Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 1, 2023

Image Emotion Analysis Combining Attention Mechanism and Multi-level Correlation

Shuxia Ren, Simin Li

School of Software, Tianjin Polytechnic University, Tianjin 300387, China

Abstract: The development of social network has brought a large amount of image information, and the research on image emotion has gradually attracted wide attention. The current image emotion analysis methods based on multi-level features simply splice the features at each level and then classify the emotions, which not only ignores the correlation between features at different levels, but also ignores the synergistic effect between global features and local features. Therefore, this paper proposes an emotion model (MAML) based on mixed attention and multi-level dependence of images, which uses spatial and channel attention mechanisms to extract local emotion region features of images. Bi-directional Long Short Term Memory network (BiLSTM) is used to establish correlation between multi-level image global features. The experimental results of MAML model on artphoto and abstract data sets prove the validity of MAML model.

Keywords: Multi-level Features; Mixed Attention; Global and Local; BiLSTM.

1. Introduction

With the increasing popularity of social media, more and more users post images on social networks to express their views and emotions, and images have gradually become an important carrier for people to express their emotions. Emotion research plays an indispensable role in education, advertising, decision-making, planning and many other activities. Therefore, image sentiment analysis has attracted much attention from researchers. However, because of the subjectivity and complexity of emotion, the task of emotion analysis for images is still very challenging.

As shown in Figure 1, image emotion is closely related to image color, texture, line and other underlying features. At the beginning of the study, the researchers predicted emotion by designing various manual features (such as color, texture, composition, etc.) [1,2,3]. With the development of Convolutional Neural Network (CNN), image emotion analysis methods based on deep learning show excellent performance [4,5,6]. At present, researchers have designed various neural networks for image emotion analysis. Peng et al. [7] extracted global features of images for emotion classification by fine-tuning the CNN model pre-trained in ImageNet. Research [8] shows that when humans observe images, emotion is mainly caused by local emotional areas. Based on descriptive visual attributes, You et al. [9] used the attention model to find the local areas that evoke the audience's emotions, and then extracted their features to improve the performance of emotion analysis. Zhao et al. [10] showed through experiments that the effect of using both local and global features is better than that of using one of them alone. However, most current studies ignore the synergistic effect of local and global features, and the attention model used in image emotion analysis tasks usually only considers spatial attention when focusing on local emotion features. Although spatial attention regulates the local spatial connectivity pattern on each channel through spatial attention weights [9,11,12], it ignores the interdependencies between different channels. However, it is very important to pay attention to the channel aspect, which can be seen as a process of selecting semantic attributes and is essentially consistent with the characteristics of CNN [13].

Previous studies have shown that image emotion is related to features ranging from low level to high level [14]. There is correlation between features at different levels. Zhu et al. [14] and Zhang et al. [15] used bidirectional GRU method and Gram method respectively to establish correlation between features at different levels. However, these methods pay too much attention to the global features of images and neglect the local emotion region features closely related to image emotion.

Fig 1. Emotional image sample

To sum up, although researchers have improved the effect of image sentiment analysis from multiple perspectives, how to use the spatial and channel attention mechanisms as well as the correlation between features at different levels of images to extract more discriminative features is still a problem to be solved. Therefore, this paper proposes an image feature extraction method that combines mixed attention and multi-level feature dependence. Multiple branches are used to extract multi-level features from low to high, and BiLSTM is used to achieve multi-level feature fusion. In addition, after the highest level of convolutional blocks, the lightweight spatial and channel attention module CBAM is used to focus on the local emotion area. Therefore, the MAML model proposed in this paper can not only extract the local features of the image that cause the viewer's emotion, but also extract the complete feature information of the image.

2. Related Work

2.1. Multi-level Features

Closing the affective gap is a major challenge in affective prediction. In the past, a lot of efforts have been focused on feature extraction. In earlier studies, image sentiment analysis methods mainly classified emotions according to manual

features of images, such as low-level features such as color and texture, intermediate features such as image composition and aesthetics, and high-level features that contain semantic information.

