Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

NLWSNet: a weakly supervised network for visual sentiment analysis in mislabeled web images*

Luo-yang XUE[1], Qi-rong MAO†‡[1,2], Xiao-hua HUANG[3,4], Jie CHEN[1]

[1]*Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China*

[2]*Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, Zhenjiang 212013, China*

[3]*School of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167, China*

[4]*Center for Machine Vision and Signal Analysis, University of Oulu, Oulu 8000, Finland*

†E-mail: mao_qr@ujs.edu.cn

Received Nov. 12, 2019; Revision accepted Feb. 3, 2020; Crosschecked July 20, 2020

Abstract: Large-scale datasets are driving the rapid developments of deep convolutional neural networks for visual sentiment analysis. However, the annotation of large-scale datasets is expensive and time consuming. Instead, it is easy to obtain weakly labeled web images from the Internet. However, noisy labels still lead to seriously degraded performance when we use images directly from the web for training networks. To address this drawback, we propose an end-to-end weakly supervised learning network, which is robust to mislabeled web images. Specifically, the proposed attention module automatically eliminates the distraction of those samples with incorrect labels by reducing their attention scores in the training process. On the other hand, the special-class activation map module is designed to stimulate the network by focusing on the significant regions from the samples with correct labels in a weakly supervised learning approach. Besides the process of feature learning, applying regularization to the classifier is considered to minimize the distance of those samples within the same class and maximize the distance between different class centroids. Quantitative and qualitative evaluations on well- and mislabeled web image datasets demonstrate that the proposed algorithm outperforms the related methods.

Key words: Visual sentiment analysis; Weakly supervised learning; Mislabeled samples; Significant sentiment regions

https://doi.org/10.1631/FITEE.1900618 **CLC number:** TP391.4

1 Introduction

Recent developments in deep convolutional neural networks (CNNs) have led to great success in a variety of visual sentiment analysis (VSA) tasks, including sentiment classification (Yang et al., 2017b; Zhao et al., 2017; Zhang FF et al., 2018b), sentiment dimension prediction (Zhao et al., 2016), sentiment region detection (Yang et al., 2018a, 2018b), and others (Jia et al., 2018; Zhang QS and Zhu, 2018). The success is generally driven by the availability of large-scale well-annotated sentiment datasets (Chen SX et al., 2018a; Zeng et al., 2018) like Flickr and Instagram (FI) (You et al., 2016) and Flickr (Borth et al., 2013). However, annotating a massive number of sentiment images is extremely labor-intensive and costly due to the high level of subjectivity in the human recognition process (Fang et al.,

‡ Corresponding author

* Project supported by the Key Project of the National Natural Science Foundation of China (No. U1836220), the National Natural Science Foundation of China (No. 61672267), the Qing Lan Talent Program of Jiangsu Province, China, the Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, China, the Finnish Cultural Foundation, the Jiangsu Specially-Appointed Professor Program, China (No. 3051107219003), the Jiangsu Joint Research Project of Sino-Foreign Cooperative Education Platform, China, and the Talent Startup Project of Nanjing Institute of Technology, China (No. YKJ201982)

ORCID: Luo-yang XUE, https://orcid.org/0000-0003-2674-4051; Qi-rong MAO, https://orcid.org/0000-0002-0616-4431

2018). Table 1 shows the available manually annotated affective datasets used for sentiment analysis. It is observed that most of the available datasets contain less than 2000 samples. To reduce the heavy lift in annotation, an alternative approach obtains sentiment annotations of images using an image search engine, e.g., Google or Flickr Image Search.

Table 1 Available affective datasets

Dataset	Category	Num
IAPSa (Mikels et al., 2005)	8	395
Abstract (Machajdik and Hanbury, 2010)	8	228
ArtPhoto (Machajdik and Hanbury, 2010)	8	806
Twitter I (Borth et al., 2013)	2	603
Twitter II (You et al., 2015)	2	1269
EmotionROI (Peng KC et al., 2016)	6	1980
Flickr (Katsurai and Satoh, 2016)	2	60 745
Instagram (Katsurai and Satoh, 2016)	2	42 856

Most of the datasets contain less than 2000 samples. Num: number of images

Specifically, for VSA, most image search engines use keywords such as happy and anger as the queries. The connection between keywords and images is established based on co-occurrences between the web image and its surrounding text. However, query keywords may not be consistent with the visual content of the target sentiment. Thus, annotations of web images obtained using a search engine inevitably contain noisy information (Fig. 1). These noisy labels will provide negative information to a classifier if we use them directly to train the classifier for VSA. However, much research (You et al., 2017; Yang et al., 2018b) focused only on weakly supervised learning for local sentiment region discovery on datasets with correct image-level labels, but did not directly use mislabeled sentiment datasets as training samples.

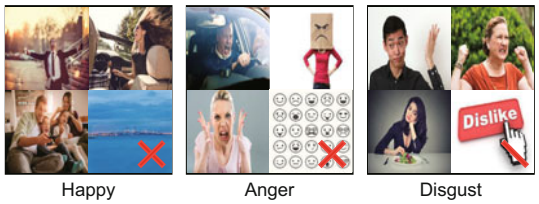

Fig. 1 Results retrieved from the image engine by keywords

Some of the results could not correctly represent the corresponding sentiment

To address the above-mentioned problem, we propose a weakly supervised network (NLWSNet), which is robust to label noise and thus can directly use web images for the training process. The NLWSNet process has three stages, i.e., reducing negative effect of mislabeled samples, discovering class-special and local significant regions, and obtaining good performance of the sentiment classifier.

Specifically, to suppress the negative influence caused by samples with incorrect labels during the training process, we propose a non-extreme channel-attention (NECA) module, which is significantly different from traditional attention modules. Importantly, the NECA module retains the effectiveness of traditional attention mechanisms while avoiding the influence of extreme distribution in attention values. This strategy can guarantee that the feature maps obtained are generally effective and that the attention values do not lead to information loss.

In the second stage, we consider the sentiment in one sample to be intricate. Therefore, we design a spatial-class activation map (SCAM) module to acquire special-class sentiment features. In particular, the framework is shared before the SCAM module. However, in the SCAM module, we allocate an independent two-class classifier for each sentiment category. Thus, the approach of sentiment analysis is translated into multiple and simple sentiment detection problems (Chen SX et al., 2017, 2018b). Meanwhile, we apply the class activation map (CAM) (Zhou et al., 2016) in each sentiment detection process. After the SCAM module, the special-class and discriminative regions are employed for final sentiment classification.

Additionally, we apply regularization to the final classification to achieve good performance. Specifically, we apply center loss and triple loss as the regularization, which is an approach widely used in the face recognition field. Through the proposed regularization, the final classifier minimizes intra-class distances while maximizing the distances between samples in different categories.

As mentioned above, our contributions in this study are summarized as follows:

1. Annotation cost for training samples is reduced in our system because mislabeled web images can be used directly for the training process by the proposed NECA module, which will reduce the negative influence of mislabeled samples in sentiment recognition.

2. We propose the SCAM module to stimulate

the network to discover the special-class and significant regions, which may assist complicated sentiment analysis.

3. We introduce the regularization for visual sentiment classification to effectively learn the discriminative representation, which could be favorable in sentiment classification.

