

Research Article

GIF Video Sentiment Detection Using Semantic Sequence

Dazhen Lin,^{1,2,3} Donglin Cao,^{1,2,3} Yanping Lv,^{1,2} and Zheng Cai^{1,2}

¹Cognitive Science Department, Xiamen University, Xiamen, China

²Fujian Key Laboratory of Brain-Inspired Computing Technique and Applications, Xiamen University, Xiamen, China

³Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China

Correspondence should be addressed to Donglin Cao; another@xmu.edu.cn

Received 11 November 2016; Revised 10 March 2017; Accepted 26 March 2017; Published 16 May 2017

Academic Editor: Simone Bianco

Copyright © 2017 Dazhen Lin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of social media, an increasing number of people use short videos in social media applications to express their opinions and sentiments. However, sentiment detection of short videos is a very challenging task because of the semantic gap problem and sequence based sentiment understanding problem. In this context, we propose a SentiPair Sequence based GIF video sentiment detection approach with two contributions. First, we propose a Synset Forest method to extract sentiment related semantic concepts from WordNet to build a robust SentiPair label set. This approach considers the semantic gap between label words and selects a robust label subset which is related to sentiment. Secondly, we propose a SentiPair Sequence based GIF video sentiment detection approach that learns the semantic sequence to understand the sentiment from GIF videos. Our experiment results on GSO-2016 (GIF Sentiment Ontology) data show that our approach not only outperforms four state-of-the-art classification methods but also shows better performance than the state-of-the-art middle level sentiment ontology features, Adjective Noun Pairs (ANPs).

1. Introduction

Nowadays, social applications (such as Facebook, Twitter, and Weibo) contain a huge number of texts, images, and video clips (GIF). With faster Internet connection, people are more willing to post GIF videos than static images to make a personalized and appealing post. According to a recent study [1], the total proportion of visual contents from all shared links on Twitter is 36%. Our statistical results on Sina Weibo, the largest microblog in China, show that 24% of multimedia posts contain GIF videos. However, despite the popularity of GIF videos in social networks, most sentiment detection approaches obtain users' opinions by using only text based sentiment analysis technology. Researches for GIF sentiment analysis are still in the beginning. There are two main challenges for GIF sentiment analysis: semantic gap problem and sequence based sentiment understanding problem. Firstly, the learning process lacks middle level features and a corresponding semantic label measure. Without semantic label measure, machine cannot learn the middle level sentiment semantic elements and their relation from low level features. Secondly, semantic sequence based sentiment expression is one of the important issues in GIF sentiment

analysis. Because sentiments are hidden in the sequence of images, machine cannot mine the impact of semantic sequence based sentiment expression from bag-of-words based features expression.

In particular, we make the following contributions to solve the above two problems:

- (1) We propose a Synset Forest method to select semantic SentiPair labels that solves the semantic gap problem in label set.
- (2) We propose a SentiPair Sequence based GIF video sentiment detection approach that solves the sequence based sentiment understanding problem.

The remainder of this paper is organized as follows. Section 2 briefly describes the background and related works in visual sentiment analysis. Section 3 presents the middle level feature, SentiPair Sequence. The algorithm and framework of SentiPair Sequence based approach are detailed in Section 4. Experimental results on GIF video dataset are given in Section 5. Section 6 draws conclusions and gives directions for future work.

2. Related Work

Traditional sentiment research works focus on text based sentiment analysis because words are the most common way of expressing opinions. According to the granularity of analysis, text based sentiment research works can be divided into three levels: document level [2], sentence level [3], and entity level [4].

With the development of mobile devices and social media, an increasing number of GIF videos were used to express opinions of users in social media. Hence, visual sentiment analysis becomes a hot topic in multimedia and social media fields. According to the visual content type, recent studies can be divided into two types: image sentiment analysis and video sentiment analysis.

For image sentiment analysis, You et al. [5] used the progressive CNN and bypassed the midlevel features. Without the midlevel ontology, the number of neurons and connections is huge due to the “abstract” nature of visual sentiment. Deep networks need huge amount of less “noisy” labeled training instances to adjust the huge amount of neurons. Otherwise, it will get stuck into local optimum. To build a robust visual sentiment ontology, Borth et al. [6] and Yuan et al. [1] proposed to employ midlevel entities or attributes as features for image sentiment analysis. In [6], 1,200 Adjective Noun Pairs (ANPs), which correspond to different levels of different emotions, are extracted. These ANPs are used as queries to retrieve images from Flickr. Then, pixel-level features of images in each ANP are employed to train 1,200 ANP detectors. The responses of these 1,200 classifiers are finally used as midlevel features for visual sentiment analysis. The work in [1] employed a similar mechanism. The main difference is that 102 scene attributes were used instead. Furthermore, Jou et al. [7] proposed a large-scale Multilingual Visual Sentiment Ontology (MVSO) which is based on VSO to solve the multilingual problem in visual sentiment expression. Campos et al. [8] used a fine-tuned CNN to improve the vision based sentiment predication. By using ANP, Cao et al. [9] proposed a visual sentiment topic model for topic level sentiment detection, and Wang et al. [10] used a bag-of-words model for cross-media sentiment detection. To solve the problem of modeling object-based visual concepts, Chen et al. [11] proposed a hierarchical framework to handle the concept classification in an object specific manner. In order to process the multimodality problem in sentiment learning, Li et al. [12] proposed a multimodal correlation model to build the correlation between modalities. Furthermore, Chen et al. [13] proposed a multimodal hypergraph learning model to bridge modalities of cross-media. You et al. [14, 15] constructed a joint visual-textual sentiment framework which utilized both the state-of-the-art visual and textual sentiment analysis techniques for joint visual-textual sentiment analysis.

