Neurocomputing 328 (2019) 56–68

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

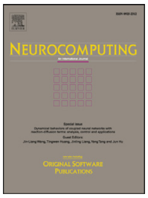


Image captioning by incorporating affective concepts learned from both visual and textual components

Jufeng Yang[a,*], Yan Sun[a], Jie Liang[a], Bo Ren[a], Shang-Hong Lai[b]

[a] *College of Computer Science, Nankai University, Tongyan Road #38, Tianjin 300458, China*
[b] *Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan*

ARTICLE INFO

Article history:
Received 13 November 2017
Revised 7 March 2018
Accepted 22 March 2018
Available online 20 August 2018

Keywords:
Image captioning
Affective concepts
Emotion recognition

ABSTRACT

Automatically generating a natural sentence describing the content of an image has been extensively researched in artificial intelligence recently, and it bridges the gap between computer vision and natural language processing communities. Most of existing captioning frameworks rely heavily on the visual content, while rarely being aware of the sentimental information. In this paper, we introduce the affective concepts to enhance the emotion expressibility of text descriptions. We achieve this goal by composing appropriate emotional concepts to sentences, which is calculated from large-scale visual and textual repositories by learning both content and linguistic modules. We extract visual and textual representations respectively, followed by combining the latent codes of the two components into a low-dimensional subspace. After that, we decode the combined latent representations and finally generate the affective image captions. We evaluate our method on the SentiCap dataset, which was established with sentimental adjective noun pairs, and evaluate the emotional descriptions with several qualitative and human inception metrics. The experimental results demonstrate the capability of our method for analyzing the latent emotion of an image and providing the affective description which caters to human cognition.

1. Introduction

In the community of artificial intelligence, automatically describing the visual content of images is a very challenging task. This kind of task is often referred as image captioning, which aims to “paint a picture in your mind’s eye” [1]. In the past few years, many researche rs have made considerable progress in the development of image captioning while large corpora with paired image and descriptions (e.g., MS COCO [2] and Flickr8k [3]) has been constructed to assist the task of image description generation. While existing approaches pay much attention to the object recognition in the image content, they lack the ability of giving emotional expressions. While object recognition reflects the objective existence in images, the emotions for an image arise in the viewer are highly subjective [4].

Emotion always serves as an important factor in human perception and helps to produce vivid images in human minds. In Fig. 1, we show comparison between captions derived from traditional content-based captioning methods and the proposed method based on affective concepts. As shown in the examples, the sentences produced by the traditional captioning methods usually do not contain emotional components in the generated descriptions. Since traditional captions focus on articulating visual contents in images, the generated sentences are usually lack of distinctiveness and attractiveness since they are formed based on similar factual concepts. In contrast, it is easy for us to find that the sentence ‘The two cats sitting on the floor looks happily’ in Fig. 1 describes the cat with orange fur. The incorporation of emotional expression makes it much more distinctive in the descriptions of images. Moreover, the expression of emotional content will enrich the expressibility of sentences and make them more attractive. Compared to the flat description that most current captioning systems have produced, the emotional captions induce attractive and distinctive descriptions.

Note there are several difficulties when combining the emotional expressions with factual descriptions [5]. At first, not all images have a strong sentiment polarity of either positive or negative emotion. According to the human perception investigation, emotion ambiguity as well as nonexistences exist in a large volume of image [6]. For these images, we should allow users to control the recognition of emotions and the polarity of sentimental captions. Then, the extracted affective concepts are discrete in the original space, which needs to be projected into a quantitative continuous space for the following operation. Finally, the set of affective concepts is coarse if we directly retrieve the large-scale textual

[*] Corresponding author.
E-mail address: yangjufeng@nankai.edu.cn (J. Yang).

https://doi.org/10.1016/j.neucom.2018.03.078
0925-2312/

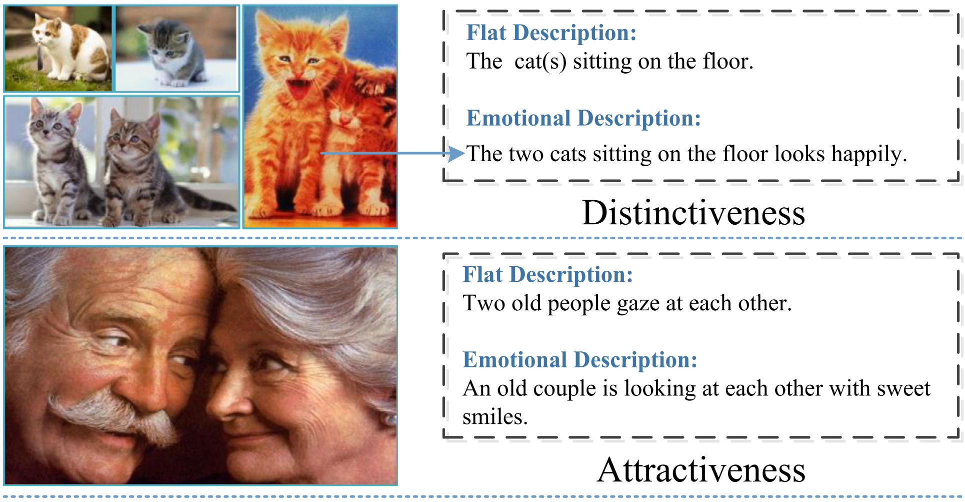

Flat Description:
The cat(s) sitting on the floor.
Emotional Description:
The two cats sitting on the floor looks happily.
Distinctiveness
Flat Description:
Two old people gaze at each other.
Emotional Description:
An old couple is looking at each other with sweet smiles.
Attractiveness

Fig. 1. The comparison between traditional content-based captioning methods and the proposed algorithm which incorporates various affective concepts. The descriptions derived from our algorithm show advantages against the content-based methods in mainly two aspects, i.e., distinctiveness and attractiveness. **Top:** The illustration of distinctiveness. Traditional caption gives a description of what exists in the image and what it is doing, which can be simply copied by many images with similar scene (e.g., the four cat images can be described with the same caption). This can be handled with the proposed affective incorporation. **Bottom:** The example of attractiveness. The traditional caption is also insipid with the factual illustration of objects and actions. The incorporation of more affective adverbs introduces a vivid component of the sentence, which attracts more attention of human viewers.

corpora. Therefore, we need to do proper pre-processing to filter the noisy and unusual concepts.

In this paper, we propose to generate vivid image descriptions, i.e., captions with distinctiveness and attractiveness, by incorporating emotional expressions. To address this problem, we incorporate the information of lingual corpora and visual representations of the emotional units. Two basic components are utilized in our system to generate vivid image descriptions: Long Short-term Memory (LSTM) [7] and affective concepts modeling. The employed LSTM module for caption generation with the CNN and RNN models follows the encoder-decoder framework [8]. We apply this module to generate vivid image captions for assembling basic sentences. In addition, we also introduce the affective concepts into the framework which generates emotional expressions with the following principles: (1) linking to emotions; (2) being frequently used; (3) can be detected reasonably. We exploit psychological theory and textual corpora to extract a candidate list of affective concepts specific to the domain of emotional expressions. With pre-trained captions and affective concepts, we map the emotional elements to the parse tree of sentence following Kuzentsova et al. [9]. Note that affective captions are related to emotional content and textual sentences, cross-modal features are demanded to satisfy the generation. The work [10] also illustrated the superiority of multi-scale evidences in many tasks, motivating us to generate emotional captions by combining lingual information and visual data to construct affective descriptions via combining emotional concepts with prepared data. To effectively model emotions, we mainly consider two key aspects as follows. First, there are extensive works which investigate image sentiments by solving the affective image classification problem. Effective emotion detectors, such as SentiBank [11], have been designed for affective predictions. A large-scale visual sentiment ontology is constructed as well based on the Plutchnik's Wheel of Emotions [12]. Moreover, as rare work models affective content in sentences, it is required to enhance the compatibility between visual sentimental features and textual representations for the generation of affective descriptions [13]. While we leverage visual detectors to capture emotional information in the content module, we further utilize a linguistic module so as to integrate both textual and visual components.