Inspired by the fact that CNN methods work well in other visual recognition tasks with the development of deep learning, researchers began to apply CNN-based methods to image sentiment analysis. The CNN-based approach can learn features automatically through its multi-level deep learning architecture, rather than manually designing image features. Peng et al. [7] tried to apply CNN model to image emotion recognition for the first time. They fine-tuned the pre-trained CNN on ImageNet[16], showing that the CNN model is superior to the manual method on the Emotion6 dataset. You et al. [17] combined the CNN model in literature [18] with support vector machine (SVM) to detect image emotion on a Web image dataset. They demonstrated that the CNN method can capture more advanced emotional features than manual methods. Image emotion representation is related to both low-level and high-level features, but these methods only use the last level of CNN high-level semantic feature vector, ignoring the importance of low-level features. Some researchers combine higher-level semantic information with lower-level visual features in different ways to guide emotion classification. Zhu et al. [14] designed a CNN-RNN model to extract visual and semantic features through underlying convolution, and then aggregate them using bidirectional cyclic convolutional neural networks (BiRNN). Rao et al. [19] used three kinds of convolutional neural networks to obtain the three hierarchical features of the original image, the prominent theme and the color respectively, and conducted a comprehensive analysis of the emotion categories, achieving good results. However, these methods often use aggregation functions to fuse the features of each level of the image for emotion classification. Sowmyayani and Rani [20] point out that prominent objects in images play an important role in determining emotion. First, a spectral significance detection model is used to detect significant objects from the entire image, and then depth features and manual features of significant objects are extracted. However, prominent objects are important for emotion classification research, but other parts of the image cannot be ignored, and some images are difficult to extract prominent objects.

The above methods based on manual features and CNN model independently consider the global features of images or the features of significant objects to predict emotion. However, Zhao et al. [10] show through experiments that it is better to use local and global features at the same time than to use one of them alone. Therefore, multi-level global features and local regional features can be combined to classify image emotion.

2.2. Mixed Attention Mechanism

The basic idea of the attention mechanism is to make the system ignore irrelevant information and focus on important information. With the development of deep learning, attention mechanism has been widely applied in the field of computer vision, such as object detection [21], image classification [22], image captions generation [23], etc.

There have been some studies on integrated visual attention to the CNN emotion classification framework. Song et al. [24] proposed an emotion network with visual attention, integrating visual attention into the emotion classification framework to locate local areas related to emotion. Yang et al. [25] proposed a weakly supervised model coupled with attention mechanism, which combined visual emotion detection and classification within a unified CNN framework. These attention mechanisms mainly focus on spatial attention and ignore channel attention, but channel attention is also a process of selecting semantic attributes. Zhao et al. [26] established a deep attention network that integrates spatial attention and channel attention with the same polarity for image emotion regression. In order to make full use of the multi-level, spatial and channel features of CNN, Li et al. [6] proposed a SCEP model, integrating spatial attention and channel attention mechanisms into a classical convolutional neural network layer structure to predict image emotion. In 2018, Woo S et al. [27] proved through experiments that the mode of channeling attention module first and then spatial attention module is more effective. Therefore, combining the attention mechanism in channel and space, they proposed a lightweight and universal modular hybrid attention mechanism model CBAM.

The existing researches pay too much attention to the spatial attention but neglect the channel attention, or ignore the importance of the global feature when focusing on the local emotion region of the image. Therefore, in this paper, when extracting multi-level image features, features of different levels are fused by BiLSTM to form global features, and a branch is used separately to obtain local features by integrating CBAM module. In this way, not only the spatial and channel attention mechanisms are used to focus on the important local emotional features, but also the global features are guaranteed to participate in the feature fusion.

2.3. BiLSTM

After extracting the multi-level features of the image, it is necessary to fuse them and then carry out the emotion classification. Rao et al. [19] proposed a multi-level deep network MldrNet, which uses Max and Avg aggregation functions to fuse features of different levels. But this approach ignores the dependencies between the multi-level features. Zhu et al. [14] proposed a unified CNN-RNN visual emotion recognition model, which utilized the features of multiple branches in CNN at different levels, and effectively integrated these features by using the dependency relationship between them through the bidirectional GRU method. Zhang et al. [28] proposed a multi-level hybrid model, which learns and integrates deep semantics and shallow visual representations. Considering that images have strong texture features, Gram matrix is used to establish the correlation between multi-level features and form Gram matrix for emotion classification.