For video sentiment analysis, Morency et al. [16] proposed a framework which utilized video sounds and facial expressions to analyze “interview clips.” They focused on the sentiment analysis towards video with fixed contents, similar patterns, and average noises. The experiment results are promising, but, due to the fact that the subject is specified, the method cannot be used to deal with large-scale GIF

videos. Jou et al. [17] proposed to use the features such as color histogram to train a framework for online GIF sentiment analysis. They also proposed a good GIF emotion dataset. However, the labels of dataset lacked temporal sequence information description which is important for understanding how an action in a GIF video yields sentiment. Cai et al. [18] proposed a spatial-temporal visual midlevel ontology and dataset. They constructed a semantic tree to label visual sentiments. However, there is no learning approach which can use those midlevel ontology labels to learn the semantic sequence for GIF sentiment analysis.

In general, the study of GIF sentiment detection is still in the beginning; semantic gap problem and sequence based sentiment understanding problem are the main challenges in this topic.

3. SentiPair Sequence

To solve the semantic gap problem, we propose a middle level sentiment representation named SentiPair Sequence. In the construction of SentiPair, we consider three important criteria: emotional correlation, universality, and detectability. Emotional correlation means that the middle level features should be related to the expressions of sentiment in videos. Universality means that the middle level features should cover most kind of visual sentiment concepts of videos. Detectability means that the middle level features should be able to be detected easily.

3.1. Emotional Correlation. For the first criterion, we introduce the SentiPair Sequence to show why it satisfies emotional correlation. Here, the SentiPair is the joint name of Adjective Noun Pair (ANP) and Verb Noun Pair (VNP). We think that there are two important sentiment expression factors in GIF videos: appearances and motions. Firstly, people often use adjective words to describe the appearances of an object which contain the subjective sentiment of users, like “lovely girl” and “cute dog.” Secondly, the motions of an object are also used to express the dynamic changes of sentiment, like “girl cry” and “girl smile.” To describe the appearances and motions, we use ANPs and VNPs, respectively.

After we obtain ANPs and VNPs, we can form a SentiPair Sequence as follows:

$$\begin{aligned} \text{SentiPair Sequence} &= (\text{Sent } E_1, \text{Sent } E_2, \dots, \text{Sent } E_n), \\ \text{Sent } E_i &= \text{ANP, VNP, Time}(\text{Sent } E_i) \\ &< \text{Time}(\text{Sent } E_j), \quad i < j, \end{aligned} \quad (1)$$

where $\text{Sent } E_i$ is the i th SentiPair and $\text{Time}(\text{Sent } E_i)$ is the time of the i th SentiPair appeared.

The above equation shows that SentiPair Sequence denotes a sequence of appearances and motions under time series. Therefore, it effectively combines two important sentiment expression factors and enriches emotion labels for learning.

More specifically, each SentiPair refers to either a concrete concept like “smile face” or a specific motion like “falling

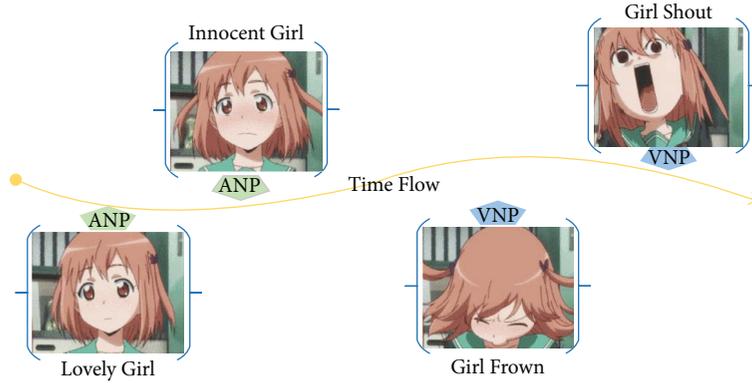


FIGURE 1: An example of SentiPair Sequence.

cup.” In a SentiPair Sequence, SentiPairs are sorted by the order of their occurrence. Figure 1 shows a typical SentiPair Sequence. As we can see, the girl in the video acts differently. In the first frame, the girl was smiling and hence the first SentiPair indicates “Lovely Girl.” In the second frame, the girl looked a bit worried, and the second SentiPair is “Innocent Girl.” With the third SentiPair indicating “Girl Frown,” we can find out that the girl looks sad, which contains a negative sentiment tendency. In the last frame, the girl failed to suppress her feeling and the SentiPair indicates “Girl Shout.” As a result, we can denote the SentiPair Sequence of this GIF video as follows: “Lovely Girl,” “Innocent Girl,” “Girl Frown,” and “Girl Shout.”

SentiPair Sequence describes the concepts associated with sentiment judgment. In general, SentiPair Sequence carries two kinds of concepts. The first one is the existing objects (by ANPs), and the second one is object’s motions (by VNPs).

3.2. Universality. For the second criterion, we introduce Synset Forest to build ANPs and VNPs word set which can cover most words which are semantically related to sentiment. The Synset Forest is a forest which consists of three trees: adjective tree, verb tree, and the noun tree. An example of all three trees can be found at Figure 2. For example, word “smile” mostly denotes the positive sentiment and it belongs to the verb tree, and word “good” also denotes positive sentiment and it belongs to the adjective tree. It shows that all words are organized in a hierarchical tree structure. Furthermore, considering the semantic meaning of each word, our Synset Forest is built from WordNet, a famous lexical database of English. In the WordNet, Synsets are interlinked by means of conceptual semantic and lexical relations. By using WordNet Synsets, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms as Figure 2 shows. Therefore, the proposed Synset Forest models a unified semantic and concept architecture related to sentiment. In the construction of SentiPair, the Synset Forest acts as a collection of candidate words for ANPs and VNPs.