2 Related works

2.1 Visual sentiment analysis

Most existing approaches for VSA can be divided roughly into shallow modeling methods and deep modeling methods.

1. Shallow modeling methods

Machajdik and Hanbury (2010) defined a combination of rich hand-crafted features based on art and psychology theory. Borth et al. (2013) proposed a mid-level concept, i.e., adjective noun pairs (ANPs), to detect image concepts instead of expressing the sentiments directly. Li et al. (2018) computed the weighted sum of the textual sentiment values of ANPs to describe an image. Yuan et al. (2013) proposed Sentribute, an image-sentiment analysis algorithm, which can easily interpret high-level understanding. Chen T et al. (2014b) built object detection models to detect six kinds of objects and proposed classification models to handle the similarity between visual sentiment concepts.

These shallow modeling methods have been proven to be effective on several small datasets, whose images are selected from a few special domains, e.g., abstract paintings and art photos (Machajdik and Hanbury, 2010).

2. Deep modeling methods

In recent years, many deep neural networks have been widely used in visual recognition systems in many fields (Girshick et al., 2014; Krizhevsky et al., 2017; Ou et al., 2018; Zhang FF et al., 2018a). The advantage of these models is that they use the back-propagation algorithm (LeCun et al., 1989) to learn high-level features from the original data input. At the same time, they use the manual features calculated by traditional recognition methods as preprocessing steps (Zhang N et al., 2014). Chen T et al. (2014a) constructed a visual sentiment classification model named DeepSentiBank, which demonstrated significantly improved classification accuracy and retrieval performance. In addition, some methods incorporate the model weights learned from large-scale universal datasets (Deng et al., 2009) and fine-tune the CNN for VSA (Campos et al., 2015, 2017).

As mentioned above, due to the expense of manual annotation of sentiment labels, the number of samples in most existing affective datasets is smaller than 2000. The scale of most datasets is thus far from the required scale for training robust deep models. Note that in this study we focus on using those samples with noisy labels to directly train the deep models. The scale of the datasets can be partially enlarged, and can be employed to improve the robustness of deep models.

2.2 Weakly supervised learning

Weakly supervised learning aims to learn a model using samples with a limited number of labels. It is widely used in different computer vision tasks, such as object detection and semantic segmentation. He and Peng (2017) proposed a weakly supervised part selection method with two spatial constraints to promote part selection and achieve the best results. Oquab et al. (2015) learned the interaction between humans and objects purely from action labeling. He et al. (2019) applied a multi-level attention mechanism to guide discriminative localization learning and an end-to-end discriminative localization network to localize discriminative regions, which not only achieved a notable classification performance but also improved the classification speed. For the phrase grounding task, Rohrbach et al. (2016) adopted an attention mechanism optimized by reconstruction of query information and avoided bounding-level labels for each query. Based on this, Xiao et al. (2017) used a continuous attention map and explored the detailed structural reconstruction of language modalities. Zhu Y et al. (2017) proposed a soft proposal network (SPN) to generate soft proposals and aggregated image-specific patterns by coupling the proposal and feature maps. There are few works focusing on sentiment analysis handled in a weakly supervised way. For VSA, Yang et al. (2018b) presented a weakly supervised coupled network (WSCNet) to integrate the detection and classification branches into a unified deep framework.

Most existing weakly supervised learning methods need correct image-level labels, and their tasks

are object detection or semantic segmentation. However, in this study, the image-level labels of the samples are noisy, and our task is to train an effective classification model without correct image-level labels, which could lead to labor savings.

2.3 Attention mechanism

Humans selectively use an important part of a sample to make a decision (Itti et al., 1998; Corbetta and Shulman, 2002). This strategy is called an attention mechanism and has been widely used in various fields (Zhu YK et al., 2016; Zagoruyko and Komodakis, 2017; Liu et al., 2018; Yu et al., 2018).

In many methods, researchers use the attention mechanism to increase the accuracy of the CNN classification model. Methods are divided mainly into spatial-attention ones and channel-attention ones. Specifically, in residual attention networks (RANs), Wang et al. (2017) used a three-dimensional (3D) spatial-attention map to enhance recognition accuracy. In the squeeze-and-excitation (SE) method, Hu et al. (2018) applied only a one-dimensional (1D) channel-attention map to accomplish the same purpose. The bottleneck attention module (BAM) (Park et al., 2018) and convolutional block attention module (CBAM) (Woo et al., 2018) increase the accuracy of the classifier using both 1D channel-attention maps and two-dimensional (2D) spatial self-attention maps. These methods all require additional trainable parameters to obtain an attention map. Peng YX et al. (2019a) proposed a visual-textual bi-attention mechanism to distinguish the fine-grained information with different levels of saliency detection from both the local and relation levels. Zhuang et al. (2017) calculated the ratio of each value within the whole feature map to obtain the spatial attention map. Peng YX et al. (2019b) proposed the spatial-temporal attention model to emphasize the salient regions of the frame for video classification. The advantage of these methods is that additional parameters are not required in the training process.

The attention mechanisms mentioned above obtain good performance in classification because they discover the significant differences. However, by training sentiment classification with mislabeled samples, these attention mechanisms could result in over-fitting due to the extreme distribution of the attention values. At the same time, the extreme distribution of the attention values introduces information loss in the correctly labeled samples, which adds negative influence to the sentiment classification. Inspired by Hu et al. (2018), our method provides non-extreme distribution of attention values, which is suitable for classification with mislabeled samples.

3 Method

NLWSNet is composed mainly of three stages (Fig. 2). In the first stage, the negative effect of mislabeled samples is suppressed using the NECA module. In the second stage, the SCAM module is proposed to automatically discover sentiment-specific regions. In the last stage, regularization is applied with the coupled center loss and triple loss for classification to learn the discriminative representation.

3.1 Non-extreme channel attention mechanisms

In Fig. 2, we consider that each sample is represented as an array of features. Let $\boldsymbol{X}^n_{i,j,m} \in \mathbb{R}^c$ denote the last convolutional layer activations from the n^{th} sample with the m^{th} convolution kernel at the spatial location, where (i, j) are the coordinates of the feature maps ($i = 1, 2, \ldots, d$, $j = 1, 2, \ldots, d$, and d is the height or width of the feature maps) and c is the number of sentiment categories. After acquiring the feature maps, we use the sigmoid function $f_{\text{sigmoid}}(x) = 1/(1 + \mathrm{e}^{-x})$ as the nonlinear activation function to obtain the 1D attention values:

$$S_m = f_{\text{sigmoid}}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}_{i,j,m} + \boldsymbol{b}), \tag{1}$$

where $\boldsymbol{W}$ and $\boldsymbol{b} \in \mathbb{R}^c$ denote the weight set and biases of the attention detector, respectively. To suppress the attention values in the feature map caused by the incorrectly labeled samples, we normalize the attention scores to [0, 1] to aggregate the feature map:

$$a_m = \frac{\exp(S_m)}{\sum_m \exp(S_m)}, \tag{2}$$

where a_m is the 1D attention score and $\sum_m a_m = 1$. Afterwards, we can obtain the attention map representation $A^n_{i,j,m}$ as follows:

$$A^n_{i,j,m} = \boldsymbol{X}^n_{i,j,m} \circ a^n_m, \tag{3}$$

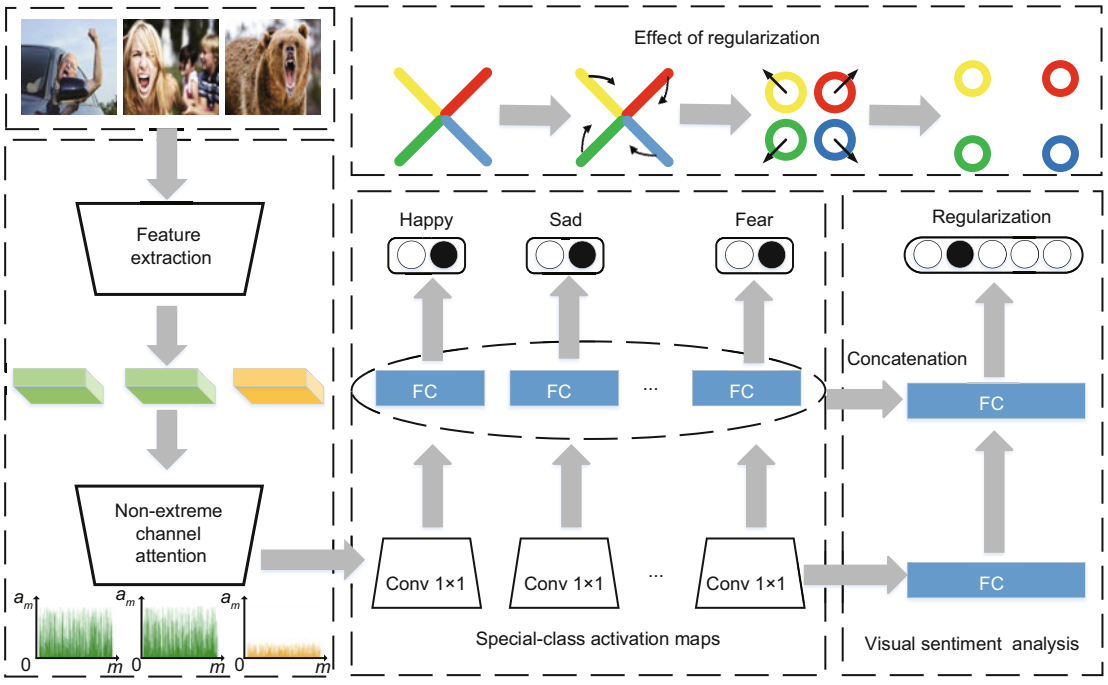

Fig. 2 Architecture of the weakly supervised learning network (NLWSNet)

NLWSNet incorporates a non-extreme channel-attention module, a spatial-class activation map module, and classification regularization. It can reduce the annotation cost tremendously and enhance the robustness of sentiment analysis on mislabeled web images. FC: fully connected layer; Conv: convolutional layer. Feature maps of correctly and incorrectly labeled samples are in green and orange, respectively. References to color refer to the online version of this figure

where "$\circ$" denotes the element-wise multiplication. For comparison with traditional attention mechanisms, we list their methods as follows:

$$a_m = f_{\text{softmax}}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^n_{i,j,m} + \boldsymbol{b}), \tag{4}$$

$$a_{i,j,m} = f_{\text{softmax}}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^n_{i,j,m} + \boldsymbol{b}), \tag{5}$$

$$a_{i,j,m} = \frac{\exp(\boldsymbol{X}^n_{i,j,m})}{\sum_i \sum_j \exp(\boldsymbol{X}^n_{i,j,m})}, \tag{6}$$

where $f_{\text{softmax}}(x) = \mathrm{e}^x / \sum \mathrm{e}^x$. Eqs. (4), (5), and (6) (Chen L et al., 2017; Zhuang et al., 2017) show the channel-wise attention, spatial attention, and the usual way to normalize attention values, respectively. As mentioned in Inception-V2 (Szegedy et al., 2016) and RAN, $f_{\text{softmax}}(x)$ may result in overfitting because the model becomes too confident about its predictions.

Thus, especially when the labels of web images are noisy, the extreme distribution of the attention scores affects the learned recognition model. Compared with traditional attention mechanisms, our method avoids the above-mentioned problem by using $f_{\text{sigmoid}}(x)$ in advance.

On the other hand, the channel attention scores in green are generally larger than those in orange in Fig. 2. Attention scores in green from the correctly labeled samples can obtain the larger attention values from the self-attention NECA module, and can help learn how to distinguish feature representations.

3.2 Special-class activation map

Because the sentiment in one sample can be complicated, it is difficult for the classifier to directly analyze the difference between various emotions. Therefore, in the SCAM module, sentiment analysis is turned into multiple sentiment detection problems. In each sentiment detection problem, we apply the class activation map approach to obtain local sentiment regions, and this has significant contributions to VSA.

Specifically, after the process of NECA, global average pooling (GAP) outputs the spatial average $V_m^{c_k}$ of the attention map of each kernel at one 1×1 convolutional layer, where m is the number of channels and c_k the corresponding sentiment detection classifier. In Fig. 3, the sentiment activation map

$M_{i,j,m}^{c_k}$ is generated using $V_m^{c_k}$ as the weight of the response attention map:

$$M_{i,j,m}^{c_k} = \sum_m V_m^{c_k} A_{i,j,m}. \tag{7}$$

Then, given the number of classes C, we allocate two-class classifiers to C by the sigmoid function, where the input to each special-class classifier c_k is the corresponding sentiment activation map $M_{i,j,m}^{c_k}$. In this process, each two-class classifier in the SCAM module captures the salient regions evoking the corresponding sentiment. Then, we couple all the local sentiment maps $\sum M_{i,j,m}^{c_k}$ with the holistic feature map $A_{i,j,m}$ as follows:

$$C_{i,j,m} = [A_{i,j,m} : M_{i,j,m}^{c_1} : M_{i,j,m}^{c_2} : \ldots : M_{i,j,m}^{c_k}], \tag{8}$$

where ":" denotes the concatenation operation. In contrast to traditional object detection methods, a weakly supervised learning approach obtains local sentiment regions in the SCAM module. In this process, the object-level label (ground truth of the sentiment regions) is not needed, and we obtain the result from the image-level label (ground truth of the sentiment category).

3.3 Regularization for classification

In the stage showing feature extraction and sentiment classification using the softmax function in Fig. 2, DenseNet pre-trained by ImageNet is fine-tuned to adjust parameters for sentiment classification. The fine-tuned DenseNet can fit the emotional

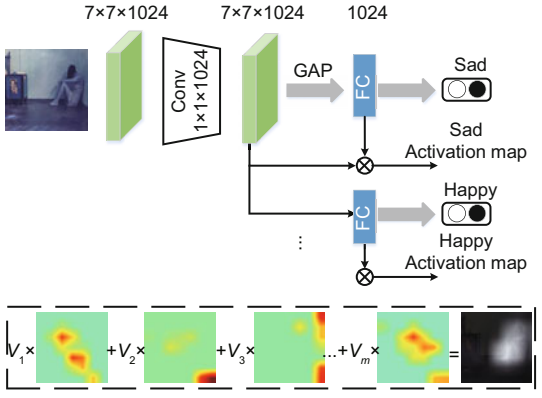

Fig. 3 Special-class activation map for allocating the individual two-class classifier for each sentiment class and obtaining special-class sentiment regions

This module transforms the correlative and complicated visual sentiment analysis into multiple independent and easy sentiment detection problems. FC: fully connected layer; Conv: convolutional layer; GAP: global average pooling

distribution from a collection of affective training samples $\{I_i, Y_i\}_{i=1}^N$ in a supervised way, where N is the size of the training set, and I_i and Y_i are the i^{th} input image and its sentiment label, respectively.