In the above methods to establish multi-level feature correlation, the bidirectional GRU method can retain fewer features per GRU, so using bidirectional GRU to establish multi-level feature correlation will affect the performance of image emotion classification. Gram matrix establishes correlation according to the texture features of image, and the weight of higher semantic features is reduced when the lower-level to higher-level features is established. Since Long Short-Term Memory network (LSTM) has more memory units, it can retain more image features, so this paper uses BiLSTM network to mine the correlation between features at different levels.

Therefore, in this paper, we propose an image sentiment analysis model MAML based on mixed attention and multi-level dependency, which combines the local features obtained by CBAM module with the multi-level global features

integrated by BiLSTM, so as to improve the performance of image sentiment analysis.

3. The Proposed Method

The emotion analysis model MAML proposed in this paper is shown in Figure 2, which mainly consists of three parts: multi-level feature extraction, local emotion feature recognition, and multi-level feature dependency establishment. Many studies have shown that image emotion is related to lower-level to higher-level features, so this paper uses the layered stacking structure of CNN to extract multi-level features of images. Considering the importance of local and global features to image emotion analysis, CBAM module is introduced in this paper, and spatial and channel attention mechanisms are used to mine regions and features with richer image emotion information. Since there is a dependency between features at different levels, this paper uses BiLSTM to integrate features at 5 different levels of images, establish the correlation between them, and refine the feature information at each level.

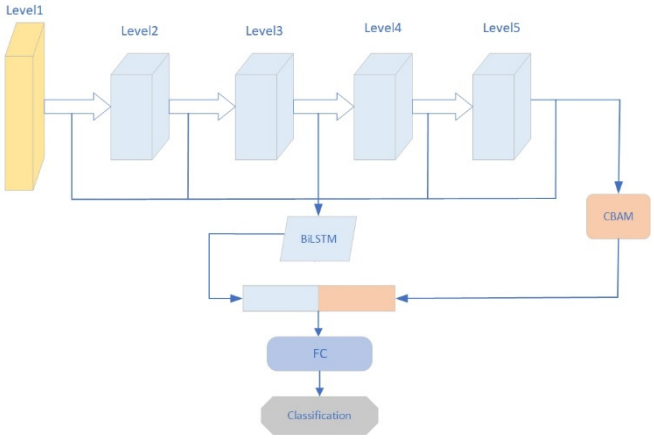

Fig 2. MAML model

3.1. Multi-level Feature Extraction

At present, convolutional neural networks have been widely used in image emotion analysis tasks, and good results have been obtained by using the last layer of high-level semantic features extracted by neural networks for emotion classification. Because of the hierarchical structure of convolutional neural networks, CNNS can capture features at different levels. The features extracted by shallow network contain more pixel information, such as color, texture, edge and so on. The deep network receptive field is increased, and the extracted features contain more abstract semantic information, but the resolution is lower and the perception of detail is poor. Consider that image emotion is not only related to higher-level features, but also to lower-level features. In this paper, Resnet18 is used to extract multi-layer features of images. transforms an image to 224×224×3 pixels is first preprocessed using Transforms, and the image is then input into the Resnet18 network, which consists of 5 parts. The convolution layer in the first part, Level1, is 7×7×64. The lowest layer feature F_0 of the image is extracted by the first convolution block and a 3×3×64 maximum pooling layer. The convolution layers contained in the other four parts Level2, Level3, Level4 and Level5 are 3×3×64, 3×3×128, 3×3×256 and 3×3×512 respectively. Each part corresponds to the features of a level, so the features of the other four levels are F_1, F_2, F_3 and F_4.

3.2. Local Emotion Feature Recognition

Attention mechanism is widely used in the field of computer vision, but the existing researches pay too much attention to the spatial attention of images and ignore the channel attention, or use the local features extracted by attention mechanism and ignore the global features. Therefore, in this paper, a hybrid attention mechanism CBAM is used to extract local features, which can generate attention feature map information in two dimensions of channel and space, and then multiply the two-feature map information with the original input feature map for adaptive feature correction to produce the final feature map. CBAM is mainly composed of channel attention module and spatial attention module. The complete CBAM module is shown in Figure 3.