Beyond that, Synset Forest can be used to improve the detection performance of SentiPair by using relative ranking which calculates the semantic distance of two entities in the Synset Forest. Although the SentiPair classification result of

a GIF video is wrong, we can modify the score by considering the semantic relation between each SentiPair.

Relative ranking means that we rerank the score of classification according to semantic relation in the Synset Forest. The calculation formula is shown as follows:

$$\text{Rank}(l) = \lambda_1 \text{Cscore}(l) + \lambda_2 \sum_{l'} \left(\frac{\text{Cscore}(l')}{\text{Semdis}(l', l)} \right)$$

$$\text{Semdis}(l', l) = \sum_{w' \in l', w \in l} \text{Minstep}(w', w) \quad (2)$$

$$w', w \in \text{samesubtree},$$

where label l is a subset of SentiPair words which are selected from Synset Forest, word w is a word in the Synset Forest, $\text{Cscore}(l)$ is the classification score of label l , $\text{Semdis}(l', l)$ is the semantic distance between label l' and l , λ_1 and λ_2 are tuning parameters, and $\text{Minstep}(w', w)$ is the minimum step that costs for working from word w' to w in the Synset Forest. For example, $\text{minstep}(\text{dog}, \text{cat})$ is 2 and $\text{minstep}(\text{dog}, \text{table})$ is 4. In our experiments, λ_1 and λ_2 are set to 0.5.

By using the above equations, we can determine whether a new label l is related to sentiment by calculating the sentiment classification score and the semantic distance between l and a positive sentiment label or negative sentiment label l' . Therefore, we can cover most words which are semantically related to sentiment. For example, if the ground truth label of an image is “cute animal” and the SentiPair detectors predict “cute dog,” by using relative ranking, the score of “cute animal” can be improved through (2). That is because “animal” is close to “dog” in Figure 2. Therefore, the second part of (2) ($\lambda_2 \sum_{l'} (\text{Cscore}(l') / \text{Semdis}(l', l))$) can be improved because the $\text{Semdis}(\text{“cute - dog”}, \text{“cute - animal”})$ is small and $\text{Cscore}(\text{“cute - dog”})$ is high.

3.3. Detectability. For the third criterion, we define two indicators: Sentiment Richness and Sentiment Appearance Probability, to determine which SentiPair has enough sentiment meaning and enough samples to learn.

3.3.1. Sentiment Richness. The calculation of Sentiment Richness comes from the score of SentiWordNet [19]. The score of

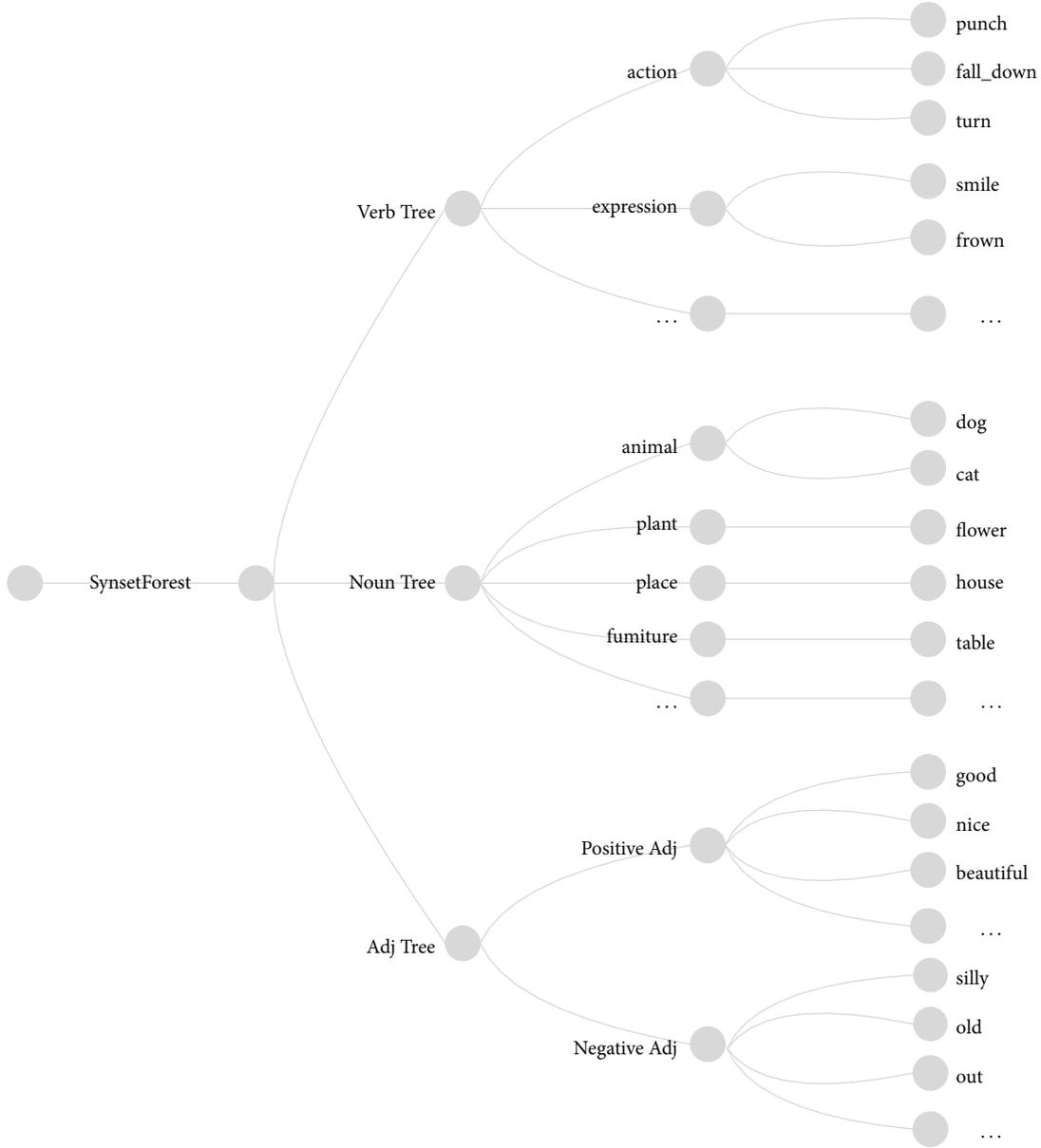


FIGURE 2: An example of Synset Forest.