Let $\boldsymbol{c}_i = \text{GAP}(\boldsymbol{C}_{i,j,m}^{\boldsymbol{c}_k})$ be the output of the last FC layer. Then sentiment classification is carried out by minimizing the softmax loss function as follows:

$$L_{\text{class}} = -\sum_{i=1}^{N} Y_i \log \frac{\exp(\boldsymbol{W}^{\text{T}}\boldsymbol{c}_i)}{\sum_{\boldsymbol{C}} \exp(\boldsymbol{W}^{\text{T}}\boldsymbol{c}_i)}, \tag{9}$$

where $\boldsymbol{W}$ is the set of model parameters and Y_i is in the form of one-hot. Then, for all the two-class classifiers in the SCAM module, the individual loss function is defined as

$$L_{\text{SCAM}} = -\sum_{\boldsymbol{C}} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1-y_i)(1-\log \hat{y}_i)], \tag{10}$$

$$\hat{y}_i = \frac{1}{1+\exp(-\boldsymbol{f}_i)}, \tag{11}$$

where y_i denotes the label of sample N for the sentiment category, being 0 or 1, and $\boldsymbol{f}_i$ is the feature vector obtained by the last fully connected layer.

To further improve the accuracy of the sentiment recognition model, we apply both the center loss and triple loss as the regularization of the final classifier. As shown in Fig. 2, sentiment analysis is briefly thought as a linear problem. Specifically, straight lines in different colors are represented as different sentiment categories. To enhance the discriminative ability of the model, we use the center loss to reduce the distance between samples within the same sentiment:

$$L_{\text{center}} = \frac{1}{2} \sum_{i=1}^{N} \left\| \boldsymbol{c}_i - \boldsymbol{C}_{Y_i} \right\|_2^2, \tag{12}$$

where $\boldsymbol{C}_{Y_i}$ denotes the i^{th} class center of feature vector $\boldsymbol{c}_i$. In the center loss, $\boldsymbol{C}_{Y_i}$ is updated as the feature vector $\boldsymbol{c}_i$ is updated, which is updated only under the $\sum_{j=1}^{\boldsymbol{C}} Y_{ij} = 1$ condition. In contrast to traditional approaches for dealing with the center loss, we build the relationship between $\boldsymbol{c}_i$ and $\boldsymbol{C}_{Y_i}$ as

$$\boldsymbol{C}_{Y_i} = \frac{Y_i \exp(\boldsymbol{W}^{\text{T}}\boldsymbol{c}_i)}{\boldsymbol{C} \sum_{\boldsymbol{C}} \exp(\boldsymbol{W}^{\text{T}}\boldsymbol{c}_i)}. \tag{13}$$

Additionally, we apply the triple loss to increase the distances between samples in different sentiments:

$$L_{\text{triple}} = \sum_{i=1}^{N} \left[\left\| g(\boldsymbol{c}_i^{\text{a}}) - g(\boldsymbol{c}_i^{\text{p}}) \right\|_2^2 - \left\| g(\boldsymbol{c}_i^{\text{a}}) - g(\boldsymbol{c}_i^{\text{n}}) \right\|_2^2 + \theta \right], \tag{14}$$

where $g(\cdot)=\left\|\frac{\exp(\boldsymbol{c}_i)}{\sum_{\boldsymbol{C}}\exp(\boldsymbol{c}_i)}\right\|_2^2$ and θ is a hyper-parameter. $\boldsymbol{c}_i^{\mathrm{a}}$ and $\boldsymbol{c}_i^{\mathrm{p}}$ are in the same class, while $\boldsymbol{c}_i^{\mathrm{a}}$ and $\boldsymbol{c}_i^{\mathrm{n}}$ are in different categories. Finally, we define the final loss function as follows:

$$L=L_{\mathrm{class}}(\boldsymbol{I},\boldsymbol{Y})+L_{\mathrm{SCAM}}(\boldsymbol{I},y)+\lambda L_{\mathrm{center}}(\boldsymbol{c})+\beta L_{\mathrm{triple}}(\boldsymbol{c}), \tag{15}$$

where λ and β are parameters to balance the contributions of L_{center} and L_{triple} functions. Since all the parameters can be derived, we can conduct an end-to-end framework using adaptive moment estimation (Adam) to minimize the final loss function.

4 Experiments

4.1 Datasets

We evaluate our method on four benchmark affective datasets, i.e., Twitter II (You et al., 2015), EmotionROI (Peng KC et al., 2016), Flickr (Katsurai and Satoh, 2016), and Instagram (Katsurai and Satoh, 2016). Additionally, we apply the public dataset Cifar10 (Krizhevsky, 2009) to obtain the mislabeled datasets.

1. Twitter II

The Twitter II dataset was collected from the social website Twitter and labeled with sentiment polarity categories by Amazon Mechanical Turk (AMT) participants. It consists of 1269 images. In this dataset, 80% of the samples are split for the training set and 20% for the testing set. Based on this dataset, the experiment is conducted 10 times.

2. EmotionROI

The EmotionROI dataset was created for a sentiment prediction benchmark. It is assembled from Flickr, resulting in 1980 images with six sentiment categories (i.e., anger, disgust, fear, joy, sadness, and surprise). In addition, each image is manually annotated into 15 regions that evoke sentiments. The original EmotionROI training set is split into two parts, i.e., 80% for the training process and 20% for validation, and its original testing set is used for tests in our experiments.

3. Flickr and Instagram

The Flickr and Instagram datasets were collected by querying positive and negative keywords from the social websites Flickr and Instagram, respectively. A group of 225 AMT participants was asked to label the images, producing 60 745 and 42 856 images and receiving at least three agreements. These two datasets are split randomly into 80%, 5%, and 15% of the samples for the training set, validation, and testing set in our experiments, respectively.

To evaluate the effectiveness of the NECA module in mislabeled sentiment datasets, we structure the noisy training set using the original training set from Flickr/Instagram and the general dataset Cifar10 in proportion (it does not have sentiment samples, and we resize it to the same size as the other samples). The reason for choosing Flickr and Instagram datasets is that the numbers of samples are much larger than those of other two datasets. We set the proportion of samples selected randomly from Cifar10 to samples from the clean training set as $\delta=N_{\mathrm{C}}/N_{\mathrm{F/I}}$, where N_{C} and $N_{\mathrm{F/I}}$ are the numbers of images in the Cifar10 and Flickr/Instagram datasets, respectively.