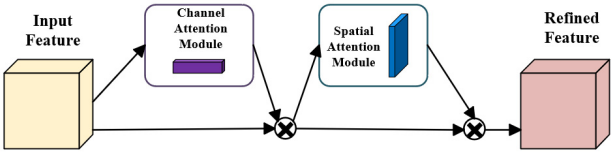

Fig 3. CBAM model

The channel attention module uses average pooling and maximum pooling to learn discriminant features, F^c_{avg} and $F^c_{\max}$ represents the features obtained after average pooling and maximum pooling, respectively. This feature is then fed into a shared multi-layer perceptron (MLP) network to generate the final channel attention feature map $M_c \in R^{C\times1\times1}$. In order to reduce the calculation parameters, a dimensionality reduction parameter r is used in MLP, $M_c \in R^{C/r\times1\times1}$, so the calculation formula of the channel attention module is

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{1}$$

The spatial attention module generates the spatial attention feature map $M_s(F) \in R^{H,W}$ based on the channel attention feature map, which is the same as the channel attention mechanism, and simultaneously uses average pooling and maximum pooling to generate discriminant features. However, the spatial attention mechanism generates 2D feature maps $F^s_{avg} \in R^{1\times H\times W}$ and $F^s_{\max} \in R^{1\times H\times W}$, so the calculation formula of the spatial attention module is

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}(F^s_{avg}; F^s_{\max})) \end{aligned} \tag{2}$$

Based on Resnet18, a total of 5 features of different levels are generated in this paper. Considering that it is better to use both local features and global features for emotion analysis than to use one of them alone, this paper inputs the features of the highest level with the richest semantic information into the CBAM module to generate features in the local emotion region. CBAM first passes the input feature $F \in R^{C\times H\times W}$ through the channel attention module to obtain 1D channel attention feature diagram $M_c \in R^{C\times1\times1}$, and then the spatial attention module obtains 2D space attention feature diagram $M_s \in R^{1\times H\times W}$ according to the channel attention feature diagram. The general process is as follows:

$$F_4 = M_c(F_4) \otimes F_4 \tag{3}$$

$$F_4 = M_s(F_4) \otimes F_4 \tag{4}$$

Where $\otimes$ represents element level multiplication. F_4 is the local emotion feature map after channel attention and spatial attention. In order to reduce overfitting, two 2×2 maximum pooling layers are continuously used to filter the features output by CBAM module and the features output by Resnet18 at five different levels to reduce the feature dimension. The specific implementation method is as follows:

$$P_i = MaxPool(F_i) \quad (5)$$

$$P_i = MaxPool(P_i) \quad (6)$$

$$X_i = Flatten(P_i) \quad (7)$$

$$F_i = RELU(W_i X_i + b_i) \quad (8)$$

Where i=0,1,2,3,4, W_i,b_i are the parameters of the full connection layer.

3.3. Multi-level Feature Dependency Establishment

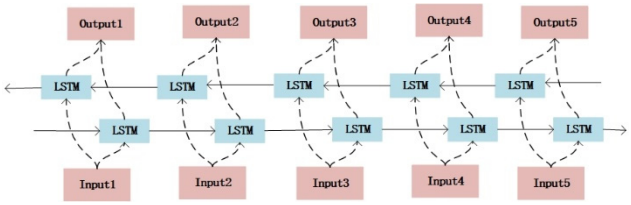

Fig 4. BiLSTM model

Five different levels of features are obtained through Resnet18, and there is a certain dependence among them, which is often ignored by existing studies. Therefore, BiLSTM is used in this paper to explore the dependency relationship between them,and the general flow of BiLSTM is shown in Figure 4.