a SentiWordNet word denotes the sentiment level of the word. The range of SentiWordNet score is $[-1, 1]$, where 0 denotes neutral sentiment and the value close to 1 and -1 means that the word contains more emotional meaning. Therefore, Sentiment Richness is formulated as follows:

$$SR(w) = |\text{SentiScore}(w)|, \quad (3)$$

where $\text{SentiScore}(w)$ is the SentiWordNet score of word w .

3.3.2. Sentiment Appearance Probability. We think that a middle level feature can be detected on the condition of enough samples. Hence, we choose high frequency words which are used to express sentiment. The calculation of Sentiment Appearance Probability is denoted as follows:

$$SAP(w) = \frac{\text{Count}(w)}{\text{MaxCount}}, \quad (4)$$

where $\text{Count}(w)$ denotes the frequency of word w in a famous GIF video website (<https://giphy.com/>) and MaxCount denotes the maximum word frequency in <https://giphy.com/>. <https://giphy.com/> is one of the biggest websites which collects GIF videos with annotations. We think that annotations are helpful in calculating Sentiment Appearance Probability.

The final threshold of choosing SentiPair is shown as follows:

$$\text{Thre}(w) = k_1 SR(w) + k_2 SAP(w), \quad (5)$$

where k_1 and k_2 are tuning parameters. In our experiments, k_1 and k_2 are set to 0.5.

TABLE 1: Distribution of SentiPair words.

Part of speech	Noun	Verb	Adj
WordNet	117,097	11,488	22,141
Selected words	889	91	375

TABLE 2: Examples of selected and unselected words.

Part of speech	Noun	Verb	Adj
Selected words	Man, cat	Laugh, cheer	Lovely, cute
Unselected words	Polka, percussion	Crusade, conscript	Zambian, last

Through the above three criteria, we can cover most words which are semantically related to sentiment and select a suitable word subset as SentiPair labels. Therefore, by learning the middle level features, machine can perceive most kinds of sentiment expressions in GIF videos. Finally, as we can see in Table 1, we select 889 noun words, 91 verb words, and 375 adjective words from WordNet. The examples of selected and unselected words can be seen in Table 2. Those unselected words not only have low scores in SentiWordNet but also have very low frequencies in <https://giphy.com/>. In our examples, although “man” and “cat” have low SentiWordNet scores, they have high frequencies in <https://giphy.com/>.

4. SentiPair Sequence Based Sentiment Detection

To effectively learn the middle level features and understand the GIF video sequences, we propose a two-step learning framework which combines the advantage of two different deep learning neural networks, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). In the first step, it learns how to obtain SentiPair features from GIF videos by using the image learning ability of CNN. In the second step, it learns how to detect GIF video sentiment by using the semantic sequence learning ability of LSTM. The framework of learning is shown in Figure 3.

From bottom to top, firstly, each frame of GIF video is fed into a 7-layer CNN to learn SentiPair features; secondly, the SentiPair Sequence is used as the input for LSTM layer to learn the semantic sequence; and, finally, the output of LSTM layer is used to determine three types of sentiment (positive, negative, and neutral) through a mean pooling layer. The details of this framework are shown in the following subsection.

4.1. Middle Level Features Learning. For the first step, the SentiPair learning is a multilabel learning problem because each image may contain more than one word. To effectively capture the leaning ability of yielding multilabel, we use sigmoid cross-entropy function as loss function in our deep neural network.

$$\text{Loss} = -\frac{1}{N} \sum_{n=1}^N (p_n \log p'_n + (1 - p_n) \log (1 - p'_n)), \quad (6)$$

where N is the number of samples, p_n is the distribution probability of ground truth for a given input, and p'_n is the distribution probability of detection.

In our CNN based SentiPair learning neural network structure, we use 7-layer neural network to learn 1,274 SentiPairs (Figure 4). This framework is similar to the work in [20]. Because we do not have enough data and annotation is a heavy work, we cannot directly use the neural network to learn SentiPairs. To obtain a robust model from few data, we use a large dataset, ImageNet, as supplement. Firstly, we use ImageNet data to learn the basic image features yielded from the 6th full connection layer. Then we fix the parameters from layer 1 to layer 6 and change the output layer from object vector to SentiPair vector. Finally, after training, layer 7 learns a mapping function from image features to SentiPair features.

4.2. Sentiment Sequence Learning. For the second step, to learn sentiment from a SentiPair Sequence, we use LSTM model which is often used to model a semantic sequence in text expression. LSTM ([21]) is a classical recurrent neural network. The best advantage of LSTM is that it alleviates the problem of gradient diffusion and explosion in sequence learning. Therefore, it can learn the long dependencies in a sequence by a memory unit and three-gate mechanism. Formally, the update formulas of LSTM are shown as follows:

$$\begin{aligned} i^{(t)} &= \sigma(W_i x^{(t)} + U_i h^{(t-1)}) \\ f^{(t)} &= \sigma(W_f x^{(t)} + U_f h^{(t-1)}) \\ o^{(t)} &= \sigma(W_o x^{(t)} + U_o h^{(t-1)}) \\ c'^{(t)} &= \tanh(W_c x^{(t)} + U_c h^{(t-1)}) \\ c^{(t)} &= f^{(t)} c^{(t-1)} + i^{(t)} c'^{(t)} \\ h^{(t)} &= o^{(t)} \tanh(c^{(t)}), \end{aligned} \quad (7)$$

where $i^{(t)}$, $f^{(t)}$, $o^{(t)}$, and $c^{(t)}$ are input gate, forget gate, output gate, and memory cell activation vector at time-step t , respectively, $h^{(t)}$ is the hidden vector, and $W_i, W_f, W_o, W_c, U_i, U_f, U_o$, and U_c are training parameters.

