

Multi-level region-based Convolutional Neural Network for image emotion classification

Tianrong Rao, Xiaoxu Li, Haimin Zhang, Min Xu*

Faculty of Engineering and Information Technology, University of Technology Sydney, Australia



ARTICLE INFO

Article history:

Received 3 July 2018

Revised 11 December 2018

Accepted 12 December 2018

Available online 1 January 2019

Communicated by Prof. Junwei Han

Keywords:

Image emotion classification

Multi-level

Region-based

Convolutional Neural Network

ABSTRACT

Analyzing emotional information of visual content has attracted growing attention for the tendency of internet users to share their feelings via images and videos online. In this paper, we investigate the problem of affective image analysis, which is very challenging due to its complexity and subjectivity. Previous research reveals that image emotion is related to low-level to high-level visual features from both global and local view, while most of the current approaches only focus on improving emotion recognition performance based on single-level visual features from a global view. Aiming to utilize different levels of visual features from both global and local view, we propose a multi-level region-based Convolutional Neural Network (CNN) framework to discover the sentimental response of local regions. We first employ Feature Pyramid Network (FPN) to extract multi-level deep representations. Then, an emotional region proposal method is used to generate proper local regions and remove excessive non-emotional regions for image emotion classification. Finally, to deal with the subjectivity in emotional labels, we propose a multi-task loss function to take the probabilities of images belonging to different emotion classes into consideration. Extensive experiments show that our method outperforms the state-of-the-art approaches on various commonly used benchmark datasets.

© 2018 Published by Elsevier B.V.

1. Introduction

With the popularity of photo-based social networks, more and more people tend to share their feelings through images on these social networks. Emotion classification has attracted increasing research interests nowadays. Psychological studies have revealed that different visual stimuli can evoke different types of humans' emotions [1,2]. Based on these studies, multimedia researchers tried to understand the emotion implied in different visual content.

Many studies investigate the relationship between images and emotions through mapping visual features to emotions. Low-level visual features are first used for image sentiment analysis [3,4]. Machajdik and Hanbury propose to combine different levels of visual features, including low-level features based on art theory and high-level image semantics, for image emotion classification [5]. Principles-of-arts features specialized for image emotion recognition are also designed to improve classification performance [6]. Recently, benefiting from the success of deep Convolutional Neural Network (CNN) on computer vision tasks [7], researchers have applied CNN, which can automatically learn deep features for

emotion classification, and demonstrated that the deep features outperform hand-crafted features on emotion classification [8,9].

However, analyzing image emotion is implicitly a challenging task compared to other traditional computer vision tasks, such as object detection and recognition, due to the two challenges of the complexity and subjectivity of emotions. For complexity, most of images can evoke different emotions rather than one pure emotion [10]. Previous methods for affective image analysis mainly rely on the single level of visual features extracted from the global perspective of the whole image, while ignoring the sentimental response of multi-level visual features from local regions which contribute to diverse emotion reaction for one image [11,12]. Fig. 1 shows sample images from different affective image datasets and the class activation maps of the sample images. We can find that image emotion is related to complex visual features from high-level to low-level and some local regions in images may contain more emotional information than other parts of images. For subjectivity, people with different cultural background may have various emotional reactions to the same image. It is unable to collect the hard emotional label of an image. Instead, emotion category is labeled with the probability is widely applied in affective image datasets. The uncertainty labels clearly improve the difficulty to build an accurate classifier for image emotion classification.

* Corresponding author.

E-mail address: Min.Xu@uts.edu.au (M. Xu).

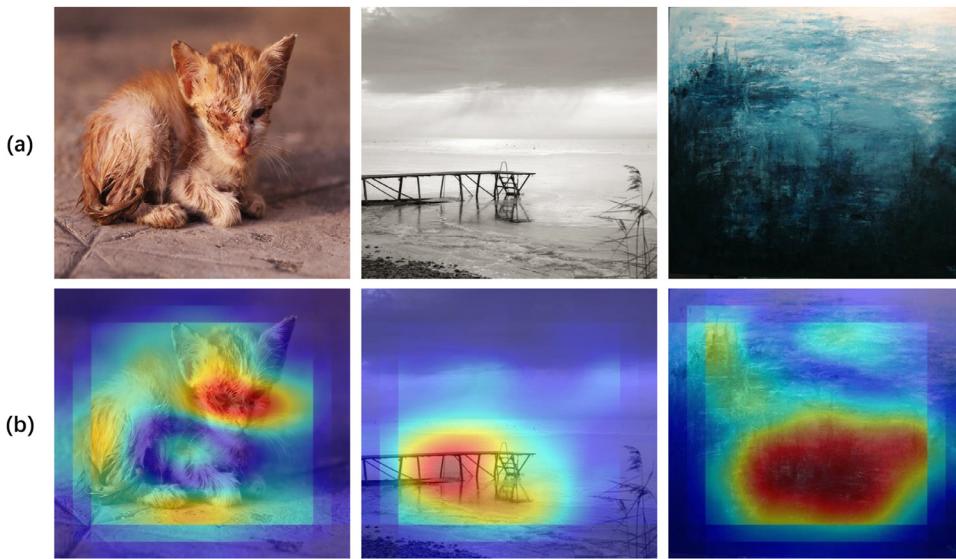


Fig. 1. (a) Sample images from different datasets that evoke the same emotion *sadness*. We can find out that image emotion is related to many factors. Left: web images whose emotions are mainly related to image semantics. Middle: IAPS photos whose emotions are mainly related to image aesthetics, such as compositions and emphasis. Right: abstract paintings whose emotions are mainly related to low-level visual features, such as texture and color. (b) Visualization of the class activation map for emotional category *sadness* of each sample image. As we can see, some regions within the image are more likely to evoke emotions than other parts of the image.

Researchers have already noticed the effect of multi-level visual features and local image regions for image emotion classification [9,13–16]. However, these methods only consider the local regions without taking a look from the global view or only focus on positive-negative emotion classification, a comprehensive consideration of both complexity and subjectivity in image emotion analysis is still a challenging problem.

Considering the aforementioned challenges, in this paper, we propose a multi-level region-based convolutional neural network that can automatically extract multi-level deep representations of local image regions. Multi-level deep features can better represent different kinds of affective images and utilizing features extracted from emotional regions can effectively avoid the noisy information containing in non-emotion regions. Moreover, a new loss function is proposed in this paper to estimate an emotion distribution derived from emotion class probability, which can effectively counteract the factor of subjectivity existing in image emotion labels. The overview of our framework is shown in Fig. 2. Emotional regions of different size are extracted based on different scales of feature maps which combine multi-level deep features. Subsequently, the local deep representations extracted from these emotional regions are combined with the global deep representations extracted from the whole image for emotion classification. Compared with existing methods mainly based on single-level visual features from a global view, the multi-level emotion information from both global and local view utilized in our model can provide a robust performance on various kinds of images.

The contributions of this paper are summarized as follows: (1) we employ a feature pyramid network(FPN) to extract multi-scale deep feature maps that related to image emotion. The multi-scale deep feature maps extracted from different convolutional layers can combine high-level semantic features with low-level deep features, and thus significantly improve the performance of emotion region detection. (2) We build a region-based CNN model that can effectively extract local emotional information from the emotional regions of the image. Ignoring the noisy information generating from non-emotional regions can significantly improve the emotion classification performance. (3) Image emotion labeling is a highly subjective task and the uncertain emotion labels will degrade the classification accuracy. Thus, we modify the loss function to consider the emotion class probability, rather than a hard class label,

into image emotion classification to overcome the subjectivity in emotion analysis.

Extensive experiments are conducted to evaluate our Multi-level R-CNN model on multiple datasets including Flickr&Instagram(FI) [8], IAPSsubset [5], ArtPhoto [5], etc. The experimental results demonstrate the effectiveness of our method for effectively detecting emotional regions with multi-level deep features and dealing with the problem of subjectivity existing in image emotion.