In summary, our contributions can be summarized in two folds. First, to the best of our knowledge, we are the first to incorporate multiple affective expressions in image captioning. Second, we gather the candidate list of affective concepts and incorporate both textual and visual components to generate distinctive and attractive captions. Extensive experiments on both automatic metrics and crowd-sourcing metrics demonstrate the efficacy of the proposed framework to generate captions with distinctiveness and attractiveness.

2. Related work

Automatically generating emotional descriptions requires both image captioning and sentiment recognition in the field of artificial intelligence. While traditional image captioning systems provide a factual description of the image contents, including the relations among each other and the activities they are involved, the recent advances of affective image classification tasks makes it possible to automatically generate vivid descriptions with attractive styles. In this section, we discuss the related works by summarizing image captioning tasks investigated in the field, and review how emotions for a given image can be identified with computer vision techniques.

2.1. Traditional image captioning systems

Automatically describing the content of a given image has become a focal task nowadays which connects the computer vision (CV) and natural language processing (NLP) [8]. Most early image captioning systems aim at generating natural language descriptions for images and their regions, which relies on the similarity retrieval with the assistances of objects and attributes [3,14,15], integrating sentence fragments (e.g., object-scene-action [14], subject-verb-object [16] and object-attribute-propositions [15]) and global image context [17].

In general, the paradigms which handle the image captioning problems leverage the well-labeled datasets, including images and their sentence descriptions, to learn about the intrinsic correspondence between the language and visual component. They can be split into two families, i.e., top-down and bottom-up based methods. The top-down based strategy targets at transforming images into sentences directly via learning a global latent representation. Contrarily, the bottom-up based method picks out appropriate words which correspond to different parts of the given image followed by combining them into a sentence via proper recurrent networks. With the development of recurrent neural networks (RNNs), the top-down paradigm shows favorable performance in various captioning tasks, which induce powerful deep formulations in the community [14].

In the generation of image caption, convolutional neural networks (CNN) bring great advances due to their strong capability to learn distinctive representations of the visual contents [18–20]. Also, for the captioning of a given image which is related to both visual recognition and description generation, the word-level language model as well as the visual formulation play very important roles in the task [8]. Recently, the progress of computation power and large-scale well-labeled datasets boosted the performance of deep language models, including RNNs [21]. Subsequently, researchers proposed to incorporate the CNN-based image representations and RNN-based language models for generating better captions. Vinvals et al. [8] suggested to seed the RNN-based models with CNN features, which makes great progress in the tasks of image description generation. Inspired by that, the incorporation of CNNs and RNNs was adopted in many works [22,23]. Generally, the CNN model serves as a high-level feature extractor while RNN models the sequence of words for generating captions for a given image. Furthermore, the CNN and RNN models are integrated into a multi-modal framework (termed as CNN+RNN in the remainder) [24–26], in which the CNN and RNN interact with each other in the multi-modal space and predict captions word by word. In addition, the long short term memory (LSTM) [7] unit has been widely adopted in RNN-based language models, which preserves long-term information and prevents the overfitting problem. Due to its computational efficacy and simplicity, we utilize the competitive system of CNN+RNN with LSTM units as our basic sentence generation engine in this paper.

Under the framework of CNN+RNN, the deep captioning tasks can be grouped into three orientations with different attentions, i.e., generating descriptions based on the global information or local components of the image, discovering new concepts beyond the image. Generating description for the global image is the most popular task investigated by researchers, which is to generate more natural and accurate description for a given image. Most works [27–29] utilized CNN models to extract features for the whole image and then fed the learned representations into RNN for descriptions. This kind of works focused on the significant objects or scenes contained in images, where the task assigns one label for each image. Recently, the rapid development of object detection induces the ability of models that efficiently identify and label multiple salient regions in an image [30]. Therefore, the systems [31,32] can focus on the salient component during each step, which trains a model to predict a set of descriptions across regions of an image for richer and interesting descriptions. For example, Justin et al. [33] proposed a fully convolutional localize network (FCLN) for dense captioning based on object detection and language models. The FCLN model detects a set of meaningful regions and associates them with captions. There are also systems [34–37] based on the local regions that expand the complexity of label space from a fixed set of categories about the sequence of words, making them able to express significantly richer concepts. Apart from the contributions with accurate descriptions and precise localization, the coverage of captioning is expanded as well. [25,38] proposed to describe the objects which never occur in the dataset. [25] constructed three novel concept datasets supplementing the information shortage of objects while [38] learned the knowledge from unpaired visual and text data.

[39] proposed to add adjectives in the descriptions on the basis of SentiBank [11]. According to [39], the emotional information is also an important component in the generation of captions, of which images in the range of some adjective nouns pairs (ANPs) are crawled from the Flickr without any checkout. With so much contribution devoted to the fields of image captioning, we proposed to introduce richer emotion concept into the image description. Unlike prior works which focused on the coverage of object recognition, we aim to convey emotional components hidden in images for a vivid expression.

2.2. Sentiment recognition

Strong sentiments embedded in images strengthen the opinion conveyed in the content, which can influence others effectively [40]. The awareness of emotion components included in a picture will benefit the understanding of the image for people. In the field of image captioning, concrete object concepts have been widely studied, of which the universal forms to model descriptions for the whole image has been adopted in a wide range of works [19,28,29]. Recently, researchers suggested that apart from visual objects, emotions contained in images should be considered and fusion on various modality has been studied as well [41,42]. The fusion works motivate us to encode sentiment recognition from both visual and textual channel in the captioning, so that the inter-modality knowledge can be well integrated. Specifically, they pointed out that abundant human intuition and emotion are hidden in creative applications such as visual art and music, but existing works lack the ability to model and utilize this information [43–45]. Image sentiment attracts more and more attention nowadays, and many researchers attempt to model emotion existence contained in images. Particularly, affective image classification is the hottest topic in emotion modeling.

Inspired by psychological and art theories [46], Machajdik et al. [4] proposed to predict emotions with low-level features, e.g., texture and color, extracted from images. Furthermore, the mid-level representations are studied to fill the semantic gaps between low-level features in human intuition. Zhao et al. [47] came up with the principle-of-arts features to predict the emotion for a given image. Meanwhile, Borth et al. [11] utilized the rich repository of image sources on the web to generate an emotion detector, which is of 1200 dimensions corresponding to the Adjective Nouns Pairs(ANPs). There are also many works [48,49] tried to improve the affective image classification variously, for example, You et al. [50,51] incorporated large-scale noisy machine labeled data to improve the performance of the recognition.

More recently, a deep neural network has been introduced extensively to deal with emotion modeling problems. For example, [52] extended the SentiBank to DeepSentiBank based on the Caffe model [53]. Besides, trial on the deep architecture[50] and reliable benchmark[54] was established as well. Images have increased rapidly in human life, not only existed as a figural narration but also a carrier of affects. Nowadays, 'kid is not only a kid' in human intuition, the mind state of the child is always attractive, but the feeling is also brought from animals in the picture. Wang et al. [46] suggested the *Emotional Semantic Image Retrival* (ESIR), which is built to mimic human decisions and provide the users with tools to integrate emotional components to their creative works. [4] classified images from the affective aspect by using low-level features, extracted texture and color information from images. [11] crawled large-scale images from the web to train a sentiment detector targeting at affective image classification to learn what emotions are associated with the pictures. Meanwhile, vividness is a test indicator in the field of mental imagery [55]. Though the [39] modeled adjective descriptions with the aid of

ANPs, sentiments haven't been presented directly or clearly since ANPs are essentially an assistant tool for emotion prediction.