Because the output dimension of LSTM increases with the increase of GIF video sequences, for each hidden vector $h^{(t)}$, we feed them into a mean pooling layer to reduce the

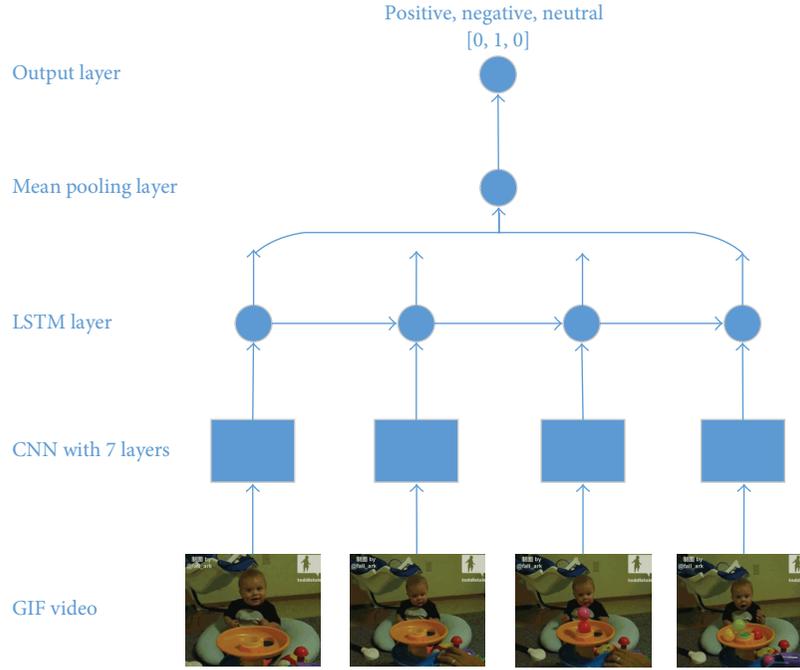


FIGURE 3: SentiPair Sequence based GIF sentiment learning framework.

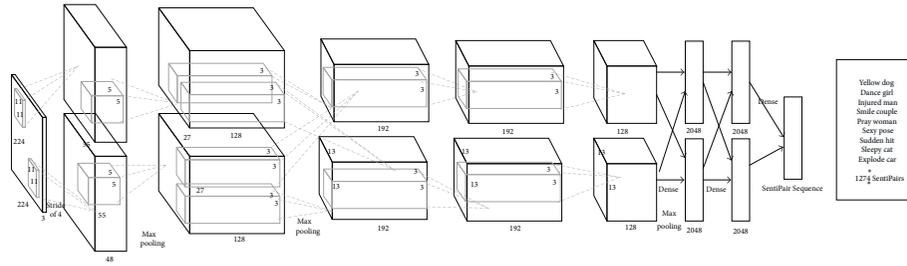


FIGURE 4: Network structure of SentiPair detection.

dimension of LSTM output as shown in Figure 3. The mean value is calculated as follows:

$$\text{Mean}^{(i)} = \frac{\sum_t^{t \pm ws} h^{(t)}}{ws}, \quad (8)$$

where ws is the window size of mean pooling. In our experiment, ws is set to be the maximum length of LSTM.

After mean pooling layer, all values are fed into softmax layer to determine the final sentiment: positive, negative, and neutral. The loss function is defined as follows:

$$\text{Loss} = -\frac{1}{N} \sum_{n=1}^N \log \frac{e^{z_{nk}}}{\sum_{i=1}^M e^{z_{ni}}}, \quad (9)$$

where N is the number of samples, z_{nk} is the detection results of sample n , and M is the number of labels.

5. Experiment

In this section, we design several experiments to verify our framework and compare the performance with other state-of-the-art algorithms. As the main contribution of our work

is a SentiPair Sequence based GIF video sentiment detection, the first experiment is conducted to show the SentiPair detection performance. Secondly, in the second experiment, four state-of-the-art bag-of-words machine learning methods are compared with our approach in both SentiPair Sequence and ANP features. Finally, some examples are compared between SentiPair Sequence and ANP features.

5.1. Experiment Setting. Since there is no suitable GIF videos datasets which are labeled with SentiPairs, we construct a new labeled dataset named GSO-2016 to train the sentiment classifiers. The GIF videos of GSO-2016 dataset came from one of the most popular microblogs. All GIF videos were posted by online users and were collected automatically. We recruited 7 workers who are undergraduate students in our university. Each worker was shown one GIF video and was expected to accomplish two tasks. Task 1 is to select suitable words from Synset Forest and form a SentiPair Sequence description for a given GIF. To be more specific, for each GIF, SentiPairs were chosen by browsing the words and tree structure of Synset Forest. Each SentiPair consists of either an adjective and a noun (ANP) or a verb and a noun (VNP). For

TABLE 3: Examples of GSO-2016 dataset. (+, -, and 0 denote positive, negative, and neutral sentiment, resp. We give two main images to show the meaning of a GIF sequence.)