2. Related work

Visual emotion analysis on still images [17–19] and videos [20,21] has attracted increasing research interests nowadays. In this section, we review the development of image emotion analysis and region-based CNN which are closely related to this work.

2.1. Affective image analysis

For visual emotion classification, existing research can be roughly divided into methods in dimensional emotion space (DES) [10,22,23] and methods in categorical emotion states (CES) [5,6,24,25]. DES models, which utilize 3-D valence-arousal-control emotion space, 3-D natural-temporal-energetic connotative space, 3-D activity-weight-heat emotion factors, and/or 2-D valence-arousal emotion space, provide predictable and flexible descriptions for emotions. In CES models, computational results are mapped directly to one of a few basic categories, such as anger, excitement, sadness, etc. Compared to DES models, CES models are easy for people to understand and label, thus have been widely applied in recent studies. To compare our results with existing work, we adopt the CES model to classify emotions into 8 categories (positive emotion *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotion *Anger*, *Disgust*, *Fear*, *Sadness*) and 2 categories(positive and negative) predefined in a rigorous psychological study [26].

The visual features used for image emotion classification are designed and extracted from different levels [18,27]. Yanulevskaya et al. [28] first proposed to categorize emotions of artworks based on low-level features, including Gabor features and Wiccest features. Solli and Lenz [29] introduced a color-based emotion-related

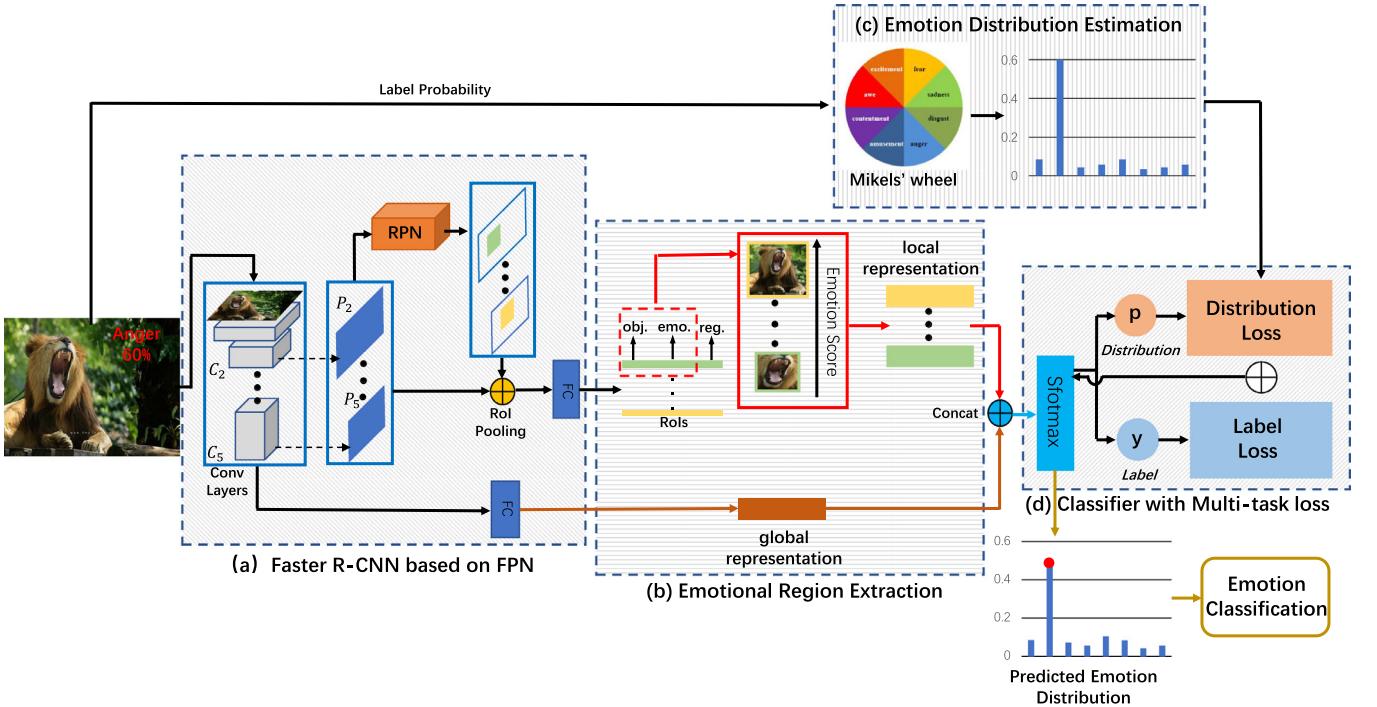


Fig. 2. The overview of the proposed framework. The framework consists 4 components: (a) faster R-CNN based on FPN, (b) emotional region extraction based, (c) emotion distribution estimation and (d) classifier with multi-task loss.

image descriptor, which is derived from psychophysical experiments, to classify images [13]. SIFT features extracted from both global view and local view were used for emotion prediction. Machajdik and Hanbury [5] defined a combination of rich hand-crafted mid-level features based on art and psychology theory, including *composition*, *color variance* and *texture*. Zhao et al. [6] introduced more robust and invariant mid-level visual features, which were designed according to art principles to capture information about image emotion. High-level adjective-noun pairs related to object detection were introduced for visual sentiment analysis in recent years [27,30]. However, these hand-crafted visual features have only been proven to be effective on several small datasets, whose images are selected from a few specific domains, e.g. abstract paintings and portrait photos. This limits the applications of image emotion classification on large-scale image datasets.

Considering the recent success of CNN-based approaches in many computer vision tasks, CNN based methods have also been employed in image emotion analysis. Peng et al. [11] first attempted to apply the CNN model in [7]. They fine-tuned the pre-trained convolutional neural network on ImageNet [31] and demonstrated that CNN model outperforms previous methods rely on different levels of handcrafted features on the Emotion6 dataset. You et al. [8] employed a progressive strategy to train a CNN model to detect image emotion on the large-scale dataset of web images. In [14], local emotional regions extracted using attention model were considered for sentiment analysis. However, most existing work only consider single-level visual features extracted from a global view, which limit their emotion classification performance due to the noisy from non-emotion regions within images and ignoring combining information from different levels including low-level visual features, mid-level image aesthetics and high-level semantics.

2.2. Region-based CNN

Our methods are based on region-based CNN (R-CNN) [32], which generates region proposals on CNN framework to localize

and classify objects in images. Then, by introducing supervised pre-training for an auxiliary and domain-specific fine-tuning, the object detection performance are significantly improved [33]. Girshick [34] further develops the R-CNN model to faster-RCNN model to reduce the training time and computing consumption, while improving the object detection accuracy. Ren et al. [35] combine the Region Proposal Network (RPN) with CNN architecture to share full-image convolutional features and predict object bounds and objectness simultaneously.

Compared to traditional region based CNNs, which are mainly used to find salient objects in images, in our method, we utilize R-CNN to find local regions that evoke emotion and use the local representations extracted from these regions as supplementary information for image emotion classification. This means that we are interested in regions with emotion rather than regions with objects. In other words, identified local regions contain not only objects and/or objects' parts, but also selected background surroundings of the objects [36].

3. Preliminaries

As shown in Fig. 2(a), candidates of emotional regions with multi-level deep features are extracted using faster R-CNN based on FPN.

3.1. Feature Pyramid Network (FPN)

To extract multi-level deep representations for image emotion analysis, a Feature Pyramid Network (FPN) [37] is employed to extract multi-scale feature maps. Compared to the existing pyramidal feature hierarchy structure in [9,19], in which the lower level feature maps are high-resolution but with low-level deep features that harm their representational capacity for object recognition and emotion classification. The detailed structure of FPN is shown in Fig. 3. As shown in the figure, FPN consists of two parts, a bottom-up pathway and a top-down pathway, between them is the lateral connections.