In fact, there are already extensive works attempting to incorporate novel concepts to enhance the descriptiveness of image captions. It is known that the large-scale data is important for captioning knowledge learning. Although there are image-caption datasets, such as the MS-COCO and Flickr8k, which provide large volume of paired images and sentences, these datasets may not be enough to introduce novel concepts into the image captioning. Mao et al. [25] proposed to handle the problem of learning new categories of objects from a handful of examples due to the importance of data transfer. They enlarge the word dictionary to cover novel concepts for the description followed by adopting a transposed weight strategy for incorporation. Hendricks et al. [38] further developed the task with deep lexical classifier and caption model. They handle objects which are unseen in paired images and sentences with independent text corpora and train the lexical classifier via fine-tuned CNNs. The captioning model integrates parameters from the lexical classifier and the language model for generating image descriptions with unseen objects. Besides extensions on cognitive-level information, Gan et al. [56] introduced the stylized factors to enrich the expressibility of caption and make it more attractive. They set this kind of tasks with humorous and romantic styles. Since users always struggle to come up with an attractive title when uploading images, it is valuable if the machine could automatically recommend attractive captions based on the content of images. A novel framework, named StyleNet, enables the production of descriptions with styles using monolingual stylized language corpus and standard captioning datasets. Inspired by these works, we propose a framework which incorporates two basic modules to generate emotional descriptions, i.e., the caption model and the affective concepts model. The caption model refers to the standard processing of language via the framework of CNN+RNN, as was introduced previously. To the best of our knowledge, this is the first work that introduces the rich affective concept indicating visual emotions proposed into the image captioning. Nevertheless, it is hard to capture the inherent connections between various emotions and corresponding sentences. To draw the novel concepts into the original sources, [57,58] propose to cope such elements as a bag and instances. In details, they discover the joint configurations whose the labels are assigned to the instances. Motivated by this strategy, we propose a candidate list of affective concepts with psychological theory and map them into word embeddings via large text corpora. In the generate of affective descriptions, we aim to pick out proper affective concepts and fuse into text sentences. For the images, we train emotion classifiers with large sentiment datasets. The final affective concepts are learned through weighted function seeded by visual features and word-embeddings.

3. Preliminary

In this paper, we deal with the generation of emotional descriptions. It is common to train the image captioning models using the paired image and sentence data, *i.e.*, the MS-COCO and Flickr8k datasets. However, there is no existing paired data considering emotional expressions. The construction of well-labeled training data costs too much human-resources and the retraining of deep captioning models is time consumption. Therefore, we explore algorithms to address the problem with relevant textual information and visual sources. Two main modules are developed in this work: the emotion classifier and the captioning model. The captioning model refers to elementary descriptions which reflects substantial visual contents while the emotional classifier recognizes emotional polarity in images. Then we produce affective descriptions by integrating the two basic components with respect to the factual information and affective concepts.

3.1. Image captioning with LSTM

We adopt the joint CNN+RNN architecture and integrate the emotion classifier for the generation of descriptions with affective concepts.

The joint CNN+RNN architecture takes an image I as input and is trained to maximize the likelihood $p(S|I)$ of producing a target sequence of words $S = \{s_1, s_2, \ldots, s_t\}$, where each word s_t is organized adequately to describe the visual contents of each image. In this framework, the CNN model captures visual information contained in images and sends the features into RNN as a latent code for generating the descriptions. The CNN model has shown great performance in producing a rich representation of the input image by embedding it to a fixed-length vector. Thus, the encoder 'CNN' creates the proper features and sends it to the decoder 'RNN' to generate sentences.

Generally, LSTM [7] is adapted to the model of RNN as an important unit dealing with long term information and solving the vanishing and exploding gradients problems of conventional RNN architectures. The core of LSTM is a memory cell c encoding the knowledge of the input at each time step that has been observed and the gates which determine when and how much information should be conveyed to other modules. The "gates" take control of the memory cell via a binary value, i.e., 0 or 1. The collaboration of three gates i.e., the input gate i_t, output gate o_t and forget gate f_t, determine what to be output and updated through the memory cell, which are defined respectively as follows:

$$i_t = sig(W_{ix}x_t + W_{ih}h_{t-1}) \tag{1}$$

$$o_t = sig(W_{ox}x_t + W_{oh}h_{t-1}) \tag{2}$$

$$f_t = sig(W_{fx}x_t + W_{fh}h_{t-1}) \tag{3}$$

where t denotes the time step under the updating rule, x_t is the current state and h_t is the hidden state at time step t for each gate, $sig(\cdot)$ refers to the sigmoid function and W denotes the parameters which need to be learned in the framework. In addition, there is a memory cell c_t which takes the previous memory c_{t-1} into account:

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{cx}x_t + W_{ch}h_{t-1}) \tag{4}$$

where $tanh(\cdot)$ is the hyperbolic tangent function, and $\odot$ represents the operation of element-wise product. h_t records the memory transferred to the hidden states and it is defined as:

$$h_t = o_t \odot c_t \tag{5}$$

The obtained h_t is then fed into a softmax layer for the probability distribution over sequential words, which is defined as follows:

$$p_{t+1} = softmax(Ch_t) \tag{6}$$

Here, p_{t+1} denotes the probability distribution over all words in the vocabulary, and C is the weight vector. The image $\boldsymbol{I}$ is input into the LSTM at the time step $t = -1$ to form the visual information. Then the corresponding words are generated one by one to minimize the loss of construction for completing the sentence.

The joint CNN+RNN strategy is commonly-used in literature [8,20,56] and we apply it to generate basic factual descriptions for each image. The basic captions will serve as prior knowledge for the following word-level embeddings of new concepts and be used as the base for generating the proposed emotional descriptions. Details will be given in the Section 4.

Table 1
Examples for the candidate set of affective concepts which are derived from retrieving the textual corpora, i.e., WordNet. We utilize keywords of the eight emotions as the exploring indexes and pick out their Synsets according to definitions and usage examples. As shown, the selected concepts are both diversified and expressive.

Emotions	Affective concepts
Amuse	Happy; happily; blessed; blessing; blissful; bright; laughing; funny; funnily;
Awe	Awesome; impressive; amazing; astonishing; baronial; dazzling; dramatically; thundering;
Content	In content; satisfied; with satisfaction; pleasing; gratifying; fulfilled
Excite	Exciting; infusive; exhilaratingly; incitant; cordial; feverish;
Anger	Aggravated; black; choleric; huffy; hot under the collar; indignant; irate; livid; smoldering
Disgust	Feel sick; unhealthy; awful; awfully; bad; repulsive; offensive
Fear	Afraid; aghast; apprehensive; frightened; horrified; terror-stricken; white-lipped; panicky
Sad	Bittersweet; doleful; heavyhearted; melancholy; pensive; tragic

3.2. Affective concept

In order to model emotion content via image descriptions, one of the key challenges is the expressibility of emotion concepts. While existing paired image data lacks the sentiment expressions in the labeled sentences, we first introduce the set of affective concepts which consist of adjectives, adverbs and phrases derived from major emotions. Generally, a sentence may consist of various words. For example, the sentence ‘There is a man with a simile on his face’ is composed of ten simple words, where only the word ‘smile’ is one of the affective concepts, as it is capable to reflecting emotional information. In this section, we investigate and apply the psychological theory, i.e., the Mikel’s Wheel of Emotions [59], as the guiding principle to build a candidate set of affective concepts. Then, we conduct operations to filter the coarse set and embed each of the concepts into a quantitative feature space for ease of computation.