GIF sequence	SentiPair	Sentiment
	shy,cat; spread_over,face	+
	sad,man; combust,money	-
	two,man; lie,man	0

example, in the second row of Table 3, the given GIF video was labeled with “sad man” (ANP) and “combust money” (VNP) according to the GIF video sequence. In Task 2, workers were expected to give an overall sentiment judgment (positive, negative, and neutral) for each image. For example, in the second row of Table 3, the GIF video is labeled with negative sentiment.

In GSO-2016, we provide labeled and unlabeled GIF videos for supervised learning, one-shot learning, and unsupervised learning. There are totally 36,039 GIF videos and 1,874 GIF videos were labeled with SentiPairs and three kinds of sentiment (positive, negative, and neutral). The dataset can be downloaded from our website (<https://pan.baidu.com/s/1hrJBSAo>). Three examples of positive, negative, and neutral sentiment GIF videos are shown in Table 3. We hope that dataset can promote the development of semantic vision understanding.

The labeled data in GSO-2016 dataset consists of 1,111 positive instances (59.2%), 164 negative instances (8.8%), and 599 neutral instances (32%). The evaluation metric in our experiment is sentiment detection accuracy. In our experiment, we use 80% and 20% labeled data as training

set and test set, respectively. The distribution of experiment dataset is shown in Table 4.

5.2. Experiment Result and Analysis

5.2.1. SentiPair Experiment. In this experiment, we have tried three methods of SentiPair detection: 7-layer CNN with single label, 7-layer CNN with single label and relative ranking (see (2)), and 7-layer CNN with multilabel and relative ranking. Single label means that we just choose the first SentiPair label in training and yield multilabel in testing according to the score of labels. Multilabel means that we use multilabel in training according to the constraint of the sigmoid cross-entropy function. The experiment results are shown in Figure 5. Relative ranking are calculated according to (2).

In Figure 5, Top n means the highest scores of n labels according to the rank scores of classification. According to the threshold of choosing SentiPairs (0.7, 0.8, 0.9), we obtain 31463, 5111, and 1274 SentiPairs, respectively. From the results, we can obtain the following conclusions: (1) Multilabel learning with relative ranking achieves the best detection performance in 1,274 SentiPairs. Multilabel learning with

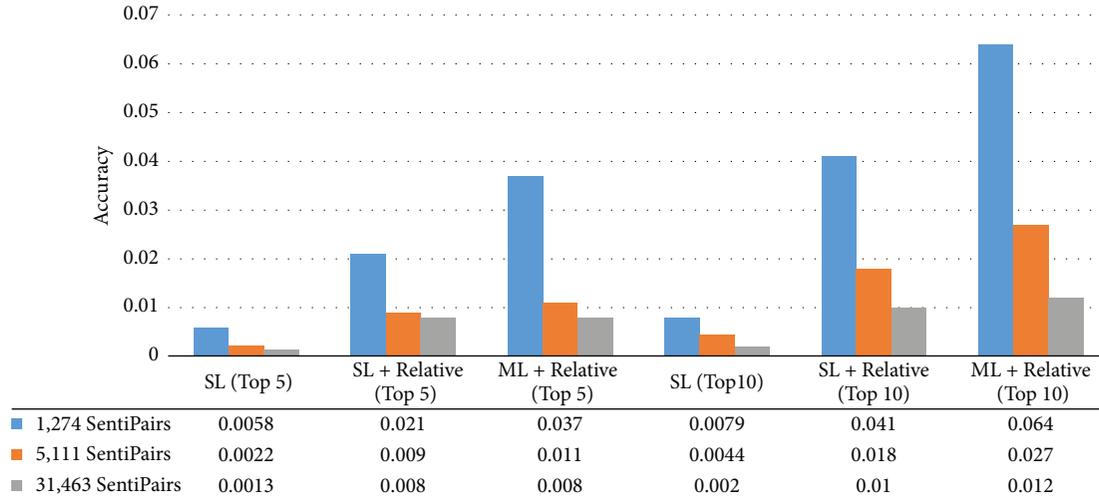


FIGURE 5: Performance of SentiPair detection (SL means single label, ML means multilabel, and Relative means relative ranking).

TABLE 4: Labeled dataset distribution.

	Training data	Test data
Positive	888	223
Negative	131	33
Neutral	479	120

relative ranking obtains 1.6% and 2.3% improvement in 1,274 SentiPairs compared with single label and relative ranking method in Top 5 and Top 10 results, respectively. Considering that the accuracy of 7-layer CNN with single label and relative ranking is 2.1% and 4.1% in Top 5 and Top 10 results, respectively, our improvement is significant. (2) The accuracy increases with the decrease of SentiPair number.

5.2.2. Sentiment Prediction Experiment. In this experiment, to show the performance of SentiPair Sequence based GIF sentiment detection, we compared our approach with four state-of-the-art classification methods (SMO, Naive Bayes, AdaBoost, and Logistic Regression) in the condition of two different middle level features (SentiPair and ANP). All four state-of-the-art classification methods used ANPs and SentiPairs through bag-of-words model and that means that they cannot use GIF sequence information. In this experiment, we choose the same learning structure (Figure 3) with only 1,874 labeled GIF videos as our baseline to show the effectiveness of middle level features. ANP detectors were trained by using AlexNet in more than 500,000 images from Flickr. SentiPair detectors were trained by using the 7-layer CNN neural network in 1,874 labeled GIF videos from GSO-2016 dataset and a large number of unlabeled images from ImageNet. The experiment results are shown in Figure 6.

From the results, we can obtain the following conclusions: (1) middle level features (SentiPair and ANP) outperform low level features (raw data); it indicates that middle level features are more robust than low level features in representing the visual sentiment. (2) LSTM outperforms the other four state-of-the-art classification methods without learning GIF

TABLE 5: The distribution of confused SentiPairs which appear in more than one kind of sentiment.