4.2 Model ablation

For our end-to-end framework, we use a weight decay of 0.0001 with a momentum of 0.9 and batch size 32, and fine-tune all the layers with Adam. We set the learning rate as 0.0001 and drop the learning rate by 10 at every 10 epochs. Our experiments are conducted using Keras and TensorFlow, and performed on an NVIDIA GTX Titan GPU with 32 GB on-board memory.

We perform model ablation on the four affective datasets. To demonstrate the effectiveness of our framework for VSA, we quantitatively evaluate NECA, SCAM, L_{center} (C), L_{triple} (T), and the couple of C and T (i.e., regularization (R)). Parameters λ and β are selected according to grid search, which is performed in the range of [0, 1] for two parameters and the stride equals 0.1. As shown in Table 2, when the method is BASED+C+T, the pair of parameters ($\lambda=0.3$ and $\beta=0.2$) that achieves the best results is chosen in our experiments.

In Table 3, we compare the sentiment recognition accuracies of the methods mentioned above. Clearly, BASED+NECA+SCAM+R achieves the highest accuracy. Additionally, the methods with NECA, SCAM, C, T, or R generally outperform the baseline one (BASED). Comparative results indicate the following: (1) Non-extreme attention can effectively suppress the influence of noise, because the

attention scores of the noisy samples are much lower than those of the correct samples. The attention model allows only the correct samples to train the followed network for sentiment recognition. (2) The special-class activation map module further captures the affective regions for each sentiment category. These affective regions make significant contribution to VSA. (3) The loss regularizer applied to the classifier makes the feature representations of the samples from the same class as close as possible and those from different classes as far away as possible.

4.3 Performance comparison

1. Classification accuracy comparison on mislabeled datasets

Table 4 presents the performance of our model influenced by δ and the comparison with attention mechanisms mentioned in Section 3.1. We apply the channel-wise attention (CWA) mechanism in Chen L et al. (2017) after acquiring the feature vector $\boldsymbol{c}_i$. The spatial attention (SA) mechanisms in Chen L et al. (2017) and Zhuang et al. (2017) are applied after acquiring the feature map $\boldsymbol{X}^n_{i,j,m}$.

We can see from Table 4 that when the noise level is less than or equal to 0.25, our model achieves stable results. Additionally, it indicates that our model can still obtain a promising performance even when 50% samples are not correctly labeled. Meanwhile, compared with other attention mechanisms, the NECA module achieves better performance. The results demonstrate that our method can significantly reduce the number of expensive, time-consuming annotations while maintaining good performance.

2. Ranking of mislabeled samples by classification scores

To intuitively show the effectiveness of our proposed method in cleaning mislabeled samples, we randomly select images in the positive category with $\delta = 0.25$ in the Flickr dataset (Fig. 4). The noisy samples are ranked in descending order based on their classification scores obtained using

Table 2 Classification accuracy comparison for parameter selection on the Flickr dataset when the method is BASED+C+T

λ	Classification accuracy (%)										
	β=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	73.46	75.98	75.00	75.98	72.55	74.02	75.00	76.47	74.02	74.51	75.49
0.1	75.98	76.47	72.06	76.47	75.49	76.47	75.98	72.55	74.02	73.04	75.49
0.2	77.94	74.51	72.06	74.02	75.98	75.49	76.96	75.00	75.98	75.49	76.47
0.3	75.49	75.49	**79.41**	77.45	74.51	75.49	74.51	72.06	74.51	76.47	74.02
0.4	72.06	76.96	76.47	75.00	75.00	76.47	72.06	73.39	71.57	76.96	73.53
0.5	75.98	75.49	73.53	72.06	74.51	77.45	72.06	74.51	74.02	73.04	74.02
0.6	75.49	73.53	76.47	75.00	76.96	73.04	76.96	74.51	74.51	74.51	69.61
0.7	77.94	75.00	78.92	74.02	76.47	75.98	73.04	73.04	75.98	75.49	75.49
0.8	75.98	75.98	77.94	76.96	74.51	75.98	74.51	75.00	77.45	75.49	73.04
0.9	77.94	75.00	74.51	74.51	73.04	73.53	74.51	76.96	75.00	75.49	72.06
1.0	77.45	75.49	74.02	77.45	75.00	73.53	74.02	73.53	71.57	72.06	74.02

Best result is in bold

Table 3 Classification accuracy comparison for model ablation on Twitter II, EmotionROI, Flickr, and Instagram datasets

Method	Classification accuracy (%)			
	Twitter II	EmotionROI	Flickr	Instagram
BASED	73.34	41.02	73.46	68.35
BASED+NECA	80.22	54.63	82.45	81.10
BASED+SCAM	80.34	55.75	81.54	79.97
BASED+NECA+SCAM	82.46	57.84	81.94	82.48
BASED+NECA+C	84.35	60.06	82.98	82.97
BASED+NECA+T	82.96	57.94	83.17	83.12
BASED+NECA+SCAM+R	**85.43**	**61.25**	**84.62**	**83.55**

The pre-trained DenseNet121 using ImageNet is used as the BASED method. Best results are in bold

Table 4 Classification accuracy comparison on the mislabeled datasets

Dataset	Method	Classification accuracy (%)										
		δ=0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Flickr	BASED	73.46	69.80	66.14	65.34	62.63	57.87	53.21	49.76	48.45	42.14	40.21
	CWA (Chen L et al., 2017)	75.54	72.23	70.10	66.43	62.12	58.46	55.16	50.13	48.61	43.17	41.84
	SA (Chen L et al., 2017)	79.34	75.01	72.14	70.12	63.26	58.93	56.33	51.11	48.95	43.44	41.92
	SA (Zhuang et al., 2017)	79.89	77.75	73.41	69.96	63.67	59.05	56.60	50.98	48.79	45.53	43.03
	Ours	84.62	84.33	80.60	82.14	79.93	80.73	75.73	73.87	72.12	70.33	67.42
Instagram	BASED	68.35	64.32	60.73	54.17	52.86	48.57	45.53	43.32	39.45	35.57	33.63
	CWA (Chen L et al., 2017)	75.22	70.01	65.88	62.53	59.84	55.41	51.75	46.92	42.46	37.18	32.98
	SA (Chen L et al., 2017)	77.13	73.26	70.85	67.69	63.24	59.60	56.06	52.70	48.28	44.07	39.24
	SA (Zhuang et al., 2017)	79.57	76.14	72.33	67.44	64.26	60.50	56.99	53.67	48.53	44.67	40.05
	Ours	83.55	79.72	78.10	78.85	81.19	77.50	74.36	71.61	67.48	69.34	66.62

Results become smaller as δ becomes larger, but our method still works when the value of δ is close to 0.50. δ is the proportion of the mixed samples from Cifar10

BASED+NECA+SCAM+R. Images in the green rectangle rank at the top of the ranking queue and images in the red rectangle rank at the bottom. We can see that images in the green rectangle come from the Flickr dataset and can correctly express the corresponding sentiment. In contrast, samples in the red rectangle from the Cifar10 dataset have no sentiment expression. Therefore, Fig. 4 clearly shows that our method can collect well-labeled datasets and can be used for sentiment learning.