There are eight basic emotions defined by in the Mikel’s theory. It is similar to the Plutchik’s wheel [12] with richer concepts, treating emotions as discrete categories and then enhancing them in three valences. Meanwhile, the Mikel’s wheel defines pairwise emotion distance as $1+k$ where k denotes “the number of steps required to reach one emotion from another on the Mikelsâ wheel” [59]. Considering the diversity of languages, we query affective concepts with 24 keywords (eight emotions enhanced with three valences, namely three detailed keywords for each emotion) in the Plutchik’s wheel to construct a sufficient set of concepts. The employed keywords for 8 emotions are shown as follows:

- Amuse: ecstasy → joy → serenity;
- Content: admiration → trust → acceptance;
- Awe: vigilance → anticipation → interest
- Excite: amazement → surprise → distraction
- Fear: terror → fear → apprehension;
- Sad: grief → sadness → pensiveness
- Disgust: loathing → disgust → boredom
- Anger: rage→ anger → annoyance

With keywords prepared, we expand a set of affective concepts in the form of adjective, adverb or phrase in the corpora of WordNet [61]. WordNet groups English words into sets of synonyms termed as the Synsets, provides short definitions and usage examples, and records a number of connections among these synonym sets or their members. It is considered as a combination of dictionary and thesaurus, and has been widely applied in the field of automatic text analysis and artificial intelligence. There are several popular linguistics-based corpora which provide alternative concepts with the guidance of emotional keywords. In this paper, we employ the Natural Language Toolkit (NLTK) [62] to create the candidate set of affective concepts via the characteristics in the WordNet, e.g., The Synsets.

3.2.1. Concepts analysis

The keywords query process provide a coarse collection of affective concepts. To give an applicable candidate set of concepts for the generation of emotional image descriptions, it is necessary to filter the original set and remove noises. We first remove the obvious stop words and perform the stemming operation for a basic filtering. Then, the concepts with a frequency of lower than 5 are discarded for the sake of comprehensible sentences, since users do not want to describe an image with tricky words. In addition, we invite ten experienced students to check the remaining concepts. Each student is aware of the emotional category for each of the candidate concept. They are asked to pick out all the confusing concepts (which often have relations to other emotions). Concepts with more than five negative marks (half of the annotators) are also eliminated from the set. Finally, we generate over 30 affective concepts for each of the 8 emotions in average, of which several examples can be found in Table 1. As shown, the first column includes eight major emotions considered in this paper. We present affective concepts with similar or closed definitions with the key emotions. Note annotators are not limited to utilize phrases included in the set of affective concepts. Descriptive words are allowed considering the diversity and smoothness of natural sentences.

The top-left and bottom plot in Fig. 2 presents the Mikel’s wheel and the chain form of the pie in this paper. The distance of emotions, termed as $dist(E_1, E_2)$, refers to the distance between the emotions E_1 and E_2, which is computed using $1+k$ as illustrated previously [63].

3.2.2. Continuous ranking

According to the Mikel’s wheel, we extract eight main emotions and induce a richer set of affective concepts than existing works [39]. In the framework of emotional descriptions generation, we aim to estimate the affective concepts for each image and embed the emotional words into sentences. However, it is difficult to distinguish the concepts of intra-category since they are semantically close and the precise selection relies on the effective prior knowledge in the linguistic community. Therefore, inspired by the pairwise emotion distance defined in the Mikel’s wheel [59], we compute the continuous emotional projection of affective concepts considering the polarity property among emotions. Specifically, since the eight basic emotions can be grouped into two polarities, i.e., positive and negative emotion, we calculate the pairwise distance for each concept pair by ignoring the influence caused by the gap between the positive and negative polarities. For example, the distance between ‘Amuse’ and ‘Awe’ should be smaller than between ‘Amuse’ and ‘Anger’, while the traditional wheel indicates the opposite result. In that case, we cut the wheel at the junction of ‘Anger’ and ‘Amuse’ and stretch it to a chain as shown in the bottom plot of Fig. 2. By modifying the computation

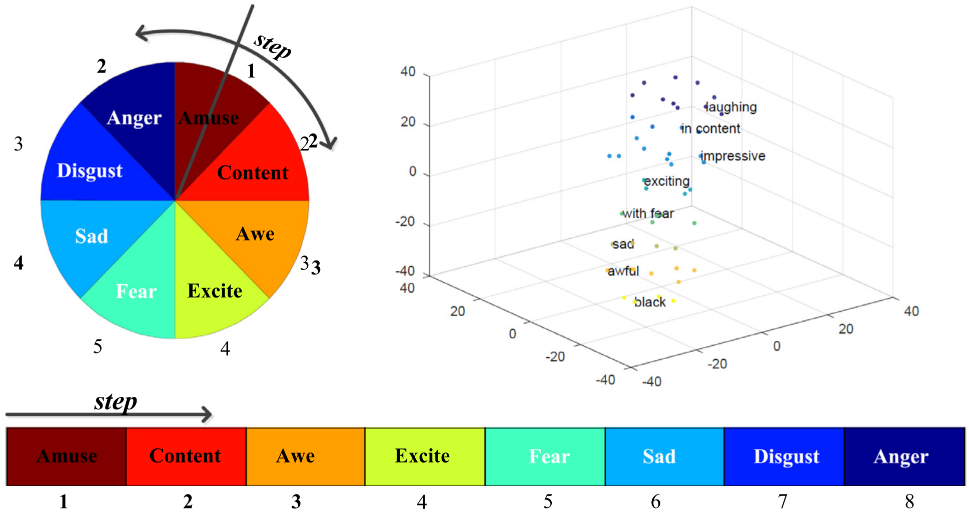


Fig. 2. The Mikel's emotion wheel [59] and the visualization of the embedding of several emotional concepts. For the Mikel's emotion wheel, emotions with black marks are positive, otherwise negative. As shown in the pie on left-top, the numbers indicate Mikel's emotion distance from 'Amuse' to other emotions, which ignore the gaps between the positive and negative. In our paper, we decompose the pie into the form of chain as shown in the bottom of the figure to make the 'Amuse' and 'Fear' disjoint. By utilizing the emotion distance, we calculate the continuous distance among affective concepts via feature embedding. The right-top plot visualizes several concepts in dimensional space using t-SNE [60]. As shown in the figure, the affective concept can be distinguished in the embedding space, especially for the concepts with different sentiment polarity.

process, the pairwise emotion distance between 'Amuse' and 'anger' is 7 instead of 1, which caters to the human cognition.

Furthermore, for affective concepts of intra-category, we propose to rank them with continuous probabilities. The candidate set of affective concepts are retrieved with emotions enhanced in continuous valence. Note that the relationship between keywords and concepts is of one-to-many fashion, so the affective concepts cannot be put into a continuous space directly. We solve this problem by employing the word embedding. Specifically, we train the Glove [64] with large text corpora (details will be given in Section 4), which generates the word embedding representations for the following operations. In addition, we append a bias term to the embedding. Hence, the affective concepts *AF* are represented as: $AF = [f_e; b]$, where f_e denotes the features from word embedding and $b \in \mathbb{R}^8$ denotes the emotional category bias derived from the improved Mikel's emotion wheel as shown in the bottom plot of Fig. 2. Each entry in b indicates the distance between the category of affective concepts and other emotion categories. For example, the bias item of an affective concept belongs to Awe is [3, 2, 1, 2, 3, 4, 5] where the distance is computed based on the sequence given by the Mikel's emotion wheel [59]. In Fig. 2, we show several example concepts in the projection space via the t-SNE tool [60] for visualization. As shown in the figure, the distributions of affective concepts in the learned projection space can properly match the cognitive distance among emotions. The phenomenon demonstrates the effectiveness of the concatenation of word embeddings and emotional bias for representing the affective concepts, which will be used in the proposed framework of emotional caption generation.

4. Proposed approach

We illustrate the framework of the proposed affective image captioning method in Fig. 3, which can compose proper affective concepts into sentences via learning a sentiment recognition model. Due to the lack of well-labeled image-caption paired dataset, in this paper, we extract emotional expressions with unpaired visual and textual repositories by learning with both content and linguistic modules. The introduction of the linguistic module is to guarantee the compatibility between visual and textual features. We utilize detectors developed from affective image classification tasks to extract visual sentimental representations, followed by learning the word vector representations for emotion expressions. For the sake of affective descriptions, we encode the cross-modal features to capture the correspondences between visual and textual channel. Sequentially, we are able to understand the emotional information and transfer it to text descriptions. Details will be given in the following sections.