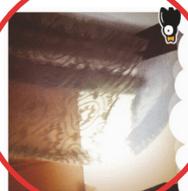
	Number of confused SentiPairs
Positive + negative + neutral	39
Positive + negative	22
Positive + neutral	136
Negative + neutral	16

sequence in both SentiPairs and ANPs; it indicates that LSTM effectively learns the impact of the time sequence information in expressing sentiment. (3) SentiPairs outperform ANPs in four learning methods except Logistic Regression; it shows that SentiPairs are better than ANPs in sentiment learning and it also indicates that the combination of ANPs and VNPs is helpful in GIF video sentiment detection.

After we compared the experiment results in Figures 5 and 6, we can see that although the prediction performance of SentiPair is bad, it still improves the sentiment prediction results. In our SentiPair experiments, there are only 1,874 SentiPairs labeled GIF videos in GSO-2016 and more than 1,274 SentiPairs need to be learned. As a consequence, it is hard to achieve a good enough SentiPair detection performance. However, in this situation, those SentiPairs are strongly related to three kinds of sentiment. According to our results from 1,274 SentiPairs in Table 5, there are only 16.7% SentiPairs which are related to more than two kinds of sentiments, and 83.3% SentiPairs are strongly related to one kind of sentiment. Although the SentiPair prediction is wrong, the prediction results of a GIF video and the ground truth labels are related to the same sentiment with a high probability.

5.2.3. Case Study. To further compare the details of SentiPairs and ANPs in sentiment detection, we show some cases in Table 6. In this table, pictures in red circles are incorrectly classified both in ANP based and in SentiPair based approaches and pictures in green boxes demonstrate

TABLE 6: SentiPairs versus ANPs (pictures in red circles are incorrectly classified both in ANP based and in SentiPair based approaches and pictures in green boxes demonstrate that SentiPairs outperform ANPs in GIF videos).

GIF Videos						
Ground truth (SentiPair)	Yellow dog, run dog	Shy cat, spread_over face	One boy, laugh boy	Lovely girl, pull_a_face face	Happy face, sing song	Angry face, throw notebook
ANP prediction	Little dog	Lovely girl	One boy	Beautiful girl	Happy man	Sad man
SentiPair prediction	Little dog, run dog	Lovely girl, Laugh girl	One boy, smile boy	Beautiful girl, smile girl	Happy man, sing man	Sad man, hit man
Ground truth (Sentiment)	Neutral	Positive	Positive	Positive	Positive	Negative
Sentiment prediction (ANP only)	Neutral	Positive	Neutral	Positive	Positive	Negative
Sentiment prediction (SentiPair)	Neutral	Positive	Positive	Positive	Positive	Negative
GIF videos						
Ground truth (SentiPair)	Cute snail, gallop snail	Injured man, weep boy	Some people, walk man	Dance girl, dark house	Fall man, skate man	Fly bird, move curtain
ANP prediction	Cute dog	Scared boy	Young people	Tearful face	Cool man	Lovely kid
SentiPair prediction	Cute dog, run dog	Scared boy, cry boy	Young people, walk man	Tearful face, dance girl	Cool man, walk man	Lovely kid, dance kid
Ground truth (Sentiment)	Positive	Negative	Neutral	Positive	Positive	Neutral
Sentiment prediction (ANP only)	Positive	Negative	Positive	Negative	Positive	Positive
Sentiment prediction (SentiPair)	Positive	Negative	Positive	Positive	Positive	Positive

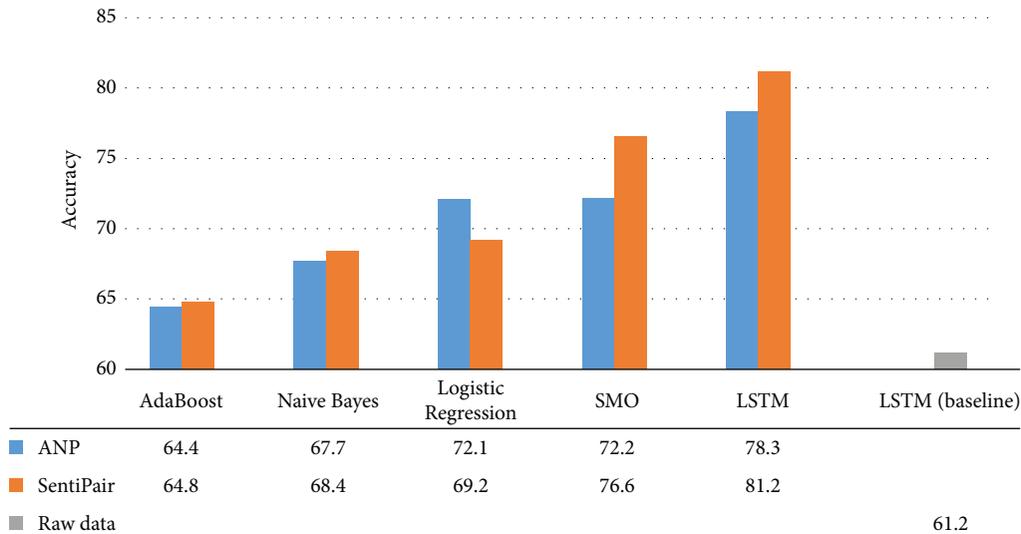


FIGURE 6: Performance of sentiment detection for GIF video. (Raw data denotes only 1,874 labeled GIF videos were used in experiment.)

that SentiPairs outperform ANPs in GIF videos. Although the SentiPair predictions are bad, there are semantic relations between ground truth labels and our predictions and they yield a good sentiment classification result. For example, “smile boy” (prediction) is similar with “laugh boy” (ground truth) because “smile” and “laugh” have similar semantic meaning; “cry boy” (prediction) is similar to “weep boy” (ground truth) because “cry” and “weep” both describe the action of crying; “cute dog” (prediction) is similar to “cute snail” (ground truth) because “dog” and “snail” both belong to “animal.” Furthermore, by combining the advantages of ANPs and VNPs, SentiPairs can outperform the ANP based approach. For example, Table 6 shows two examples with green boxes which show the advantage of SentiPair based approach. Although the predicted ANPs, “one boy” and “tearful face,” show neutral and negative sentiment, respectively, our predicted SentiPairs obtain “smile boy” and “dance girl” to modify the wrong sentiment.