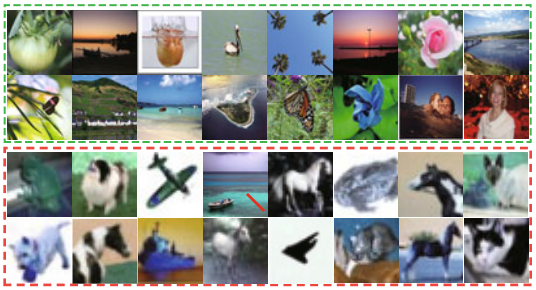

Fig. 4 Examples of samples in mislabeled sentiment datasets

Samples are sorted in descending order according to their ranks in the classification scores. Images in the green rectangle have higher classification scores than images in the red rectangle. The noise parameter $\delta = 0.25$. References to color refer to the online version of this figure

3. Classification accuracy comparison on well-labeled datasets

To demonstrate the effectiveness of our method in handing well-labeled sentiment datasets, we compare it with several related methods for image sentiment classification: hand-crafted, CNN-based, and CAM-related methods. Comparison results are shown in Table 5, where the results of the compared methods are obtained directly from the literature.

(1) Hand-crafted methods. Zhao et al. (2014) extracted 27-dimensional principle-of-art features from the affective images and used LIBSVM (Chang and Lin, 2011) for classification. SentiBank (Chen T et al., 2014a) extracted 1200-dimensional mid-level representations with the adjective noun pair (ANP) detector, while DeepSentiBank (Chen T et al., 2014a) applied the pre-trained DeepSentiBank to extract 2089-dimensional features.

(2) CNN-based methods. CNN-based methods include VGG16 (Simonyan and Zisserman, 2014) and DenseNet121 (Huang et al., 2017) pre-trained by ImageNet, fine-tuned VGG16, and DenseNet121. Sun et al. (2016) selected the top 1 affective region to extract local features and combined holistic features with local features to conduct sentiment classification. Yang et al. (2017a) used label distribution learning and developed a multi-task deep framework to generate local affective regions for each sentiment category. Then, the local affective regions were integrated with the holistic features to recognize emotions. In this method, the local region labels for the training process need to be annotated manually. Since the methods of Sun et al. (2016) and Yang et al. (2017a) are proposed for binary classification and multi-class classification, respectively, these two methods with incompatible numbers of classes in the datasets cannot be evaluated. Therefore, we denote the results from Sun et al. (2016) on EmotionROI and those from Yang et al. (2017a) on Twitter II, Flickr, and Instagram as "–."

(3) CAM-related methods. CAM aims to

find the object regions with only image-level labels in a weakly supervised way. We evaluate our method with WILDCAT (Durand et al., 2017), SPN (Zhu Y et al., 2017), and WSCNet (Yang et al., 2018b). Specifically, by WILDCAT, Durand et al. (2017) learned multiple localized features related to different class modalities, and pooled these features to predict the sentiment categories. By SPN, Zhu Y et al. (2017) aggregated image-specific patterns by coupling the candidate affective regions and the holistic feature maps of the images to sentiment classification. By WSCNet, which is based on CAM, Yang et al. (2018b) unified the detection and classification tasks into one framework for sentiment recognition.

Table 5 clearly shows that our method obtains the highest accuracies on four benchmark affective datasets, since our method extracts features by fine-tuned CNN and provides sentiment-specific local regions. Importantly, in the process of discovering the sentiment-specific regions, the problem of VSA is transferred into easier sentiment detection problems.

Fig. 5 Region visualization of samples discovered by our weakly supervised SCAM module on EmotionROI

One sample corresponds to one activation map, which corresponds to the label of the sample

4. Visualization of sentiment-specific regions

To intuitively show the sentiment-specific regions discovered by the SCAM module, we randomly visualize samples with their activation maps corresponding to their labels (Fig. 5), and all sentiment category regions come from one sample (Fig. 6). Here, feature maps are visualized as grayscale images according to their feature values. In these grayscale images, areas with higher feature values are brighter, and this is helpful in sentiment classification.

As shown in Fig. 5, sentiment activation maps of samples corresponding to labels capture the local and significant sentiment regions, where the sentiment could obviously be expressed. In Fig. 6, we can see that on the EmotionROI dataset, part of the sentiment activation maps (such as anger, disgust,

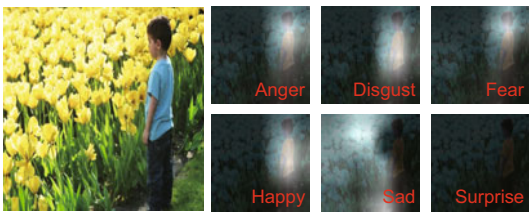

Fig. 6 Visualization of sentiment-specific category activation maps for one sample from EmotionROI

Different activation maps capture different regions evoking the corresponding sentiment

Table 5 Classification accuracy comparison with several baseline methods on Twitter, EmotionROI, Flickr, and Instagram

Category	Method	Classification accuracy (%)			
		Twitter II	EmotionROI	Flickr	Instagram
Hand-crafted	Zhao et al. (2014)'s	67.92	34.84	66.61	64.17
	SentiBank (Borth et al., 2013)	66.63	35.24	69.26	66.53
	DeepSentiBank (Chen T et al., 2014a)	71.25	42.35	70.16	67.13
CNN-based	ImageNet VGG16 (Simonyan and Zisserman, 2014)	67.49	37.26	69.88	63.44
	ImageNet DenseNet121 (Huang et al., 2017)	73.34	41.02	73.46	68.35
	Fine-tuned VGG16 (Simonyan and Zisserman, 2014)	76.99	45.46	78.14	77.41
	Fine-tuned DenseNet121 (Huang et al., 2017)	78.81	52.63	81.23	80.12
	Sun et al. (2016)'s	81.06	–	79.85	78.67
	Yang et al. (2017a)'s	–	52.40	–	–
CAM-related	WILDCAT (Durand et al., 2017)	79.53	55.05	80.67	80.31
	SPN (Zhu Y et al., 2017)	81.67	52.70	79.71	79.53
	WSCNet (Yang et al., 2018b)	84.25	58.25	81.36	81.81
	Ours	**85.43**	**61.25**	**84.62**	**83.55**

Datasets with incompatible class numbers cannot be evaluated and the classification accuracies are denoted as "–." Best results are in bold

fear, and happy) are focused on the person in samples, part of the sentiment activation maps (such as sadness) are focused on the background in samples, and the remaining sentiment activation maps are not focused on one point. This represents that one sample may express different sentiments and that each CAM module of a two-class classifier captures the corresponding sentiment activation map.

5. Visualization of feature distributions

To gain intuitive understanding of the influence of the regularization applied to the classifier, we visualize the distributions of the features learned in our model in different epochs. We can see from Figs. 7b–7d that samples in the same category stay as close as possible. Furthermore, from Figs. 7c and 7d, we can see that distances between samples in different categories become larger. From Fig. 7, we can see that our method learns a discriminative (between different sentiments) and compact (within one sentiment) feature representation, ultimately leading to superior classification performance.