The LSTM mainly includes the forgetting gate, the input gate, the output gate and the recording unit of the last moment. The multi-level image features are obtained in the multi-level feature extraction part, and they are used as the input of each moment of LSTM unit. The output process of the entire LSTM unit is as follows:

$$F_t = \delta(W_f \cdot (h_{t-1}, p_i) + b_f) \quad (9)$$

$$I_t = \delta(W_i \cdot (h_{t-1}, p_i) + b_i) \quad (10)$$

$$\widetilde{C}_t = \tanh(W_c \cdot (h_{t-1}, p_i) + b_c) \quad (11)$$

$$C_t = F_t * C_{t-1} + I_t * \widetilde{C}_t \quad (12)$$

$$O_t = \delta(W_o \cdot (h_{t-1}, p_i) + b_o) \quad (13)$$

$$h_t = O_i * \tanh(C_t) \quad (14)$$

Where F_t, I_t, $\widetilde{C}_t$, C_t and O_t are respectively the forgetting gate, the input gate, the candidate storage unit value, the memory unit value of the current moment, and the output gate. h_t is the output information of the LSTM unit. After the forward LSTM model and the reverse LSTM model, the forward hidden feature $\vec{h}_t$ and the reverse hidden feature $\overleftarrow{h}_t$ are obtained for each layer of image features. The features of five different levels of the image are fused by BiLSTM and input into LayerNorm layer to obtain the final global feature *F*.

After obtaining the local features extracted by CBAM module and the global features after BiLSTM fusion of multi-level features, this paper uses cat () function to concatenate these two features to obtain feature F'. Finally, the dimension is further reduced through the full connection layer and sent to the softmax layer for classification:

$$c = soft\max(W_2 \operatorname{Re}lu(W_1(Flatten(F')) + b_1) + b_2) \quad (15)$$

4. The Experiments

4.1. Dataset

This paper evaluates our approach using two datasets, ArtPhoto and Abstract [34], the details of which are shown in Table 1.

ArtPhoto: It is a selection of 806 photos from an art sharing website, which are taken by artists who consciously manipulate the position, brightness, color, etc. of the image to evoke a certain emotion in the viewer.

Abstract: This dataset consists of 228 abstract paintings. Unlike images in the ArtPhoto dataset, images in the abstract dataset represent emotions through overall colors and textures, rather than some emotional object. In this dataset, each painting was voted on by 14 different people to determine its emotional category. Select the emotion category with the most votes as the emotion category for the image.

Table 1. Statistical data of image emotion dataset

Dataset	Positive	Negative	Sum
ArtPhoto	378	428	806
Abstract	139	89	228

4.2. Experiment Settings

The model in this article is based on a pre-trained Resnet18 network and implemented using the PyTorch framework, with 40 epochs trained on NVIDIA GeForce MX450. The batch size and learning rate are set to 4 and 0.0001, respectively. During model training, images were randomly cropped to different sizes and aspect ratios and scaled to 224×224 pixels, and then randomly flipped horizontally with a probability of 0.5 to expand the data and prevent overfitting. Finally, the image is standardized according to the channel, so as to accelerate the convergence speed of the model. ArtPhoto and Abstract were randomly divided into 80% training datasets and 20% test datasets.

4.3. Compare with the Previous Methods

The method presented in this paper is compared with the following different baselines:

GCH [15]: The global view of the image is composed using a 64-bit binary color histogram feature.

Sentibank [29]: A 1200-dimensional intermediate feature called adjective-noun pair (ANPs) was proposed to describe the relationship between image content and emotion. This study is an important work to explore the correspondence between early semantic information and emotion.

Rao [30]: Early explorations in the analysis of local areas related to emotion. The image is segmented into different blocks by image segmentation, called multi-scale blocks, and the vision bag features based on sift contain both local and global information extracted from the image blocks.

PCNN [31] : A progressive training framework based on VGGNet. They used large amounts of weakly supervised data to make the model learn some common visual features to reduce the difficulty of training visual emotion datasets.

AR [32] : This study proposes a new concept of emotion region to explore local region and emotion arousal, and uses ready-made object detection technology as local information for analysis in combination with VGG model.

CNNGSR [33]: Xiong et al proposed R-CNNGSR obtained the initial emotion prediction model by using group sparse regularization through CNN, and then obtained a compact neural network, then combined the underlying features and emotion features to automatically detect the emotion region, and finally integrated the whole image and emotion region to predict the overall emotion of the image.

Zhang [28]: Zhang et al. proposed a multi-level hybrid model that learns and integrates deep semantics and shallow visual representations for emotion classification.