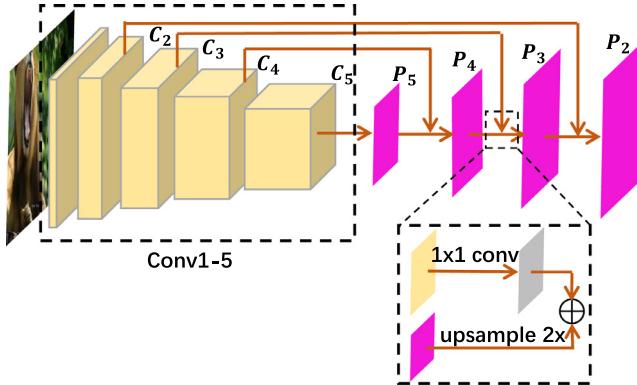


Fig. 3. Structure of Feature Pyramid Network (FPN).

The bottom-up pathway is the feed-forward computation of normal backbone convolutional network (e.g., [7,38,39]). In this paper, we use ResNet101 [39] as the backbone network. From the bottom-up pathway, feature hierarchy which contains feature maps in different size can be computed. The output of the last layer of each bottleneck in the ResNet101 is selected as the reference set of feature maps to create the pyramid. The output of these bottlenecks is defined as \$\{C_2, C_3, C_4, C_5\}\$ for conv2, conv3, conv4 and conv5 outputs, and note that conv1 is excluded for the pyramid due to the large memory consuming for the massive feature map.

The top-down pathway is used to combine different levels of feature maps extracted from bottom-up pathway. Feature map from the highest pyramid level, which is semantically stronger but spatially coarser, is upsampled to fit the size of lower-level feature maps in feature pyramid, which are higher resolution but only contains low-level deep features. The upsampled map is then merged with the corresponding bottom-up maps (a \$1 \times 1\$ convolutional layer is added behind each bottom-up map to reduce channel dimensions) by element-wise addition. The process is iterated until the last (finest resolution) merged map is generated. The set of final feature maps is defined as \$\{P_2, P_3, P_4, P_5\}\$, which is corresponding to \$\{C_2, C_3, C_4, C_5\}\$ in the same spatial size respectively.

3.2. Faster R-CNN

Detecting concrete visual objects in images has been widely studied in computer vision [35,40,41]. We briefly review the Faster R-CNN model [35], which is used to extract emotional region from the image in this work. Faster R-CNN is a two-stage detector mainly consisting of three major parts: shared bottom convolutional layers which is FPN in our model, a region proposal network (RPN) and a classifier built for region-of-interest (ROI). The detailed structure is shown in the left part of Fig. 2.

First, an input image is represented as multi-scale feature maps which combine different levels of deep features by FPN. Then, RPN generates candidate object proposals based on the feature maps. Since we replace the single-scale feature map using in Faster R-CNN with multi-scale feature maps, single-scale anchors with size \$\{32^2, 64^2, 128^2, 256^2\}\$ pixels are applied for multi-level feature maps \$\{P_2, P_3, P_4, P_5\}\$ with different receptive fields, respectively. Finally, ROI-pooling is used to extract features representing ROI and ROI-wise classifier predicts the category label based on the features. The training loss is composed of two terms:

$$L_{det} = L_{obj} + L_{reg} \quad (1)$$

here \$L_{obj}\$ is the classification loss over two classes (if the candidate object region contains an object or not). \$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)\$ is the regression loss on the box coordinates for better localization, in which \$t_i\$ is a 4-d vector representing the coordinates of the

predicted bounding box, \$t_i^*\$ is the coordinates of the ground-truth box and \$R\$ is the robust loss function(L1 smooth) defined in [34]. More detailed information about the architecture and training procedure of Faster R-CNN can be found in [35].

4. Emotion analysis using multi-level R-CNN

4.1. Emotional region extraction

Detecting concrete visual objects in images has been widely studied in computer vision [35,40,41]. However, compared to object detection, detecting emotional content is extremely challenging. The main difficulty is that both the concrete objects and the surrounding background contribute to image emotions [30,42]. Due to the strong co-occurrence relationships between objects and local emotional regions [43], we could still utilize object detection methods to select potential emotional regions. However, to select proper emotional regions from the candidate object proposals generating through RPN, we modify the Faster R-CNN to contain emotional information as shown in Fig. 2(b).

Following the definition of objectness score \$S_{obj}\$ in [35], which measures the membership to set of object classes vs. background, we define emotion score \$S_{emo}\$ to evaluate the probability of a region evoking emotions. To compute emotion score, a binary class label (of emotional region or not) is assigned to each anchor in RPN. The positive label is assigned to an anchor with the highest Intersection-over-Union (IoU) overlap with a ground-truth emotional region or an anchor that has an IoU overlap higher than 0.7 with any ground-truth emotional region. The negative label is assigned to anchor with IoU overlap lower than 0.3 with any ground-truth emotional regions. Using the samples collected from RPN, a softmax classifier can be trained to predict to probability \$p_{emo}\$ of the region evoking emotions.

In Faster R-CNN, we introduce the emotional region classifier into the ROI-wise classifier and fix the object classifier. The new training loss function is:

$$L_{det}^* = L_{emo} + L_{reg} \quad (2)$$

where \$L_{emo}\$ is the classification loss of the ROI being an emotional region or not. Therefore, the ROI-wise classifier can compute both the probability of the ROI evoking emotions \$p_{emo}\$ and the probability of the ROI containing an object \$p_{obj}\$. The Faster R-CNN can train on the loss \$L_{det}^*\$ related to emotional regions.

As we mentioned before, considering the emotional region is related to the probability of the region containing an object \$p_{obj}\$ and the probability of the region evoking emotions \$p_{emo}\$, the emotion score of the region can be computed considering both probabilities:

$$S_{emo} = \sqrt{p_{emo}^2 + p_{obj}^2} \quad (3)$$

The proposed emotion score \$S_{emo}\$ can reflect how likely a region evoking emotions. The 10 regions with highest emotion score are selected as the emotional regions of the image and used for image emotion classification.

4.2. Emotion distribution estimation

The majority voting strategy is widely employed to obtain the ground truth emotional label for most of affective image datasets [5,8]. Many images in these datasets have emotional labels with probabilities instead of hard emotional labels. To handle the impact of labeling image emotions with emotion class probabilities, we consider to either estimate an emotion distribution based on label probabilities or directly import label probabilities into a loss function for training. For emotion distribution estimation, since the

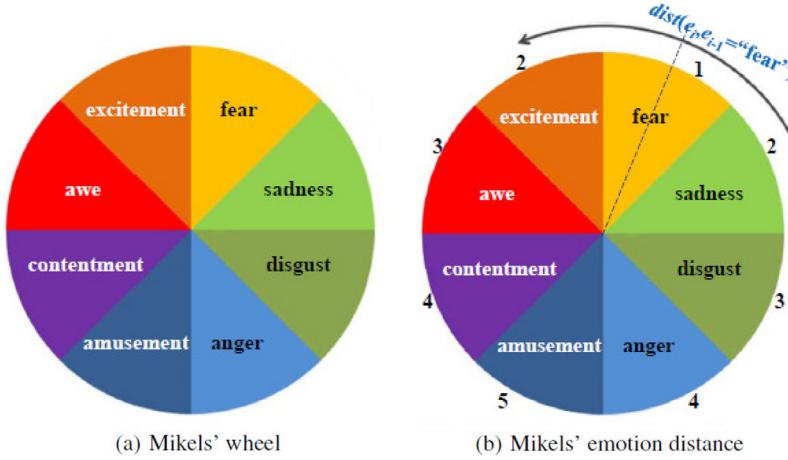


Fig. 4. Mikels' emotion wheel and example of emotion distance [10].

subjectivity existing in humans' emotional response to images, the emotional response to an image is more likely a distribution of several emotions rather than a single emotion.

