-270, 2019.

[42] Wollmer M, Weninger F, Knaup T, et al. “YouTube movie reviews: Sentiment analysis in an audio-visual context.” IEEE Intelligent Systems, vol.28, no.3, pp: 46-53,2013.

[43] Poria S, Chaturvedi I, Cambria E, et al. “Convolutional MKL based multimodal emotion recognition and sentiment analysis.” In 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016, pp: 439-448.

[44] Poria S, Peng H, Hussain A, et al. “Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis.” Neurocomputing, vol.261, no.10, pp: 217-230, 2017.

[45] Byrne S P, Mayo A, O’Hair C, et al. “Facial emotion recognition during pregnancy: Examining the effects of facial age and affect.” Infant Behavior and Development, vol.54, no.2, pp: 108-113,2019.

[46] Zadeh A, Chen M, Poria S, et al. “Tensor Fusion Network for Multimodal Sentiment Analysis.” In Empirical methods in natural language processing, 2017, pp: 1103-1114.

[47] Li YG, Zhou XG, Sun Y, Zhang HG. “Research and Implementation of Chinese Microblog Sentiment Classification.” Journal of Software, vol.28, no.12, pp: 3183-3205, 2017.

[48] 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. ACM, 2013, pp: 223-232.

[49] Maurya C G, Gore S, Rajput D S. “A Use of Social Media for Opinion Mining: An Overview (With the Use of Hybrid Textual and Visual Sentiment Ontology).” in Proceedings of International Conference on Recent Advancement on Computer and Communication. Springer, Singapore, 2018, pp: 315-324.

[50] Li Z, Fan Y, Liu W, et al. “Image sentiment prediction based on textual descriptions with adjective noun pairs.” Multimedia Tools and Applications, vol.77, no.1, pp: 1115-1132, 2018.

[51] Borth D, Chen T, Ji R, et al. “SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content.” in ACM International Conference on multimedia, 2013, pp: 459-460.

[52] Jiang TengJiao, Wan Changxuan, Liu Dexi et.al. “Extracting Target-opinion pairs based on semantic analysis.” Chinese Journal of Computers. vol.40, no.3, pp: 617-633,2017.

[53] Tsai Y H, Yeh Y, Wang Y F, et al. “Learning Cross-Domain Landmarks for Heterogeneous Domain Adaptation”. In the conference on computer vision and pattern recognition, 2016, pp: 5081-5090.

[54] Huang X, Peng Y, Yuan M, et al. “Cross-modal Common Representation Learning by Hybrid Transfer Network.” In international joint conference on artificial intelligence, 2017, pp: 1893-1900.

[55] Schmitter P, Steinrucken J, Romer C, et al. “Unsupervised domain adaptation for early detection of drought stress in hyperspectral images.” Isprs Journal of Photogrammetry and Remote Sensing, vol.131, no.9, pp: 65-76, 2017.

[56] Ji R, Chen F, Cao L, et al. “Cross-Modality Microblog Sentiment Prediction via Bi-Layer Multimodal Hypergraph Learning.” IEEE Transactions on Multimedia, vol.21, no.4, pp: 1062-1075, 2019.

[57] Jie Xu, et al. “Sentiment analysis of social images via hierarchical deep fusion of content and links.” Appl. Soft Comput. vol.80, no 4, pp387-399, 2019.

[58] Van Opbroek A, Achterberg H C, Vernooij M W, et al. “Transfer Learning for Image Segmentation by Combining Image Weighting and Kernel Learning.” IEEE Transactions on Medical Imaging, vol.38, no.1, pp: 213-224, 2019.

[59] Wu F, Wang Z, Zhang Z, et al. “Weakly Semi-Supervised Deep Learning for Multi-Label Image Annotation.” IEEE Transactions on Big Data, vol.1, no.3, pp: 109-122, 2015.

[60] Huang X, Peng Y, Yuan M, et al. “Cross-modal Common Representation Learning by Hybrid Transfer Network.” In International joint conference on artificial intelligence(IJCAI), 2017, pp: 1893-1900.

[61] Jie Xu, Feiran Huang, Xiaoming Zhang, et al. “Visual-textual sentiment classification with bi-directional multi-level attention networks”. Knowledge-Based Systems, vol.178, no.4, pp:61-73,2019.

[62] Majumder N, Poria S, Gelbukh A, et al. “IARM: Inter-Aspect Relation Modeling with Memory Networks in Aspect-Based Sentiment Analysis”. Empirical methods in natural language processing, 2018, pp:3402-3411.

[63] Kingma D P , Ba J. “Adam: A Method for Stochastic Optimization.” In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015.

[64] You Q, Luo J, Jin H, et al. “Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia.” In conference of web search and data mining, 2016, pp: 13-22.

[65] Tan J, Xu M, Shang L, et al. “Sentiment analysis for images on microblogging by integrating textual information with multiple kernel learning.” In pacific rim international conference on artificial intelligence, 2016, pp: 496-506.

[66] Le Q V, Mikolov T. “Distributed Representations of Sentences and Documents.” In international conference on machine learning, 2014, pp: 1188-1196.

[67] Siersdorfer S, Minack E, Deng F, et al. “Analyzing and predicting sentiment of images on the social web.” In Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010, pp: 715-718.

[68] 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. ACM, 2013, DOI: 10.1145/2502069.2502079.

[69] A. Zadeh, M. Chen, S. Poria, et al. “Tensor fusion network for multimodal sentiment analysis”. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, 2017, pp. 1103–1114.

[70] Ren S, He K, Girshick R, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, no.6, pp: 1137-1149, 2017.

[71] K. Simonyan, A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. Computer vision and pattern recognition CoRR (2014) arXiv:1409.1556.