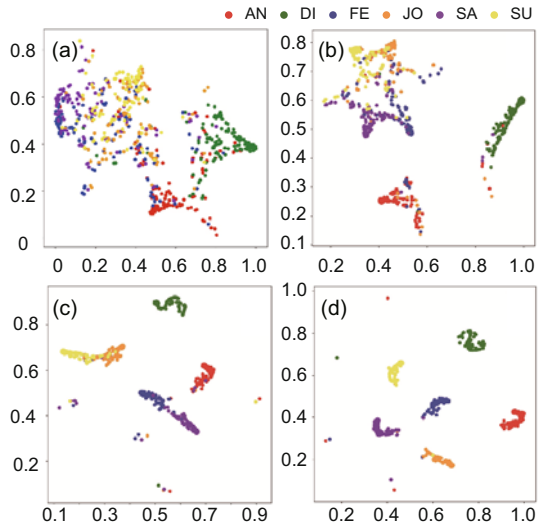

Fig. 7 Feature distributions of our model at every 15 epochs, representing one interval starting from the beginning on EmotionROI: (a) 15th epoch; (b) 30th epoch; (c) 45th epoch; (d) 60th epoch

Features are extracted from the outputs of the last hidden layer and mapped to a two-dimensional space using t-SNE (Hinton, 2008). References to color refer to the online version of this figure

5 Conclusions

In this study, we have proposed a weakly supervised network (NLWSNet) to learn visual sentiment representation from a mislabeled dataset. The proposed NLWSNet can effectively handle label noise through a non-extreme attention mechanism and a special-class activation map module. Through the attention mechanism, the proposed method suppresses the negative influence of the noisy samples and ensures that the information from the correct samples is not lost. Furthermore, it captures the affective regions for each sentiment category, and this local information is helpful in sentiment classification. Finally, the regularization helps the sentiment classifier learn feature representation by adjusting the distance between samples. The efficacy of our method has been demonstrated by extensive experiments.

Contributors

Luo-yang XUE designed the research and drafted the manuscript. Qi-rong MAO and Xiao-hua HUANG helped organize the manuscript. Jie CHEN participated in the experiments. Luo-yang XUE and Qi-rong MAO revised and finalized the paper.

Compliance with ethics guidelines

Luo-yang XUE, Qi-rong MAO, Xiao-hua HUANG, and Jie CHEN declare that they have no conflict of interest.

References

Borth D, Ji RR, Chen T, et al., 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. Proc 21st ACM Int Conf on Multimedia, p.223-232. https://doi.org/10.1145/2502081.2502282

Campos V, Salvador A, Giró-i-Nieto X, et al., 2015. Diving deep into sentiment: understanding fine-tuned CNNs for visual sentiment prediction. https://arxiv.org/abs/1508.05056

Campos V, Jou B, Giró-i-Nieto X, 2017. From pixels to sentiment: fine-tuning CNNs for visual sentiment prediction. *Image Vis Comput*, 65:15-22. https://doi.org/10.1016/j.imavis.2017.01.011

Chang CC, Lin CJ, 2011. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*, 2(3):27. https://doi.org/10.1145/1961189.1961199

Chen L, Zhang HW, Xiao J, et al., 2017. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6298-6306. https://doi.org/10.1109/CVPR.2017.667

Chen SX, Zhang CJ, Dong M, et al., 2017. Using ranking-CNN for age estimation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.742-751. https://doi.org/10.1109/CVPR.2017.86

Chen SX, Zhang CJ, Dong M, 2018a. Coupled end-to-end transfer learning with generalized Fisher information. Proc IEEE/CVF Conf on Computer Vision and Pattern

Recognition, p.4329-4338.
https://doi.org/10.1109/CVPR.2018.00455

Chen SX, Zhang CJ, Dong M, 2018b. Deep age estimation: from classification to ranking. *IEEE Trans Multim*, 20(8):2209-2222.
https://doi.org/10.1109/TMM.2017.2786869

Chen T, Borth D, Darrell T, et al., 2014a. DeepSentiBank: visual sentiment concept classification with deep convolutional neural networks.
https://arxiv.org/abs/1410.8586

Chen T, Yu FX, Chen JW, et al., 2014b. Object-based visual sentiment concept analysis and application. Proc 22nd ACM Int Conf on Multimedia, p.367-376.
https://doi.org/10.1145/2647868.2654935

Corbetta M, Shulman GL, 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci*, 3(3):201-205. https://doi.org/10.1038/nrn755

Deng J, Dong W, Socher R, et al., 2009. ImageNet: a large-scale hierarchical image database. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.248-255.
https://doi.org/10.1109/CVPRW.2009.5206848

Durand T, Mordan T, Thome N, et al., 2017. WILDCAT: weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5957-5966.
https://doi.org/10.1109/CVPR.2017.631

Fang Y, Tan H, Zhang J, 2018. Multi-strategy sentiment analysis of consumer reviews based on semantic fuzziness. *IEEE Access*, 6:20625-20631.
https://doi.org/10.1109/ACCESS.2018.2820025

Girshick R, Donahue J, Darrell T, et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.580-587.
https://doi.org/10.1109/CVPR.2014.81

He XT, Peng YX, 2017. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. Proc 21st AAAI Conf on Artificial Intelligence, p.4075-4081.

He XT, Peng YX, Zhao JJ, 2019. Fast fine-grained image classification via weakly supervised discriminative localization. *IEEE Trans Circ Syst Video Technol*, 29(5):1394-1407.
https://doi.org/10.1109/TCSVT.2018.2834480

Hinton GE, 2008. Visualizing high-dimensional data using t-SNE. *Vigil Christ*, 9(2):2579-2605.

Hu J, Shen L, Sun G, 2018. Squeeze-and-excitation networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7132-7141.
https://doi.org/10.1109/CVPR.2018.00745

Huang G, Liu Z, van der Maaten L, et al., 2017. Densely connected convolutional networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2261-2269.
https://doi.org/10.1109/CVPR.2017.243

Itti L, Koch C, Niebur E, 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Patt Anal Mach Intell*, 20(11):1254-1259.
https://doi.org/10.1109/34.730558

Jia XB, Jin Y, Li N, et al., 2018. Words alignment based on association rules for cross-domain sentiment classification. *Front Inform Technol Electron Eng*, 19(2):260-272. https://doi.org/10.1631/FITEE.1601679

Katsurai M, Satoh S, 2016. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.2837-2841.
https://doi.org/10.1109/ICASSP.2016.7472195

Krizhevsky A, 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report TR-2009, University of Toronto, Toronto, Canada.

Krizhevsky A, Sutskever I, Hinton GE, 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 60(6):84-90.
https://doi.org/10.1145/3065386

LeCun Y, Boser B, Denker JS, et al., 1989. Backpropagation applied to handwritten zip code recognition. *Neur Comput*, 1(4):541-551.
https://doi.org/10.1162/neco.1989.1.4.541

Li ZH, Fan YY, Liu WH, et al., 2018. Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multim Tools Appl*, 77(1):1115-1132.
https://doi.org/10.1007/s11042-016-4310-5

Liu GL, Reda FA, Shih KJ, et al., 2018. Image inpainting for irregular holes using partial convolutions. Proc 15th European Conf on Computer Vision, p.89-105.
https://doi.org/10.1007/978-3-030-01252-6_6

Machajdik J, Hanbury A, 2010. Affective image classification using features inspired by psychology and art theory. Proc 18th ACM Int Conf on Multimedia, p.83-92.
https://doi.org/10.1145/1873951.1873965

Mikels JA, Fredrickson BL, Larkin GR, et al., 2005. Emotional category data on images from the international affective picture system. *Behav Res Methods*, 37(4):626-630. https://doi.org/10.3758/BF03192732

Oquab M, Bottou L, Laptev I, et al., 2015. Is object localization for free?—Weakly-supervised learning with convolutional neural networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.685-694.
https://doi.org/10.1109/CVPR.2015.7298668

Ou WH, Luan X, Gou JP, et al., 2018. Robust discriminative nonnegative dictionary learning for occluded face recognition. *Patt Recogn Lett*, 107:41-49.
https://doi.org/10.1016/j.patrec.2017.07.006

Park J, Woo S, Lee JY, et al., 2018. BAM: bottleneck attention module. Proc British Machine Vision Conf, Article 147.