Inspired by the study of emotion theory [44], the degree of similarity between two emotions, which determines the relationship of the two emotions, from similar to complete opposite, can be represented through Mikels' Wheel [10]. Fig. 4 shows Mikels' wheel and the method to compute emotion distance revealing the similarity between two emotions. The low distance d_{ij} between emotion i and emotion j indicates that the two emotions are similar to each other. Using Mikels' Wheel as weak prior knowledge, we can assign the probability of different emotion classes to an image based on the dominant emotion class of that image. Therefore, if we denote the image has an dominant emotion j with probability p_j^* , the emotion distribution for the i th emotion of the image can be generated through triangular distribution as shown in Fig. 2(c):

$$f(i) = \begin{cases} p_j^* & i = j \\ \frac{1}{d_{ij}}(1 - p_j) & i \neq j \\ \sum_{i \neq j} \frac{1}{d_{ij}} & \end{cases} \quad (4)$$

in which, the emotion classes being closer to the dominant emotion class are assigned with higher probabilities. The sum of all emotion class probabilities $\sum f(i)$ is normalized to 1.

4.3. Classifier and loss function

Through Faster R-CNN, a set of local deep representations of emotional regions is collected $\{X_{local}\}_{j=1}^K$, where K is the number of emotional regions extracted from one image. Considering that an image may not contain many local emotion regions, we only include top-ranked major emotion regions for classification, by setting $K = 10$. The global deep representation of the whole image X_{global} extracted from the ResNet101 is concatenated with the local deep representations $\{X_{local}\}_{j=1}^K$:

$$X = [X_{global}, \{X_{local}\}_{j=1}^K] \quad (5)$$

Followed by a softmax layer, X is transformed into a probability distribution of different emotions, where the emotion category with the highest probability is considered as the predicted label of the image. Considering the two methods (as discussed in Section 4.2) deal with the subjectivity existing in emotions, two loss functions can be applied in our approach.

Multi-task loss: taking both emotional label and estimated emotion distribution into account, the multi-task loss function consists

two terms:

$$L_{multi} = (1 - \lambda)L_{cls} + \lambda L_{ed} \quad (6)$$

where L_{cls} is the traditional classification loss, which can be computed as:

$$L_{cls} = -\sum_i y_i \log(p_i) \quad (7)$$

where $y = \{y_i | y_i \in \{0, 1\}, i = 1, \dots, n, \sum_{i=1}^n y_i = 1\}$ indicates the ground-truth label of the image, and p_i is the probability of an image belonging to the i th emotion category.

L_{ed} is the loss from emotion distribution $f(i)$. We employ the KL loss defined in [45]. λ controls the trade-off between the two weights. The KL loss is the measurement of the similarity between the emotion distribution $f(i)$ and the predicted emotion distribution p_i :

$$L_{ed} = -\sum_i f(i) \log(p_i) \quad (8)$$

The loss function can be optimized by stochastic gradient descent(SGD). We define $\{a_i | i = 1, 2, \dots, N\}$ to be the activation values of class i in the last fully connected layer. The gradient can be computed by:

$$\begin{aligned} \frac{\partial L}{\partial a_i} &= (1 - \lambda) \sum_i \frac{\partial L_{cls}}{\partial p_i} \frac{\partial p_i}{\partial a_i} + \lambda \sum_i \frac{\partial L_{ed}}{\partial p_i} \frac{\partial p_i}{\partial a_i} \\ &= p_i + (1 - \lambda)y_i + \lambda f(i) \end{aligned} \quad (9)$$

Loss with probability: another instinctive thought to deal with label with probability is to directly introduce label probability into loss function. Similar to [46], the classification loss with probability L_p can be defined as:

$$L_p = -\sum_i y_i \log(p_i^\theta), p_i^\theta = \frac{\exp(p_j^{*2} \cdot p_i)}{\sum_i \exp(p_j^{*2} \cdot p_i)} \quad (10)$$

The class prediction is weighted by the label probability p_j^* . By introducing the explicit simplifying assumption $p_j^* \sum_i \exp(p_j^{*2} \cdot p_i) \approx (\sum_i \exp(p_i))^{p_j^*}$ which becomes equal when $p_j^* \rightarrow 1$, the classification loss with probability can be rewritten as:

$$\begin{aligned} L_p &= -\sum_i \log(\exp(p_j^{*2} \cdot p_i)) + \log \left(\sum_i \exp(p_j^{*2} \cdot p_i) \right) \\ &\approx -p_j^{*2} \sum_i y_i \log(p_i^\theta) + \log \left(\frac{1}{p_j^{*2}} \right) \end{aligned} \quad (11)$$

The label with lower probability P_j^* will reduce the contribution of the classification loss. With the above equation, label probability is introduced into loss for training.

5. Experiments and results

In this section, we evaluate our model against state-of-the-art emotion classification methods through comprehensive experiments to demonstrate the effectiveness of our framework for different emotion classification tasks.

5.1. Dataset

Our experiments are carried out on normally used image emotion datasets:

Flickr and Instagram (FI) (8 categories) [8]: this dataset is collected from social websites using the names of emotion categories as searching keywords. Workers from Amazon Mechanical Turk (AMT) are then hired to further label the images. Finally, 23,308 well-labeled images are collected for emotion recognition.¹

EmotionROI (2 categories) [36]: the dataset contains 1,980 affective images from Flickr with labeled emotional regions. This dataset can be used for training the R-CNN.

IAPSsubset (8 categories) [5]: the *International Affective Picture System* (IAPS) is a standard stimulus image set, which has been widely used in affective image classification. IAPS consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, puppies, babies, animals, landscapes and others. Among all IAPS images, Mikels et al. [26] selected 395 images and mapped arousal and valence values of these images to the above mentioned eight discrete emotion categories.

ArtPhoto (8 categories) [5]: in the ArtPhoto dataset, 806 photos are selected from some art sharing sites by using the names of emotion categories as the search terms. The artists, who take the photos and upload them to the websites, determine emotion categories of the photos. The artists try to evoke a certain emotion for the viewers of the photo through the conscious manipulation of the emotional objects, lighting, colors, etc. In this dataset, each image is assigned to one of the eight aforementioned emotion categories. Considering the generalization of our model, we demonstrate the effectiveness of the components and adjust the parameters in our model on **FI** dataset, which contains over 20,000 different kinds of affective images, rather than other 4 small datasets, each of which only contains several hundreds of affective images for a specific domain.

Abstract (8 categories) [5]: this dataset consists of 228 abstract paintings. Unlike the images in the IAPS-Subset and ArtPhoto dataset, the images in the Abstract dataset represent the emotions through overall color and texture, instead of some emotional objects. In this dataset, each painting was voted by 14 different people to decide its emotion category. The emotion category with the most votes was selected as the emotion category of that image.

5.2. Implementation details

The backbone network of our model is the FPN [37]. In this paper, we follow a two-step training strategy. At the first step, we use the same strategy as in [37] to fine-tune the multi-level R-CNN on COCO pretrained weights using the EmotionROI dataset. Note, the aspect ratio of anchor is set to {1:1}. At the second step, the learning rate of the last two fully-connected layers are initialized as 0.001 and fine-tuned by SGD. The batch size is 128 and a total

of 100 epochs are run to update the parameters. All the experiments are carried out on four NVIDIA GTX 1080 GPUs with 32GB of GPU memory.

5.3. Baseline

We compare the proposed framework with the state-of-the-art methods for image emotion classification, which use various features, including hand-crafted features and deep features.