4.1. Content module

For the sake of emotional description generation, we first extract the sentiment representations for each image using the convolutional neural network (CNN). Then we employ the off-the-shelf recurrent neural network (RNN) on the learned latent space to generate affective sentences. In this subsection, we exploit the emotion category for each image.

The commonly-used CNN model is pre-trained on the large-scale visual recognition challenge, i.e., ImageNet [65], which provides cognitive knowledge from 1000 categories. To analyze and understand image emotions in measurable representations, we carry the fine-tuning of a CNN model on the large-scale emotion dataset to process affective knowledge. We employ the large-scale emotion dataset established by You et al. [54], termed as the FI dataset, which supplies sufficient data to prompt the learning of novel information on other communities. As shown in Table 2, the FI dataset consists of 90,000 images with weak labels (each image is either positive or negative), among which over 23,000 images are well labeled with eight emotions. According to Morina et al. [54], the construction of FI follows the Mikel's emotion definition system by querying the image search engines from the Internet using the eight emotion as keywords. Similarly, we collect a candidate list of affective concepts from WordNet using the queries from Mikel's emotion definition system. We then conduct the filtering and continuous ranking process on the list as illustrated in Section 3.2.2. The basic CNN model is the VGGNet [66], which is pre-trained on the training split of ILSVRC-2012 dataset [65]. We implement the procedure of fine-tuning on the well-labeled subset of the FI datasets to induce the attractiveness of emotional expressibility. The features of the fully connected layer is employed

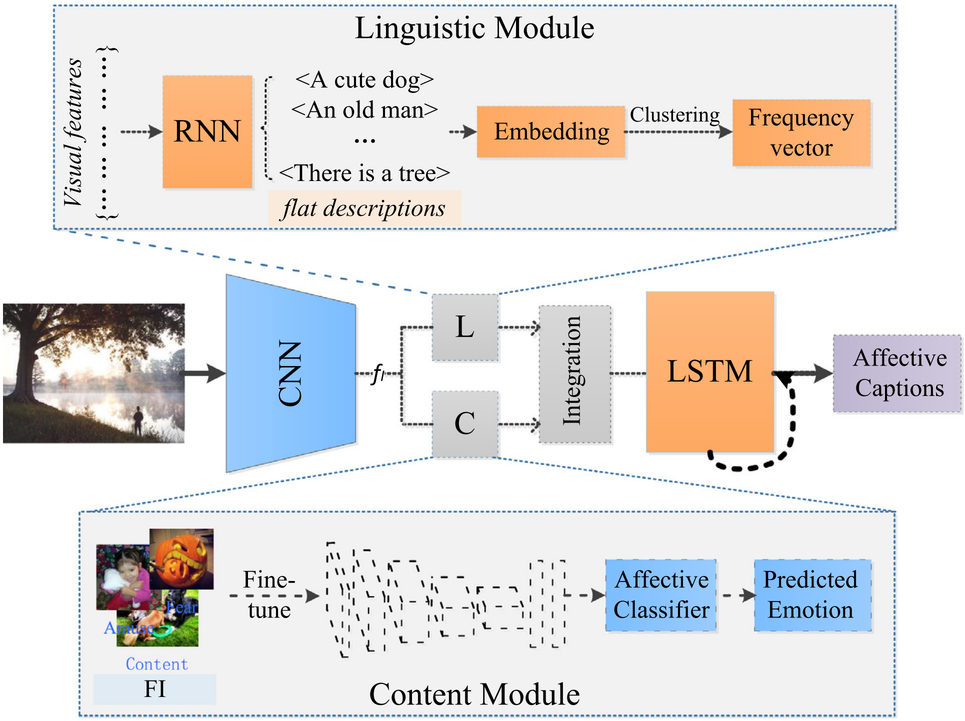


Fig. 3. The framework (**Middle**) of the proposed affective image captioning algorithm which mainly consists of two parts, *i.e.*, content module (**Bottom**) and linguistic module (**Top**). The content module maps images to deep representations to predict the emotions via knowledge transfer, which is trained on a large benchmark of affective images. The linguistic module takes the deep representations as input, followed by embedding the learned words into latent representation space. It then captures the nearest concepts by using the off-the-shelf clustering methods and get the frequency vector as output. The weights learning component incorporates the analysis from both visual and textual representations to generate the affective sentences.

Table 2
Statistics of the large-scale FI dataset. The ‘Weakly’ label denotes that each image is annotated by either positive or negative emotion. The ‘Well’ label indicates that the image is annotated by eight kinds of emotions according to the Mikel’s wheel. As shown, each of the eight emotion categories consists of over 1,000 images, which is sufficient for learning the distribution of emotion.

Label	Amuse	Content	Awe	Excite	Fear	Sad	Disgust	Anger	All
Weakly	11,000	11,000	11,000	11,000	13,000	110,00	11,000	11,000	90,000
Well	4942	5374	3151	2963	1032	2922	1658	1266	23,308

for the emotion representations of each image, named as f_I. As the lack of affective concepts in traditional tasks, it requires to integrate the collected affective concepts into original factual descriptions. The render of the linguistic module guarantees the compatibility between visual sentimental representations and affective concepts.

4.2. Linguistic module

In the module of content analysis, we recognize the emotional category with visual representations. Since we incorporate the most appropriate concepts to the descriptions for images, it is required to refine affective concepts rather than simple classifications on eight emotions. Though we predict emotions from the visual data, the generation of emotional captions is limited since the absence of affective concepts in traditional descriptions. The inter-modality gap requires further integration to reconcile the visual sentimental representations with affective concepts of textual formation. Considering that the content module is limited to the benchmark of eight emotion categories, the large volume of text corpora and the projection to vector space representations allows the algorithm to enhance the expression effectively and strengthen the compatibility between affective content and textual descriptions.

For the computation of the linguistic module, we leverage the word embeddings to calculate the textual features. First, we extract the deep representation of images from the basic CNN model and input it into the RNN to obtain the flat sentences. To discover the relationship between inter-modal features, [57] adopt the analogue strategy to regard representations from different modality as bag and its instance. Motivated by the analogue strategy, we analyze the original descriptions and assign affective concepts to each sentence by clustering the word embeddings following the bag-of-visual words algorithm. In other word, we generate a vector of affective concepts for each image in the linguistic module based on the original sentences. We denote representations of affective concepts as $A = \{A_1, A_2, \ldots, A_m\}$, where m denotes the number of affective concepts, A_j is a 50-dimensional embedding which is illustrated in Section 3.2. The word vector representations are extracted by GloVe [64], which is trained on a large-scale corpus, *i.e.*, Text8. We calculate the Euclidean distance from each word to affective concepts so that each word w is assigned to the nearest concept. For each image, we statistic the frequency of affective concepts in a m-dimensional vector $a = \{a_1, a_2, \ldots, a_m\}$. With the m-dimension frequency vector prepared, serving as the textual features output into the RNN to generate a sequence of words for description.