Peng KC, Sadovnik A, Gallagher A, et al., 2016. Where do emotions come from? Predicting the emotion stimuli map. Proc IEEE Int Conf on Image Processing, p.614-618. https://doi.org/10.1109/ICIP.2016.7532430

Peng YX, Qi JW, Zhuo YK, 2019a. MAVA: multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism. *IEEE Trans Image Process*, 29:2728-2741. https://doi.org/10.1109/TIP.2019.2952085

Peng YX, Zhao YZ, Zhang JC, 2019b. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Trans Circ Syst Video Technol*, 29(3):773-786.
https://doi.org/10.1109/TCSVT.2018.2808685

Rohrbach A, Rohrbach M, Hu RH, et al., 2016. Grounding of textual phrases in images by reconstruction. Proc 14th European Conf on Computer Vision, p.817-834.
https://doi.org/10.1007/978-3-319-46448-0_49

Simonyan K, Zisserman A, 2014. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

Sun M, Yang JF, Wang K, et al., 2016. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. Proc IEEE Int Conf on Multimedia and Expo, p.1-6. https://doi.org/10.1109/ICME.2016.7552961

Szegedy C, Vanhoucke V, Ioffe S, et al., 2016. Rethinking the inception architecture for computer vision. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2818-2826. https://doi.org/10.1109/CVPR.2016.308

Wang F, Jiang MQ, Qian C, et al., 2017. Residual attention network for image classification. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6450-6458. https://doi.org/10.1109/CVPR.2017.683

Woo S, Park J, Lee JY, et al., 2018. CBAM: convolutional block attention module. Proc 15th European Conf on Computer Vision, p.3-19. https://doi.org/10.1007/978-3-030-01234-2_1

Xiao FY, Sigal L, Lee YJ, 2017. Weakly-supervised visual grounding of phrases with linguistic structures. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5253-5262. https://doi.org/10.1109/CVPR.2017.558

Yang JF, She DY, Sun M, 2017a. Joint image emotion classification and distribution learning via deep convolutional neural network. Proc 26th Int Joint Conf on Artificial Intelligence, p.3266-3272. https://doi.org/10.24963/ijcai.2017/456

Yang JF, Sun M, Sun XX, 2017b. Learning visual sentiment distributions via augmented conditional probability neural network. Proc 31st AAAI Conf on Artificial Intelligence, p.224-230.

Yang JF, She DY, Sun M, et al., 2018a. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Trans Multim*, 20(9):2513-2525. https://doi.org/10.1109/TMM.2018.2803520

Yang JF, She DY, Lai YK, et al., 2018b. Weakly supervised coupled networks for visual sentiment analysis. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7584-7592. https://doi.org/10.1109/CVPR.2018.00791

You QZ, Luo JB, Jin HL, et al., 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. Proc 29th AAAI Conf on Artificial Intelligence, p.381-388.

You QZ, Luo JB, Jin HL, et al., 2016. Building a large scale dataset for image emotion recognition: the fine print and the benchmark. Proc 30th AAAI Conf on Artificial Intelligence, p.308-314.

You QZ, Jin HL, Luo JB, 2017. Visual sentiment analysis by attending on local image regions. Proc 31st AAAI Conf on Artificial Intelligence, p.231-237.

Yu JH, Lin Z, Yang JM, et al., 2018. Generative image inpainting with contextual attention. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5505-5514. https://doi.org/10.1109/CVPR.2018.00577

Yuan JB, Mcdonough S, You QZ, et al., 2013. Sentribute: image sentiment analysis from a mid-level perspective. Proc 2nd Int Workshop on Issues of Sentiment Discovery and Opinion Mining, p.1-8. https://doi.org/10.1145/2502069.2502079

Zagoruyko S, Komodakis N, 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. https://arxiv.org/abs/1612.03928

Zeng SN, Gou JP, Yang X, 2018. Improving sparsity of coefficients for robust sparse and collaborative representation-based image classification. *Neur Comput Appl*, 30(10):2965-2978. https://doi.org/10.1007/s00521-017-2900-4

Zhang FF, Mao QR, Shen XJ, et al., 2018a. Spatially coherent feature learning for pose-invariant facial expression recognition. *ACM Trans Multim Comput Commun Appl*, 14(1s):27. https://doi.org/10.1145/3176646

Zhang FF, Zhang TZ, Mao QR, et al., 2018b. Joint pose and expression modeling for facial expression recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3359-3368. https://doi.org/10.1109/CVPR.2018.00354

Zhang N, Donahue J, Girshick R, et al., 2014. Part-based R-CNNs for fine-grained category detection. Proc 13th European Conf on Computer Vision, p.834-849. https://doi.org/10.1007/978-3-319-10590-1_54

Zhang QS, Zhu SC, 2018. Visual interpretability for deep learning: a survey. *Front Inform Technol Electron Eng*, 19(1):27-39. https://doi.org/10.1631/FITEE.1700808

Zhao SC, Gao Y, Jiang XL, et al., 2014. Exploring principles-of-art features for image emotion recognition. Proc 22nd ACM Int Conf on Multimedia, p.47-56. https://doi.org/10.1145/2647868.2654930

Zhao SC, Yao HX, Gao Y, et al., 2016. Predicting personalized emotion perceptions of social images. Proc 24th ACM Int Conf on Multimedia, p.1385-1394. https://doi.org/10.1145/2964284.2964289

Zhao SC, Ding GG, Gao Y, et al., 2017. Approximating discrete probability distribution of image emotions by multi-modal features fusion. Proc 26th Int Joint Conf on Artificial Intelligence, p.4669-4675. https://doi.org/10.24963/ijcai.2017/651

Zhou BL, Khosla A, Lapedriza A, et al., 2016. Learning deep features for discriminative localization. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2921-2929. https://doi.org/10.1109/CVPR.2016.319

Zhu Y, Zhou YZ, Ye QX, et al., 2017. Soft proposal networks for weakly supervised object localization. Proc IEEE Int Conf on Computer Vision, p.1859-1868. https://doi.org/10.1109/ICCV.2017.204

Zhu YK, Groth O, Bernstein M, et al., 2016. Visual7W: grounded question answering in images. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.4995-5004. https://doi.org/10.1109/CVPR.2016.540

Zhuang BH, Liu LQ, Li Y, et al., 2017. Attend in groups: a weakly-supervised deep learning framework for learning from web data. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2915-2924. https://doi.org/10.1109/CVPR.2017.311