5.3.1. Hand-crafted features

- **GCH/LCH/GCH+BoW/LCH+BoW** [4]: 64-bin color histogram features for global view(GCH) and local view(LCH), and with SIFT-based bag-of-words features.
- **Zhao** [6]: low-level and mid-level features based on principle of art.
- **Rao(a)** [13]: SIFT-based bag-of-visual features for both global and local view based on the image blocks extracted from images.
- **SentiBank** [27]: 1200-dim adjective noun pairs(ANPs) features as mid-level representation with linear SVM classifier.

5.3.2. Deep features

- **AlexNet** [7]: AlexNet fine-tuned on ImageNet pre-trained weights.
- **VGG-16** [38]: VGGNet fine-tuned on ImageNet pre-trained weights.
- **ResNet101** [39]: ResNet Fine-tuned on ImageNet pre-trained weights.
- **DeepSentiBank** [47]: 2,089-dim ANPs features based on CNN.
- **PCNN** [48]: a novel progressive CNN architecture based on VG-GNet [38].
- **Rao(b)** [9]: a CNN architecture based on AlexNet with side branch to utilize multi-level deep features.
- **Zhu** [19]: a unified CNN-RNN architecture for visual emotion recognition.

5.4. Experimental validation

For methods using deep features, we first fine-tune them on the large scale dataset (**FI**). The **FI** dataset is split randomly into 80% training, 5% validation and 15% testing sets. For the 4 datasets (**FI**, **IAPSsubset**, **ArtPhoto**, **Abstract**), with 8 emotional categories(positive emotion *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotion *Anger*, *Disgust*, *Fear*, *Sadness*), we can convert them to 2 emotional categories with labeling 4 positive emotions as positive and 4 negative emotions as negative. To compare the results for all datasets, we present the classification results for both 8 emotional categories and 2 emotional categories.

5.4.1. The effectiveness of local emotional region

To demonstrate the effectiveness of considering the proposed local emotional regions. We design experiments performed on the **FI** dataset to compare: (1) ResNet101 [39] only using the global feature extracted from the last convolutional layer; (2) our framework only with features extracted from object regions extracted using Faster R-CNN with FPN [37]. (3) our framework only with features extracted from emotional regions; (4) our framework with object regions extracted using Faster R-CNN with FPN; and (5) our framework with features extracted from both the whole image and emotion regions. Table 1 shows the performance of the five different methods on the test set of **FI**. As shown in Table 1, compared to ResNet101, our method with object regions improves the performance by 7.65% for 8 classes **FI** and 7.08% for 2 classes **FI** and our method with emotional regions improves the performance by 14.64% and 12.84%. This reveals that emotional information from

¹ We have 23,164 labeled images as some images no longer exists on the Internet.

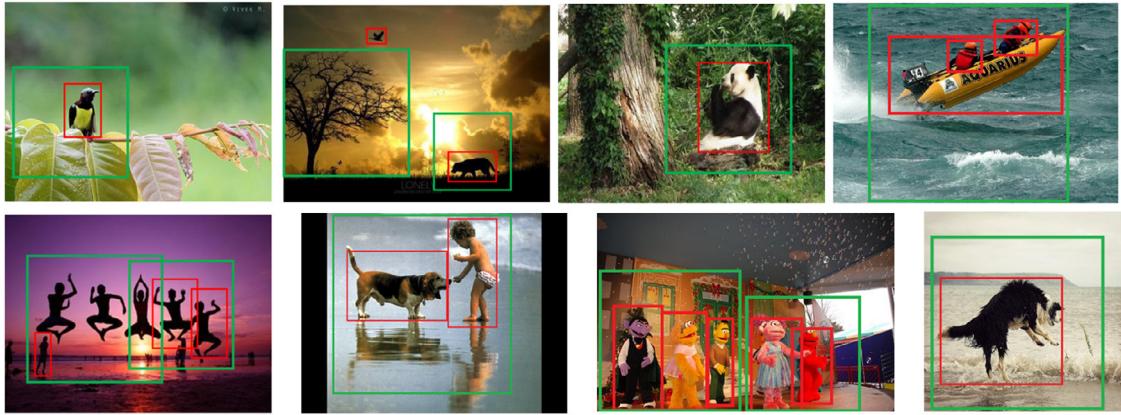


Fig. 5. Examples of object regions with highest objectness scores (red bounding box) and emotional regions with the highest emotion probability (green bounding box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1

Classification accuracy for both 8 classes and 2 classes on the test set of **FI**. Our method with different configurations, i.e., combining with object region and emotional region is compared with single column ResNet101 without local information and using object region and emotional region as local information only.

Method	FI (8 classes) (%)	FI (2 classes) (%)
ResNet101	60.82	74.67
Object regions only	54.82	88.44
Emotional regions only	59.78	72.57
Ours+object regions	68.47	81.75
Ours+emotional regions	75.46	87.51

local regions can largely improve the emotion classification accuracy than a single-column CNN-based global features extracted from the whole image. However, the emotion recognition performances reduced significantly without using features extracted from global view. This demonstrates the effectiveness of the global features extracted from the whole image.

Although both our framework with object regions and that with emotional regions improve the emotion classification performance, it is clear that using emotional regions in our method outperforms than using object regions by 6.99% and 5.76% for 8 classes **FI** and 2 classes **FI**, respectively. The results of our methods only using local features extracted from object regions and emotional regions also indicate that emotional regions contain more emotional information than object regions.

Fig. 5 shows examples of object regions and emotion regions. we can find that emotional regions are larger than object regions by containing objects and the surrounding background which may evoke emotions.

In Fig. 6, we show the confusion matrix of ResNet101 and our method with different configuration. It is clearly that applying local information in image emotion classification can improve the performance and provide a more balanced classification result for each emotion category. Especially applying emotional regions as local information in our method achieves the best classification result on most of the emotion categories. This also demonstrate the effectiveness of the emotional region.

5.4.2. The effectiveness of multi-level features

Previous methods have already indicated that multi-level features can significantly improve the image emotion classification performances [9,19]. However, the effectiveness of multi-level features in emotional region detection still needs to be proved. Fig. 7 performs the detection results on the testset of EmotionROI dataset. We notice that multi-level features improve

Table 2

Classification accuracy for both 8 classes and 2 classes on the test set of **FI** using popular CNN models and our method with traditional softmax loss(L_{cls}), multi-task loss(L_{multi}) and loss with probability (L_p).

Method	FI (8 classes) (%)	FI (2 classes) (%)
AlexNet+ L_{cls}	58.61	70.44
ResNet101+ L_{cls}	60.82	74.67
Ours+ L_{cls}	73.05	85.94
AlexNet+ L_p	57.44	68.72
ResNet101+ L_p	59.28	74.15
Ours+ L_p	73.58	86.07
AlexNet+ L_{multi}	60.32	72.83
ResNet101+ L_{multi}	62.77	77.15
Ours+ L_{multi}	75.46	87.51

both performances of object region detection and emotional region detection. The reason is that multi-level framework provides features maps with different scales of respective fields, which can effectively detect objects with different size in an image. What is more, the multi-level features can improve the accuracy of predicting emotional score [9], which can further promote the emotional region detection performance.

5.4.3. Choice of the loss functions

As we have discussed earlier, the subjectivity of the emotion is one of the main challenges for visual emotion recognition. Compared to traditional softmax loss L_{cls} widely used in different CNN models, the two loss functions we introduced before both taking label probability into account. We conduct experiments on the **FI** dataset for popular CNN model and our method using the aforementioned loss functions. The results are shown in Table 2. Though both L_{multi} and L_p introduce label probability into loss function, the performances of them are quite different. For L_p , the classification performance is worse than using L_{cls} while the performance of using L_{multi} is 2% better than that of using L_{cls} . The main reason is that, compared to L_p , L_{multi} introduces inter-class relationship, rather than simply abandon the low-probability labels, which contribute to the overall classification performance. Therefore, L_{multi} is more suitable for emotion classification and we apply the multi-task loss function in the following experiments.