4.3. Incorporating the visual and textual components

In order to learn the distinctive factor for the generation of emotional description, we introduce a fusion module which serves as a major block for preparing the context vector combining either visual and affective information. As the framework which automatically generates emotional descriptions concerns the textual expression of visual emotions, it is necessary to cope with intrinsic relations in various modalities [58,67]. We add the linguistic module to exploit the connections between visual emotions and textual descriptions. From the content analysis, we obtain *fc*7 features f_I as deep visual representations for the input of LSTM. Besides, we let the *fc*8 feature vector $E = \{e_1, e_2, \ldots, e_8\}$ represent the emotion distribution [68]. Each entry in the 8-dimensional vector denotes one of the eight emotions separately. For each entry in e, it reflects the probability whether the referring emotion is the right category for the generation of descriptions, or the importance to give to the emotion in simultaneous consideration. On the other hand, we explore the frequency of affective concepts a to introduce finer semantic descriptions. Once the weights of emotions and probability of affective concepts are prepared, the context vector x_t is computed by:

$$x_t = \varphi(x_t, a, e), t \in \{0, 1, \ldots, N-1\} \tag{7}$$

where φ is a function that returns an aligned vector for the set of word sequence information and visual content, and N denotes the number of iterations in LSTM. The goal of this function is to integrate emotion information into the context vectors x_t. Most of existing works in machine learning and computer vision project various representations from the original space to another [69]. The data of both modalities are usually converted from the original to another space for integration [67,69]. In this work, we simplify the integration by concatenating visual and textual features and extracting the principal vector whose length is aligned to the visual representations. In detail, we execute the φ based on the Principal Component Analysis (PCA) [70] for capturing major information. We assign the 512-dimension of x_t according to facilitate the training process and preserve the efficient components as well as representation in the limited dimension. As a result, we are able to consider the extra affective information provided by the emotion distribution e and frequency vectors of affective concepts a in the unit of LSTM to enhance the emotional influence. In the formulation of emotion, we further introduce the judgment on whether an emotion exists or not. Given the sentiment probability produced by the fine-tuned CNN model, we introduce a threshold λ to control the addition of emotions. While the maximum probability of the emotion probability $max(e) > \lambda$, we perform the fusion of emotions and context. Otherwise, the fusion is interrupted. In summary, it is formulated as follows:

$$\begin{aligned} x_t &= \varphi(x_t, C(a, e)) \\ C &= \begin{cases} 1, max(e) \geq \lambda \\ 0, max(e) < \lambda \end{cases} \end{aligned} \tag{8}$$

5. Experimental results

In this section, we conduct experiments to evaluate the quality of the affective captions derived from the proposed method.

5.1. Datasets

To evaluate the proposed method in the generation of emotional descriptions, we conduct experiments on the subset of Microsoft COCO dataset (MS-COCO, [2]), namely the SentiCap, which is annotated with adjective noun pairs (ANPs) in [39]. Note for each image, the SentiCap dataset provides six sentences with different sentiment polarities, *i.e.*, three for positive and the rest for negative. The ANPs are utilized in the description to express the sentiments in each image. In our work, we propose to predict the intrinsic emotion contained in image contents but not interchangeable sentiments with different annotators in mutual works. We reannotate the SentiCap as well, resulting in 1593 positive image and 632 negative images according to the annotations. Annotators are asked to vote for the emotion category an image belongs to and then describe it with proper concepts.

5.2. Evaluation metrics

In this paper, we propose to evaluate our EmoCap method with both automatic metrics and crowd-sourced judgments following the same fashion in [3]. Note evaluating the performance of the subjective generation remains to be a challenging problem in literature. Most existing works report the performance under common evaluations, e.g., BLEU_1 or METEOR, against emotional captions which are re-annotated with affective concepts (as described in Section 5.1. For the aspect of crowd-sourcing metrics, we ask participants to give a subjective score on the two properties, i.e., distinctiveness and attractiveness, of each description for a given image.

5.2.1. Automatic metrics

Following the same routine of the previous works, we employ comprehensive automatic evaluations, i.e., BLEU, ROUGE, METEOR and CIDEr metrics, from the Microsoft COCO evaluation software [71]. For all of the four metrics, higher score values indicate better performance. BLEU evaluates a given transcription using the perplexity of the model which calculates the geometric structure of the inverse probability for each predicted word. We also report results on METEOR, ROUGE and CIDEr as well for comprehensive evaluations.

5.2.2. Crowd-sourcing metrics

We conduct a series of human evaluation experiments following the criteria proposed in [72]. We employ five annotators for each annotations of the image and report the average score for comparison. For each caption given its corresponding image, it is required to be rated with a 4-point scale: Excellent, Good, Bad, and Embarrassing. In the evaluation, we inform the raters that 'Excellent' means the caption contains all of the exact emotional descriptions for details presented in the picture; 'Good' refers that the generated caption contains some biased descriptions like the confusions of amuse and awe, which turns to be not bad in acceptance; 'Bad' means the caption may be misleading such as the improper use of affective concepts; 'Embarrassing' means that the caption is totally wrong, or may upset the owner or subject of the image. In addition, the employers are asked to rate the descriptions in two aspects: descriptiveness and attractive. The descriptiveness inclines to the objective quality of sentences while the attractiveness is more incline to subjectivity.

5.3. Implementation details

We implement the EmoCap using Torch. The affection classifier utilized in the module of content analysis is trained separately. Besides the CNN model which is fine-tuned from the large-scale affection benchmark FI, we train the affection classifier using the famous International Affective Picture System (IAPS) [73], which is the same emotion definition system with FI. The frequency of affection concepts is built on top of the bag-of-visual strategy. The experiments are carried out with a TITAN X GPU. We implement our training processing against the pre-trained model from NIC [8] given the limitation the Senticap dataset, which is used to generate flat descriptions as well. We adopt a fixed learning rate of

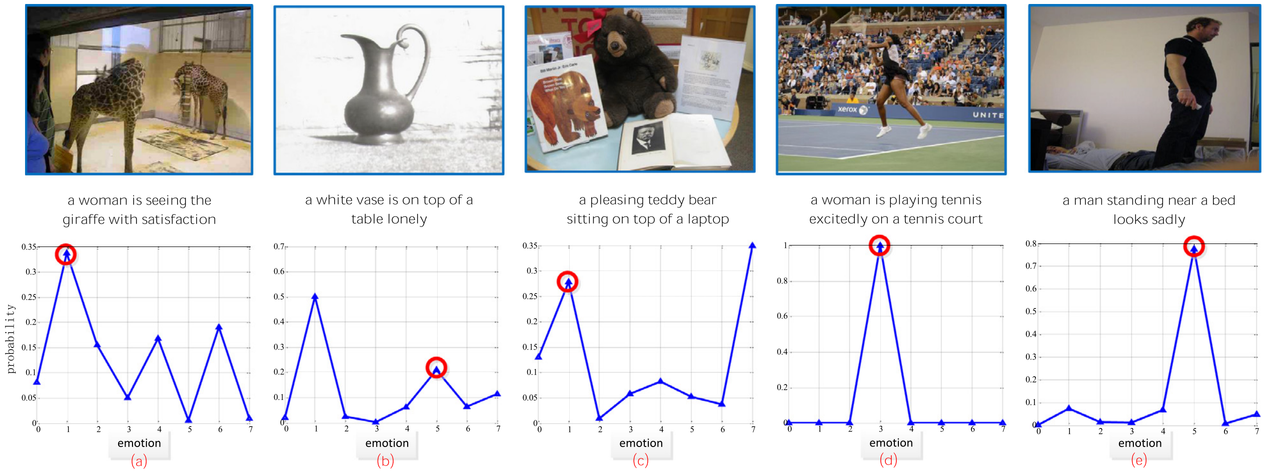


Fig. 4. Visualization of the predicted sentiment distribution (The curves), together with the generated caption (The middle text) with the preference of a specific emotion category (The red circle). Each point in the curve indicates a specific emotion, i.e., amuse, content, awe, excite, fear, sadness, disgust and anger for the indexes 1–8, respectively. Rather than the single peak in (d) and (e), the images in (a), (b) and (c) have multiple peaks which indicate a complicated sentiment contained in the image. We show an example caption for each of them. It is clear that the caption can express the proper emotion adaptively as is given by the users.

Table 3
Overall comparisons among the captions generated with sentiment derived by four methods. The 'EmoCap-T' and 'EmoCap-I' denote the baselines which only consider the text and image information, respectively. We employ metrics including BLEU, METEOR, ROUGE and CIDEr for evaluation. We compare our model with the SentiCap as related concepts (ANPs) are accounted in the algorithm. $\lambda = 0$ refers to experiments without the judgments on neural or emotional images. $\lambda = 0.6$ is the threshold for judging emotions.