4.4. Experimental Result

In the experiment, two datasets, ArtPhoto and Abstract, were randomly divided into a training set and a test set at a ratio of 8:2. In the above experimental environment, the performance effects of MAML emotion model on the two data sets are shown in Table 3. As can be seen from Table 3, the accuracy of MAML model on ArtPhoto and Abstract datasets reached 79.38% and 82.14% respectively, higher than the baseline method, thus verifying the effectiveness of the proposed method.

Table 2. Classification results of different methods

Method	Dataset	
	ArtPhoto	Abstract
GCH	66.53	67.33
Sentibank	67.33	64.30
Rao	71.53	67.82
PCNN	70.96	70.84
AR	74.80	76.03
R-CNNGSR	75.02	75.89
Zhang	75.63	77.85
Ours	79.38	82.14

4.5. Contrast Experiment

Our method is compared with several previous methods on two datasets, ArtPhoto and Abstract. As can be seen from Table 3, the method based on deep learning has better performance than the method based on traditional manual features. However, the method based on sift visual bag features of Rao et al. [30], which extracts local and global information from image blocks, has 0.57% higher accuracy in emotion prediction on ArtPhoto data set than that of PCNN method [31]. It shows that manual features play an important role in image sentiment analysis. AR method [32] and R-CNNGSR method [33] automatically detect the local emotion region, and then combine the features of the local emotion region with the overall image features to produce the final emotion prediction. Different from the previous methods, Zhang et al. [28] used a multi-level mixed model to achieve better effect in emotion classification by combining features of different levels from low to high. In this paper, we make full use of the multi-level features of convolutional neural network to extract features at different levels from low to high, then use BiLSTM to establish the correlation between features at different levels, and use CBAM module to extract local features related to image and emotion. The accuracy of the proposed method on ArtPhoto and Abstract data sets is improved compared with previous methods, which proves that the proposed method is effective in image sentiment analysis.

4.6. Ablation Experiment

In order to verify the effectiveness of the mixed attention mechanism and BiLSTM module in the MAML model, an ablation experiment was conducted on the abstract dataset, and the experimental results are shown in Table 4.It can be seen from Table 4 that the performance of the model is improved after the CBAM and BiLSTM modules are added to Resnet18 respectively, and the performance is the best after the two modules are added at the same time, thus verifying the effectiveness of the proposed method in this paper.

Table 3. Results of ablation experiments on an abstract dataset

Method	Accuracy
Resnet18	73.47
CBAM	78.57
BiLSTM	80.63
CBAM+BiLSTM	82.14

5. Summary

In this paper, we propose an image sentiment analysis model, MAML. It uses multiple branches in Resnet18 to extract multi-level features, and uses the dependency relationship between features at different levels through BiLSTM method to effectively integrate these features, and focuses on the main features in the emotion region through the mixed attention mechanism of space and channel, so as to make the final emotion features more discriminative. This method shows excellent performance on both Artphoto and abstract datasets. In the future, we will consider designing a more reasonable feature extraction network to solve the problem of sample imbalance and the global and local emotional features in images can be mined more accurately through deep learning methods, so as to further improve the effect of image emotion analysis.

Acknowledgments

This work was financially supported by Tianjin Natural Science Foundation of China (19JCYBJC18700).

References

[1] Machajdik J, Hanbury A. Affective image classification using features inspired by psychology and art theory[C]//Proceedings of the 18th ACM international conference on Multimedia. 2010: 83-92.

[2] Yanulevskaya V, van Gemert J C, RothK, et al. Emotional valence categorization using holistic image features[C]//2008 15th IEEE international conference on Image Processing. IEEE, 2008:101-104.

[3] Zhao S, Gao Y, Jiang X, et al. Exploring principles-of-art features for image emotion recognition[C]//Proceedings of the 22nd ACM international conference on Multimedia. 2014: 47-56.

[4] Xu L, Wang Z, Wu B, et al. Mdan: Multi-level dependent attention network for visual emotion analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9479-9488.

[5] Rao T, Li X, Zhang H, et al. Multi-level region-based convolutional neural network for image emotion classification [J]. Neurocomputing, 2019, 333: 429-439.