5.4.4. Choice of parameter λ

The parameter λ controls the two portion of the proposed loss function. $\lambda = 0$ means the proposed loss function is equal to cross entropy loss and $\lambda = 1$ means the proposed loss function is equal to KL loss. Considering the estimate emotion distribution \hat{p}_i generated only using label probability and weak prior knowledge of

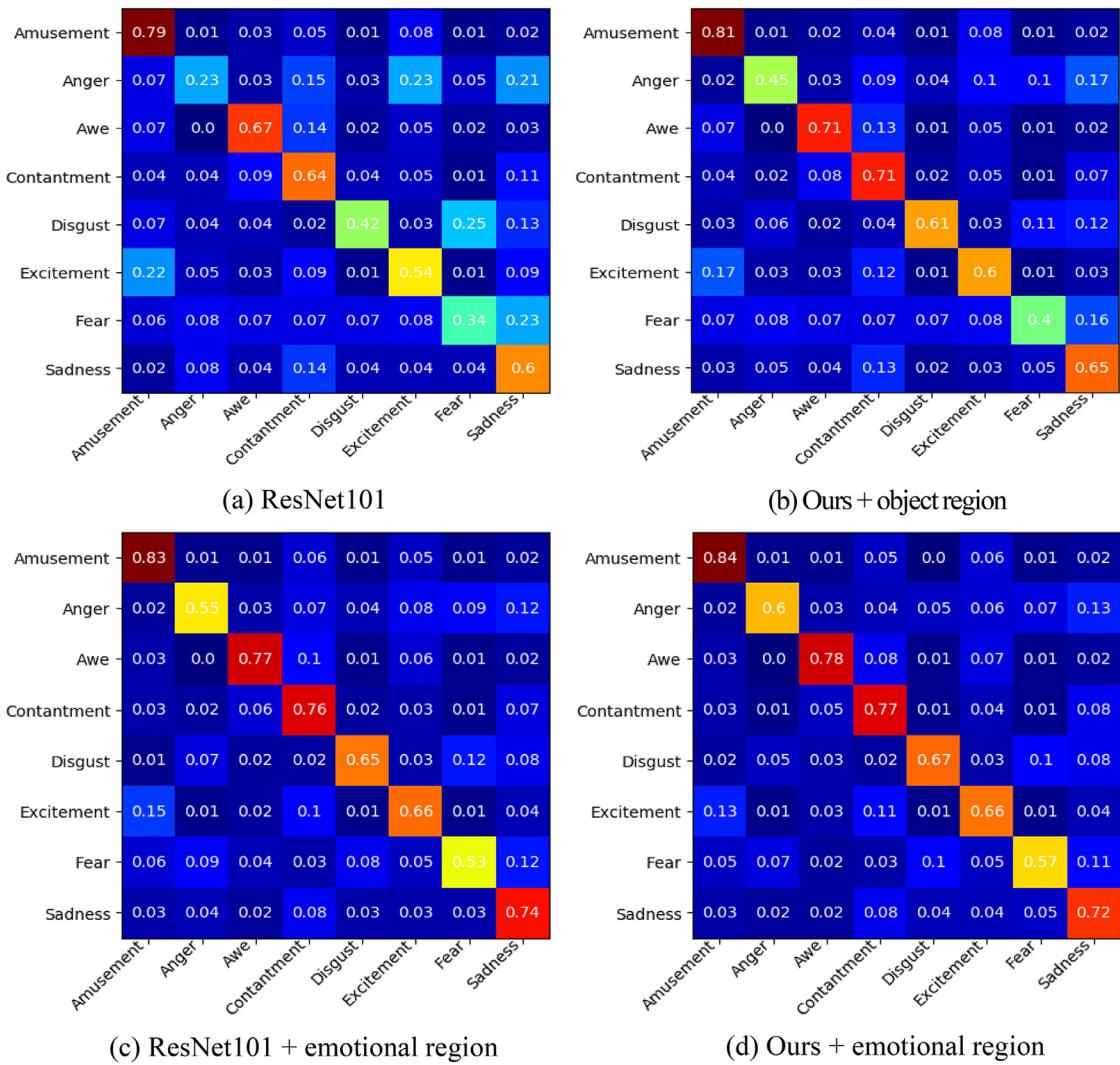


Fig. 6. Confusion matrix for our method with different configurations and ResNet101.

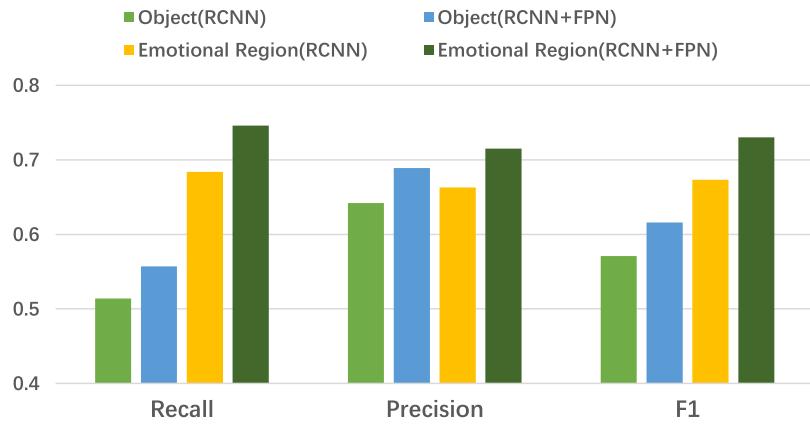


Fig. 7. Comparison of Emotional region detection performance on the test set of EmotionROI dataset using object detection methods and emotional region detection methods with single level features and multi-level features.

emotion distance defined in Mikels' wheel (Fig. 4), we do not recommend to set the parameter λ too high. Fig. 8 shows the effectiveness of parameter λ in the proposed loss function. When λ increases from 0 to 0.4, the classification performance is improved dramatically. However, further increasing over 0.5 leads to significant decreasing of the accuracy, since the large weight of L_{ed} introduces excess ambiguity. Therefore, we choose $\lambda = 0.4$ in all our

experiments for a comprehensive considering of the hard emotional label and emotion distribution.

5.5. Comparison with state-of-the-art methods

We represent the results of our method and state-of-the-art methods on the aforementioned 5 datasets (**FI**, **EmotionROI**,

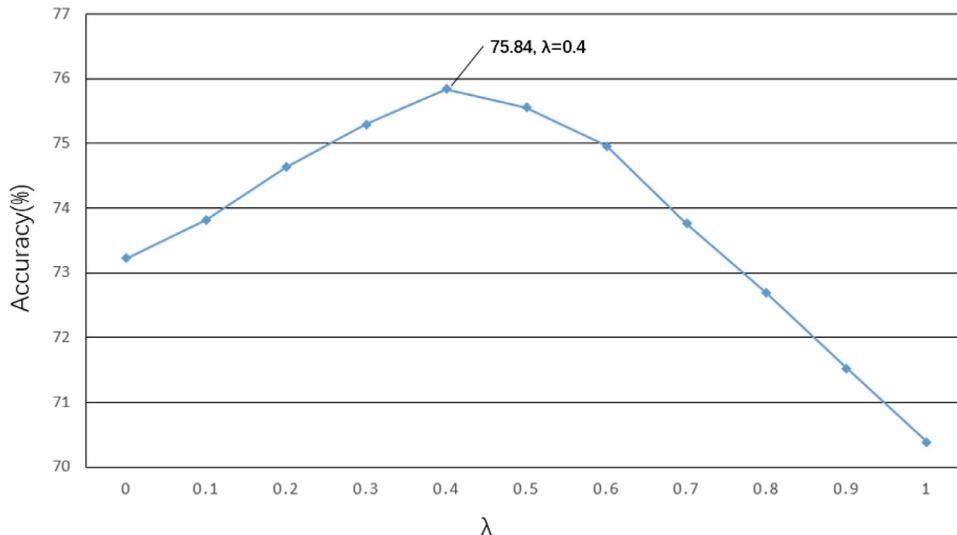


Fig. 8. Impact of different λ on the validation set of the **FI** dataset. $\lambda = 0.4$ achieves the best performance and we choose it in all our experiments.