λ	Methods	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE	CIDEr
0	SentiCap	47.2	23.7	13.4	12.4	11.0	26.3	62.0
	EmoCap-T	50.1	24.8	21.8	12.7	14.9	32.0	45.1
	EmoCap-I	53.5	28.0	24.4	16.7	15.0	33.6	51.7
	EmoCap	51.9	31.8	**28.9**	19.1	16.0	33.0	55.1
0.6	SentiCap	–	–	–	–	–	–	–
	EmoCap-T	55.5	32.1	28.4	12.8	15.8	34.0	55.6
	EmoCap-I	57.2	33.6	29.7	18.4	15.7	34.8	57.8
	EmoCap	**60.7**	**39.3**	25.0	**20.1**	**16.7**	**35.9**	**62.4**

Table 4
Comparison of evaluations for images with specific emotions. We first split the dataset into subsets corresponding to their predicted emotion category. Then we evaluate the emotional descriptions of each subset independently. 'P' in the brackets means positive while 'N' refers to the negative. As shown, the images with emotions 'Excite' and 'Disgust' present relatively better performances in each polarity, respectively.

Emotions	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE	CIDEr
Amuse(P)	41.7	27.4	20.6	16.7	16.1	30.7	53.2
Content(P)	47.4	32.0	18.8	13.5	15.7	32.9	54.9
Awe(P)	49.1	33.4	24.7	19.4	16.0	33.9	58.7
Excite(P)	49.1	34.2	21.4	21.2	16.1	33.9	60.1
Fear(N)	47.6	32.0	23.9	18.3	16.0	31.6	58.6
Sadness(N)	47.8	32.6	24.5	19.4	16.0	33.4	61.7
Disgust(N)	48.9	33.6	25.7	20.8	16.3	34.1	60.9
Anger(N)	47.9	31.2	24.0	20.1	16.3	33.4	65.9

1e–3 and use stochastic gradient descent with momentum 0.9. We use the perplexity as the stopping criteria, which is the geometric mean of the inverse probability for each predicted word. The learning procedure takes 68 minutes at around 133 image-sentence pairs per second. To evaluate the efficacy of emotion analysis, we first compare the traditional captions with the emotional captions in three steps. We extract emotional sentence in three conditions. Two simple baselines are text-based EmoCap and image-based EmoCap, respectively. In the text-based EmoCap (EmoCap-Text), we skip the module of content analysis and just extract the frequency vector of affective concepts. Due to the considerations on an image without emotions, the emotional vector $\boldsymbol{e}$ in EmoCap-Text is set up by an average distribution of probabilities rather than the zero vector. In the image-based EmoCap (EmoCap-Image), we implement the model without computations from the module of linguistic analysis. The frequency vector of affective concepts is set up similarly as in EmoCap-Text. We denote the overall system as EmoCap which jointly integrates the word-level sequence and emotion probability. (Fig. 4)

On top of that, we further develop the examination of emotions and compare the quality of descriptions related to different emotions. We train the affection classifier in one-vs-all strategy and evaluate the sentence under corresponding emotion class. Comparisons among various emotion categories are shown in Table 4. To do so, we analyze whether the descriptions are different and how they influence the results. Based on that, we propose to judge

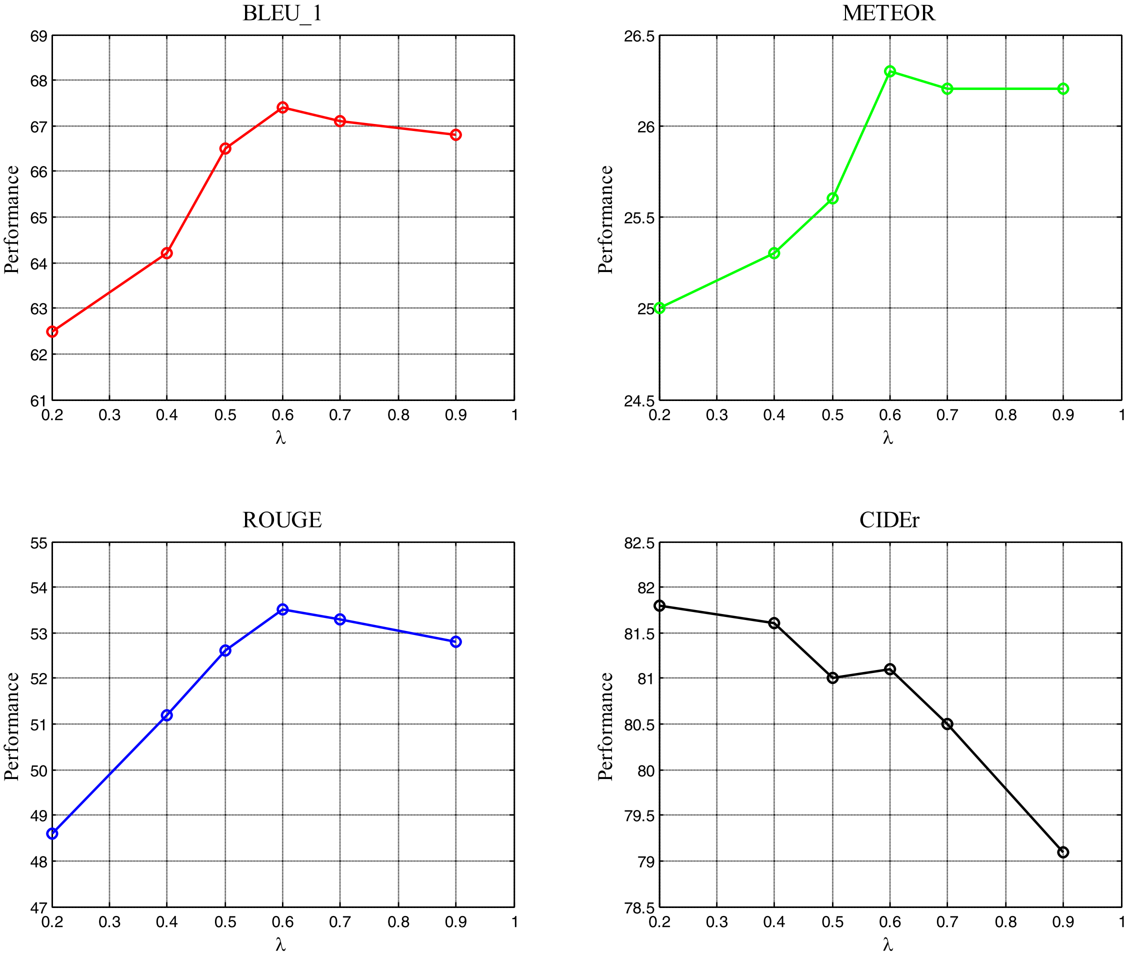


Fig. 5. Evaluation on parameter λ against the overall performance with the four kinds of metrics, i.e., BLEU_1, METEOR, ROUGE and CIDEr. We test the set of λ values: [0.2, 0.4, 0.5, 0.6, 0.7, 0.9] for each metric. Under the valuation of BLEU_1, METEOR, EmoCap achieves the best performance while we have $\lambda = 0.6$. Though the overall performances on the CIDEr tend to decrease with the increasing of λ, the performance of $\lambda = 0.6$ is comparable to the best. Therefore, in this paper we set $\lambda = 0.6$.

images with emotions or not since some images turn out to be neutral with no strong emotions. For example, a picture showing an empty office brings stimuli insignificantly into human states. It would be somewhat comic to describe such images with affective tags compulsively. A distinction on whether the image is sentimental or not becomes necessary.