[6] Li B, Ren H, Jiang X, et al. SCEP—A new image dimensional emotion recognition model based on spatial and channel-wise attention mechanisms[J]. IEEE Access,2021, 9: 25278-25290.

[7] Peng K C, Chen T, Sadovnik A, et al. Amixed bag of emotions: Model, predict, and transfer emotion distributions

[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 860-868.

[8] Compton R J. The interface between emotion and attention: A review of evidence from psychology and neuroscience[J]. Behavioral and cognitive neuroscience reviews, 2003, 2(2): 115-129.

[9] You Q, Jin H, Luo J. Visual sentiment analysis by attending on local image regions[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).

[10] Zhao S, Yao X, Yang J, et al. Affective image content analysis: Two decades review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6729-6751.

[11] Song K, Yao T, Ling Q, et al. Boosting image sentiment analysis with visual attention[J]. Neurocomputing,2018, 312: 218-228.

[12] Yang J, She D, Lai Y K, et al. Weakly supervised coupled networks for visual sentiment analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7584-7592.

[13] Chen L, Zhang H, Xiao J, et al. Sca-cnn:Spatial and channel-wise attention in convolutional networks for image captioning [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5659-5667.

[14] Zhu X, Li L, Zhang W, et al. Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition [C]// IJCAI. 2017: 3595-3601.

[15] Siersdorfer S, Minack E, Deng F, et al. Analyzing and predicting sentiment of images on the social web[C]// Proceedings of the 18th ACM international conference on Multimedia. 2010: 71 5-718.

[16] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[17] You Q, Luo J, Jin H, et al. Building a large scale dataset for image emotion recognition:The fine print and the benchmark [C]//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[18] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[19] Rao T, Li X, Xu M. Learning multi-level deep representations for image emotion classification[J].Neural processing letters, 2020, 51: 2043-2061.

[20] Sowmyayani S, Rani P A J. Salient object based visual sentiment analysis by combining deep features and handcrafted features[J]. Multimedia Tools and Applications, 2022, 81(6): 7941-7955.

[21] Li W, Liu K, Zhang L, et al. Object detection based on an adaptive attention mechanism[J].Scientific Reports,2020, 10 (1): 11307.

[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[23] Chen L, Zhang H, Xiao J, et al. Sca-cnn:Spatial and channel-wise attention in convolutional networks for image captioning [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5659-5667.

[24] Song K, Yao T, Ling Q, et al. Boosting image sentiment analysis with visual attention[J]. Neurocomputing,2018, 312: 218-228.

[25] Yang J, She D, Lai Y K, et al. Weakly supervised coupled networks for visual sentiment analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7584-7592.

[26] Zhao S, Jia Z, Chen H, et al. PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression [C]// Proceedings of the 27th ACM international conference onmultimedia. 2019: 192-201.

[27] Woo S, Park J, Lee J Y, et al. Cbam: Convolutionalblock attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[28] Zhang H, Xu D, Luo G, et al. Learning multi-level representations for affective image recognition[J]. Neural Computing and Applications, 2022, 34(16): 14107-14120.

[29] Borth D, Ji R, Chen T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C]// Proceedings of the 21st ACM international conference on Multimedia. 2013: 223-232.

[30] Rao T, Xu M, Liu H, et al. Multi-scale blocks based image emotion classification using multiple instance learning[C]// 2016 IEEE International Conference on ImageProcessing (ICIP). IEEE, 2016: 634-638.

[31] You Q, Luo J, Jin H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks [C]//Proceedings of the AAAI conference on Artificial Intelligence. 2015, 29(1).

[32] Yang J, She D, Sun M, et al. Visual sentiment prediction based on automatic discovery of affective regions[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2513-2525.

[33] Xiong H, Liu Q, Song S, et al. Region-based convolutional neural network using group sparse regularization for image sentiment classification[J]. EURASIP Journal on Image and Video Processing, 2019, 2019(1): 1-9.

[34] Machajdik J, Hanbury A. Affective image classification using features inspired by psychology and art theory. Proceedings of the 18th ACM international conference on Multimedia, 2010. 1, 5.