Table 3

Classification results for different state-of-the-art methods on 5 different datasets. For **FI**, **IAPSsubset**, **Artphoto** and **Abstract**, we present classification results for both 2 classes and 8 classes.

Method	FI		IAPSsubset		ArtPhoto		Abstract		EmotionROI
	8 classes (%)	2 classes (%)	2 classes (%)						
GCH [4]	34.76	47.95	55.15	69.96	52.14	66.53	54.74	67.33	66.85
LCH [4]	32.42	45.37	43.15	52.84	50.41	64.33	55.45	70.93	63.79
GCH+BoW [4]	36.63	50.05	57.18	71.63	57.41	71.30	54.26	68.92	67.48
GCH+BoW [4]	34.58	48.26	47.61	56.07	52.05	66.72	58.39	72.48	65.67
Zhao [6]	46.52	58.42	63.61	65.77	66.37	68.42	60.60	66.23	73.45
Rao(a) [13]	51.67	62.79	70.32	78.34	69.74	71.53	62.17	67.82	74.51
SentiBank [27]	44.49	56.47	73.58	80.57	53.96	67.33	50.68	64.30	65.73
AlexNet [7]	58.61	68.63	72.24	84.58	67.03	69.27	61.96	65.49	71.60
VGG-16 [38]	59.75	73.95	74.78	87.20	68.16	70.48	62.41	65.88	72.49
ResNet101 [39]	61.82	75.76	75.09	88.15	69.36	71.08	63.56	66.64	73.92
DeepSentiBank [47]	53.16	64.39	75.88	86.31	68.54	70.26	66.46	69.07	70.38
PCNN [48]	56.16	73.59	76.87	88.65	68.93	71.47	67.17	70.26	74.06
Rao(b) [9]	67.24	79.54	78.08	90.53	69.75	74.83	67.81	71.96	78.99
Zhu [19]	73.03	84.26	82.39	91.38	71.63	75.50	68.45	73.88	80.52
Ours	75.46	87.51	84.71	93.66	74.58	78.36	70.77	77.28	82.94

IAPSsubset, **ArtPhoto** and **Abstract**). For fair comparison with **EmotionROI** dataset, which only has two emotional classes, we show the classification performance of the other 4 datasets for both 8 classes and 2 classes. The label conversion method is introduced in [Section 5.4](#). For the small-scale datasets (**IAPSsubset**, **ArtPhoto**, **Abstract** and **EmotionROI**), we can transfer the parameters of deep learning methods on the **FI** dataset. We follow the same experimental settings described in [\[5\]](#). Due to the imbalanced and limited number of images per emotion category, we employ the “one against all” strategy to train the classifier. The image samples from each category are randomly split into five batches and 5-fold cross validation strategy is used to evaluate the different methods.

[Table 3](#) shows the comparisons of our methods to several state-of-the-art methods, including methods using hand-crafted features and deep features. It is clear that methods using deep features outperform methods using hand-crafted feature on large-scale dataset **FIs**. However, hand-crafted features show their effectiveness for some specific kinds of images on small-scale dataset.

For hand-crafted features, low-level feature like color are very suitable to classify abstract paintings, which mainly consist of color and texture. While for other kinds of images, simple color feature seems not enough for emotion classification. Multi-level features are combined in Zhao’s method [\[6\]](#) and achieve acceptable result for the small-scale dataset. The reason is that image emotion is

related to various kinds of visual features from different levels, comprehensive consideration of different visual features can benefit the classification result. In Rao(a) [\[13\]](#), local emotional region is extracted using image segmentation method and represented with SIFT feature and bag-of-words. SIFT feature is a texture representation, which can be used to detect concrete objects, e.g. face, building, animal etc.. The performance of the method demonstrates the effectiveness of both concrete objects and local regions for image emotion analysis.

For deep features, the performances of three popular CNN frameworks, which are AlexNet [\[7\]](#), VGGNet [\[38\]](#) and ResNet [\[39\]](#) are first compared. We can find that as the CNN goes deeper, the emotion classification accuracy just slightly improves. The results show that high-level image semantics cannot be used for image emotion classification independently. Other deep methods utilize only one kind of features, like DeepSentiBank [\[47\]](#) and PCNN [\[48\]](#) also show limited performance. Both Rao(b) [\[9\]](#) and Zhu [\[19\]](#) utilize the multi-level deep features extracted from different level of CNN and achieve relatively high performance. Except the multi-level features, the regional information contained in lower levels of convolutional layers also contributes to the improvement.

Employed both multi-level deep features and local emotional regions, our framework outperforms both hand-crafted feature based methods and deep approaches in all datasets. Also our

method shows a robust performance on different kinds of images, such as abstracting paintings consisting of color and texture and images from **IAPSsubset** whose emotions are evoked by certain objects. This means our method effectively combine different levels of visual features from both global and local view.

6. Conclusion

In this paper, we investigate the problem of image emotion recognition. Inspired by the observation that multi-level features and local regions with high emotional response contribute much to image emotion, we propose a framework to automatically detect emotional regions on multi-level deep feature maps. The local emotional information extracted from emotional regions is combined with global information extracted from the whole image for image emotion classification. We also utilize the label probability of the affective images to leverage the ambiguity and subjectivity of the emotional labels. The experimental results show that our method outperforms the state-of-the-art methods on different affective image datasets.

As shown before, the detected affective regions contain not only the object, but also the surrounding areas of the object. Therefore, the emotion can be used as an anchor in the detection tasks like saliency detection [49], co-saliency detection [50], and object detection [51,52] in future research. Besides, an image may evoke multiple emotions in the same time. For future study, we plan to exploit relationship between different emotions to predict emotion distribution more precisely. We will also try to apply multi-label learning in image emotion analysis.