6. Conclusion

GIF video sentiment detection is a challenge. Considering the function of GIF video sequence and motion in sentiment expression, in this paper, we propose a SentiPair Sequence based approach for GIF video sentiment detection. The SentiPair Sequence not only bridges the low level image features and high level sentiment semantic spaces but also supervises the learning process to learn the sentiment expression for motions and video sequences. The experiments suggest that the prediction accuracy is 81.2% which is significant for the other four state-of-the-art classification methods and the state-of-the-art middle level features, ANP. We also released our dataset GSO-2016 to the public. GSO-2016 contains 1,874 manually labeled GIF videos selected from more than 30,000 candidates. Each video was labeled with both sentiments and SentiPair Sequences. We believe it will be helpful for further research. Furthermore, the performance of SentiPair detection is not good enough to help in sentiment detection;

how to enhance the CNN is one of the important issues in our future works.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the Nature Science Foundation of China (no. 61402386, no. 61305061, no. 61502105, no. 61572409, no. 81230087, and no. 61571188), Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (no. MJUKF201743), Education and Scientific Research Projects of Young and Middle-Aged Teachers in Fujian Province under Grant no. JA15075, and Fujian Province 2011 Collaborative Innovation Center of TCM Health Management and Collaborative Innovation Center of Chinese Oolong Tea Industry—Collaborative Innovation Center (2011) of Fujian Province.

References

- [1] J. Yuan, Q. You, S. McDonough, and J. Luo, “SentrIBUTE: Image sentiment analysis from a mid-level perspective,” in *proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '13)*, ACM, August 2013.
- [2] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP '02)*, vol. 10, pp. 79–86, Association for Computational Linguistics, July 2002.
- [3] L. Zhou, B. Li, W. Gao, Z. Wei, and K.-F. Wong, “Unsupervised discovery of discourse relations for eliminating intra-sentence

- polarity ambiguities,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pp. 162–171, Association for Computational Linguistics, 2011.
- [4] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 168–177, ACM, August 2004.
- [5] Q. You, J. Luo, H. Jin, and J. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” in *Proceedings of the 29th AAAI Conference on Artificial Intelligence (IAAI '15)*, pp. 381–388, January 2015.
- [6] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *proceedings of the 21st ACM International Conference on Multimedia (MM '13)*, pp. 223–232, October 2013.
- [7] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, “Visual affect around the world: a large-scale multilingual visual sentiment ontology,” in *proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*, pp. 159–168, October 2015.
- [8] V. Campos, A. Salvador, X. Giró-I-Nieto, and B. Jou, “Diving deep into sentiment: understanding fine-tuned CNNs for Visual sentiment prediction,” in *Proceedings of the 1st International Workshop on Affect and Sentiment in Multimedia (ASM '15)*, pp. 57–62, ACM, 2015.
- [9] D. Cao, R. Ji, D. Lin, and S. Li, “Visual sentiment topic model based microblog image sentiment analysis,” *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 8955–8968, 2016.
- [10] M. Wang, D. Cao, L. Li, S. Li, and R. Ji, “Microblog sentiment analysis based on cross-media bag-of-words model,” in *proceedings of the 6th International Conference on Internet Multimedia Computing and Service (ICIMCS '14)*, pp. 76–80, July 2014.
- [11] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, “Object-based visual sentiment concept analysis and application,” in *proceedings of the 2014 ACM Conference on Multimedia (MM '14)*, pp. 367–376, November 2014.
- [12] L. Li, D. Cao, S. Li, and R. Ji, “Sentiment analysis of Chinese micro-blog based on multi-modal correlation model,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP '15)*, pp. 4798–4802, Quebec City, Canada, September 2015.
- [13] F. Chen, Y. Gao, D. Cao, and R. Ji, “Multimodal hypergraph learning for microblog sentiment prediction,” in *proceedings of the IEEE International Conference on Multimedia and Expo (ICME '15)*, July 2015.
- [14] Q. You, J. Luo, H. Jin, and J. Yang, “Joint visual-Textual sentiment analysis with deep neural networks,” in *proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*, pp. 1071–1074, October 2015.
- [15] Q. You, J. Luo, H. Jin, and J. Yang, “Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia,” in *proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM '16)*, pp. 13–22, February 2016.
- [16] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis: harvesting opinions from the web,” in *proceedings of the 2011 ACM International Conference on Multimodal Interaction (ICMI '11)*, pp. 169–176, November 2011.
- [17] B. Jou, S. Bhattacharya, and S.-F. Chang, “Predicting viewer perceived emotions in animated GIFs,” in *proceedings of the ACM Conference on Multimedia (MM '14)*, pp. 213–216, November 2014.
- [18] Z. Cai, D. Cao, D. Lin, and R. Ji, “A spatial-temporal visual mid-level ontology for GIF sentiment analysis,” in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '16)*, pp. 4860–4865, IEEE, Vancouver, Canada, July 2016.
- [19] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining,” in *Proceedings of the International Conference on Language Resources and Evaluation (Lrec '10)*, pp. 83–90, Valletta, Malta, May 2010.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, no. 2, p. 2012, 2012.
- [21] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.



Hindawi

Submit your manuscripts at
<https://www.hindawi.com>