References

- [1] P.J. Lang, A bio-informational theory of emotional imagery, *Psychophysiology* 16 (6) (1979) 495–512.
- [2] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J.Z. Wang, J. Li, J. Luo, Aesthetics and emotions in images, *IEEE Signal Process. Mag.* 28 (5) (2011) 94–115.
- [3] H.-B. Kang, Affective content detection using HMMS, in: Proceedings of the ACM MM, 2003.
- [4] S. Siersdorfer, E. Minack, F. Deng, J. Hare, Analyzing and predicting sentiment of images on the social web, in: Proceedings of the ACM MM, 2010, pp. 715–718.
- [5] J. Machajdik, A. Hanbury, Affective image classification using features inspired by psychology and art theory, in: Proceedings of the ACM MM, 2010, pp. 83–92.
- [6] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, X. Sun, Exploring principles-of-art features for image emotion recognition, in: Proceedings of the ACM MM, 2014.
- [7] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the NIPS, 2012.
- [8] Q. You, J. Luo, H. Jin, J. Yang, Building a large scale dataset for image emotion recognition: The fine print and the benchmark, in: Proceedings of the AAAI, 2016.
- [9] T. Rao, M. Xu, D. Xu, Learning multi-level deep representations for image emotion classification, arXiv:1611.07145 (2016).
- [10] S. Zhao, H. Yao, Y. Gao, R. Ji, G. Ding, Continuous probability distribution prediction of image emotions via multi-task shared sparse regression, *IEEE Trans. Multimedia* 19 (3) (2017) 632–645.
- [11] K.-C. Peng, T. Chen, A. Sadovnik, A.C. Gallagher, A mixed bag of emotions: Model, predict, and transfer emotion distributions, in: Proceedings of the CVPR, 2015.
- [12] B. Li, W. Xiong, W. Hu, X. Ding, Context-aware affective images classification based on bilayer sparse representation, in: Proceedings of the ACM MM, 2012, pp. 721–724.
- [13] T. Rao, M. Xu, H. Liu, J. Wang, I. Burnett, Multi-scale blocks based image emotion classification using multiple instance learning, in: Proceedings of the ICIP, 2016.
- [14] Q. You, H. Jin, J. Luo, Visual sentiment analysis by attending on local image regions., in: Proceedings of the AAAI, 2017, pp. 231–237.
- [15] J. Yang, D. She, M. Sun, M.-M. Cheng, P. Rosin, L. Wang, Visual sentiment prediction based on automatic discovery of affective regions, *IEEE Trans. Multimedia* 20 (9) (2018) 2513–2525.
- [16] K. Song, T. Yao, Q. Ling, T. Mei, Boosting image sentiment analysis with visual attention, *Neurocomputing* 312 (2018) 218–228.
- [17] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, S.-F. Chang, Visual affect around the world: a large-scale multilingual visual sentiment ontology, in: Proceedings of the ACM MM, 2015, pp. 159–168.
- [18] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, T.-S. Chua, Predicting personalized emotion perceptions of social images, in: Proceedings of the ACM MM, 2016, pp. 1385–1394.
- [19] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, D. Xu, Dependency exploitation: a unified CNN-RNN approach for visual emotion recognition, in: Proceedings of the IJCAI, 2017, pp. 3595–3601.
- [20] C. Chen, Z. Wu, Y.-G. Jiang, Emotion in context: Deep semantic feature fusion for video emotion recognition, in: Proceedings of the ACM MM, 2016, pp. 127–131.
- [21] H. Yu, L. Gui, M. Madaio, A. Ogan, J. Cassell, L.-P. Morency, Temporally selective attention model for social and affective state recognition in multimedia content, in: Proceedings of the ACM MM, 2017, pp. 1743–1751.
- [22] M. Xu, J.S. Jin, S. Luo, L. Duan, Hierarchical movie affective content analysis based on arousal and valence features, in: Proceedings of the ACM MM, 2008.
- [23] S. Benini, L. Canini, R. Leonardi, A connotative space for supporting movie affective recommendation, *IEEE Trans. Multimedia* 13 (6) (2011) 1356–1370.
- [24] S. Zhao, G. Ding, Y. Gao, X. Zhao, Y. Tang, J. Han, H. Yao, Q. Huang, Discrete probability distribution prediction of image emotions with shared sparse learning, *IEEE Trans. Affect. Comput.* (1) (2018) 1.
- [25] S. Zhao, G. Ding, Y. Gao, J. Han, Learning visual emotion distributions via multi-modal features fusion, in: Proceedings of the ACM MM, 2017, pp. 369–377.
- [26] J.A. Mikels, B.L. Fredrickson, G.R. Larkin, C.M. Lindberg, S.J. Maglio, P.A. Reuter-Lorenz, Emotional category data on images from the international affective picture system, *Behav. Res. Methods* 37 (4) (2005) 626–630.
- [27] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: Proceedings of the ACM MM, 2013.
- [28] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, J.-M. Geusebroek, Emotional valence categorization using holistic image features, in: Proceedings of the ICIP, 2008.
- [29] M. Solli, R. Lenz, Color based bags-of-emotions, in: Proceedings of the CAIP, 2009.
- [30] T. Chen, F.X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, S.-F. Chang, Object-based visual sentiment concept analysis and application, in: Proceedings of the ACM MM, 2014.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the CVPR, 2009.
- [32] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the CVPR, 2014, pp. 580–587.
- [33] R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (1) (2016) 142–158.
- [34] R. Girshick, Fast r-cnn, arXiv:1504.08083 (2015).
- [35] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Proceedings of the NIPS, 2015.
- [36] K.-C. Peng, A. Sadovnik, A. Gallagher, T. Chen, Where do emotions come from? Predicting the emotion stimuli map, in: Proceedings of the ICIP, 2016, pp. 614–618.
- [37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the CVPR, 2017, p. 4.
- [38] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *CoRR* (2014). abs/1409.1556.
- [39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the CVPR, 2016, pp. 770–778.
- [40] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the CVPR, 2016, pp. 779–788.
- [41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, in: Proceedings of the ECCV, 2016, pp. 21–37.
- [42] M. Sun, J. Yang, K. Wang, H. Shen, Discovering affective regions in deep convolutional neural networks for visual sentiment prediction, in: Proceedings of the ICME, 2016, pp. 1–6.
- [43] J. Wang, J. Fu, Y. Xu, T. Mei, Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks, in: Proceedings of the IJCAI, 2016, pp. 3484–3490.
- [44] R. Plutchik, The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, *Am. Sci.* 89 (4) (2001) 344–350.
- [45] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, *IEEE Trans. Image Process.* 26 (6) (2017) 2825–2838.
- [46] Y. Gal, Uncertainty in deep learning, Ph.D. thesis, 2016.
- [47] T. Chen, D. Borth, T. Darrell, S.-F. Chang, Deepsentibank: visual sentiment concept classification with deep convolutional neural networks, arXiv:1410.8586 (2014).
- [48] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks, in: Proceedings of the AAAI, 2015, pp. 381–388.
- [49] J. Han, H. Chen, N. Liu, C. Yan, X. Li, CNNS-based RGB-D saliency detection via cross-view transfer and multiview fusion, *IEEE Trans. Cybern.* (2017).
- [50] J. Han, G. Cheng, Z. Li, D. Zhang, A unified metric learning-based framework for co-saliency detection, *IEEE Trans. Circ. Syst. Video Technol.* 28 (10) (2018) 2473–2483.
- [51] G. Cheng, J. Han, P. Zhou, D. Xu, Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection, *IEEE Trans. Image Process.* 28 (1) (2019) 265–278.

- [52] D. Zhang, J. Han, L. Zhao, D. Meng, Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework, *Int. J. Comput. Vis.* (2018) 1–18.



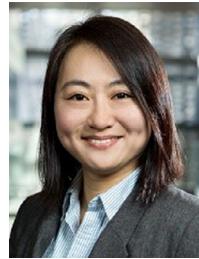
Tianrong Rao received the B.E. and M.S. degrees from the University of Science and Technology of China, Hefei, China, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree in computer science with the University of Technology Sydney, Ultimo, NSW, Australia. His current research interests include multimedia, computer vision, and machine learning.



Haimin Zhang received the Bachelors degree from Zhejiang Sci-Tech University, Hangzhou, China, and the Masters degree from Nankai University, Tianjin, China. He is currently working toward the Ph.D. degree at the University of Technology Sydney, Ultimo, NSW, Australia. His research interests include multimedia content analysis, computer vision, and machine learning.



Xiaoxu Li received Master's degree and Bachelors degree from University of Science and Technology of China, respectively. He is currently pursuing his Ph.D. degree at the University of Technology Sydney, Ultimo, NSW, Australia. His research interests focus in multimedia data analytics, computer vision and machine learning.



Min Xu received Ph.D. degree of IT from University of Newcastle, Australia, Master degree of Science (Computing) from National University of Singapore, and Bachelor degree of Engineering from University of Science and Technology of China respectively. She is currently a Senior Lecturer with the University of Technology Sydney, Ultimo, NSW, Australia. Her current research interests include multimedia processing, computer vision and machine learning.