5.4. Results analysis

5.4.1. Comparisons of direct and emotional analysis

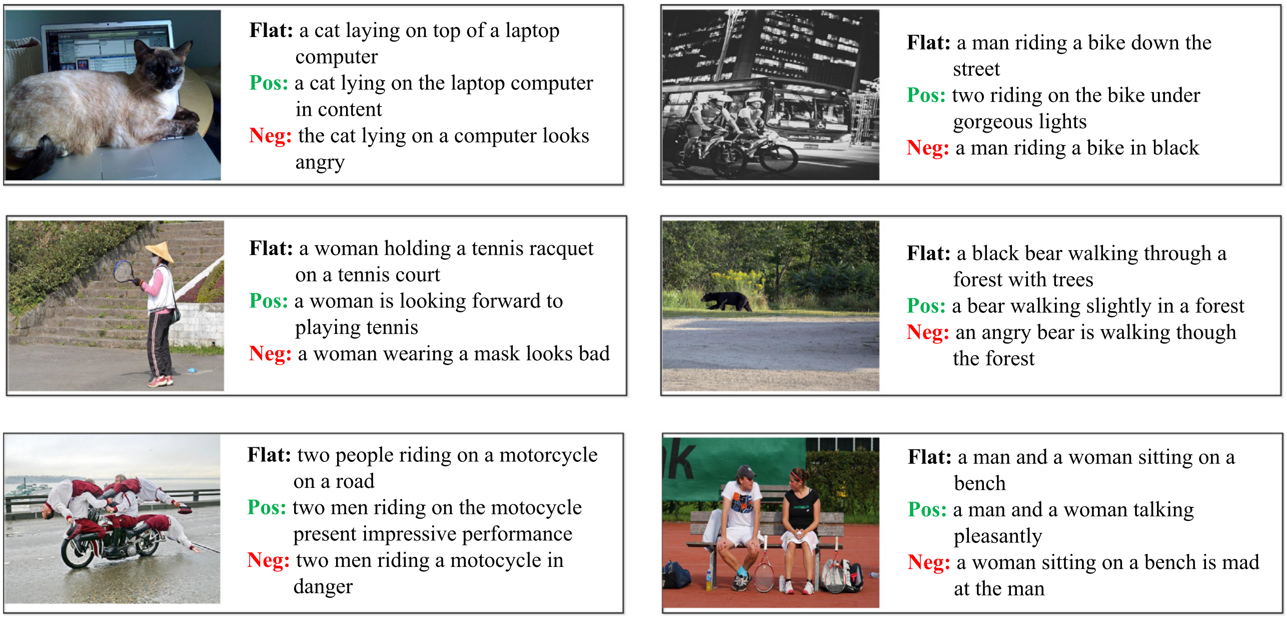
Table 3 reports automatic evaluation metrics for descriptions produces by SentiCap [39] and Our methods (EmoCap). Under the condition of $\lambda = 0$, these models take no account of judgment for neutral images. Both the SentiCap and EmoCap contains sentimental words in their descriptions. While EmoCap captures emotional expression with emotion analysis, the SentiCap is trained on annotations with both positive and negative annotations. Our methods show better performances in the automatic evaluations. We assume that the usage of ANPs brings sentiments into image captioning but ignores the analysis of intrinsic emotions. There is hardly negative sentiments exist in a positive image while the SentiCap provides each image with annotations in either sentiment. As for the comparisons between text-based EmoCap and image-based EmoCap, we find that the image-based EmoCap shows better results than text-based EmoCap in average. Given that the goal of emotional descriptions is to enable emotional expressibility for the picture, the analysis based on image content results in more accurate emotion predictions. From our experimental results, the EmoCap shows the best performances with most metrics. The linguistic analysis turns out to be not sufficient but supplementary to the final decision of emotions. Conclusively, the combination of both visual and textual content (EmoCap) results in the best performances, which indicates the multimodal fusion captures more adequate information than considerations on the single aspect. The addition of neutral image judgement ($\lambda = 0.6$) will be discussed in Section 5.4.3

5.4.2. Emotional distinctions

In most of the affective classification tasks, confusions exist in various emotions and lead to higher error rate in prediction [74]. In this paper, we propose to investigate how images with different emotions (predicted) can be influenced. To conduct the comparison, we replace the training strategy of an affective classifier with the one-vs-all method. The tested emotion is treated as true while others are assigned with false. For each examined emotion category, we maintain images predicted in the category and evaluate their performances respectively. Table 4 summarizes the quality of the descriptions for the eight categories. The results indicate that images in the category of excite and disgust show better performances relatively. Excite and Disgust are emotions with high arousal, which can stimulate human states easily. Their characteristics make it sensible in human recognition and descriptions. Overall, differentiations exist in emotions and lead to various

Table 5
Crowd-sourcing evaluation results for the distinctiveness and attractiveness of generated image descriptions. The annotation rule can be found in Section 5.2.2. As shown, the captions derived from the proposed EmoCap method achieve the best human preferences. The performance (%) in the attractiveness is about five points higher than the SentiCap, which indicates the description with emotions improves the understanding of images and attracts human attention.

	Rates	NIC	SC(1)	SC(2)	SC(3)	SC(4)	SC(5)	SC(6)	EmoCap
Distinctiveness	Excel	22.3	30.2	29.8	30.1	30.3	29.9	31.0	32.5
	Good	24.8	28.9	30.8	29.4	27.3	28.5	29.7	31.3
	Bad	30.2	25.7	26.3	24.3	26.0	25.5	26.9	24.8
	Emb	22.7	15.2	13.1	16.2	16.4	16.1	12.4	11.4
Attractiveness	Excel	16.7	23.1	24.5	23.8	25.4	23.9	24.1	29.4
	Good	20.1	25.2	24.9	25.6	25.0	24.7	26.0	30.6
	Bad	34.2	31.4	30.8	29.9	31.0	30.5	31.2	25.3
	Emb	29.0	20.3	19.8	20.7	18.6	20.9	18.7	14.7



Flat: a cat laying on top of a laptop computer
Pos: a cat lying on the laptop computer in content
Neg: the cat lying on a computer looks angry

Flat: a man riding a bike down the street
Pos: two riding on the bike under gorgeous lights
Neg: a man riding a bike in black

Flat: a woman holding a tennis racquet on a tennis court
Pos: a woman is looking forward to playing tennis
Neg: a woman wearing a mask looks bad

Flat: a black bear walking through a forest with trees
Pos: a bear walking slightly in a forest
Neg: an angry bear is walking though the forest

Flat: two people riding on a motorcycle on a road
Pos: two men riding on the motocycle present impressive performance
Neg: two men riding a motocycle in danger

Flat: a man and a woman sitting on a bench
Pos: a man and a woman talking pleasantly
Neg: a woman sitting on a bench is mad at the man

Fig. 6. Comparisons of three kinds of captions on insipid images. For each example, 'Flat' means the traditional captions derived from content-based algorithms. 'Pos' or 'Neg' indicates that the caption generators prefer a positive or negative sentence, respectively. We change the emotion probability vector which is fed into the LSTM to produce image descriptions with desired sentimental polarities. As shown, we can generate reasonable captions with different polarities for the flat images, which is more attractive and distinctive than the original description. In addition, the alternative caption generator is interpretable and controllable with subjective opinions given by users.

performances in captioning. More precision analysis is needed for the development of emotional descriptions.

5.4.3. Emotion judgement

From the evaluation of emotional descriptions, we further analyze the judgment of neutral images (refer to images with no emotions). While we annotate images with affective concepts compulsively, it would be improper to describe neutral images with emotional sentences. Sequentially, we introduce the parameter λ to control the addition of emotional expressions as given in (8). For the prediction of emotional probabilities ranged from 0 to 1, where $\sum e_i = 1$, we recognize the image whose highest probability greater than λ as images with emotions. When the highest emotion probability is greater than λ, the representation $\{e, a\}$ will be integrated in the context vector. $\lambda = 0$ (in Table 3) means every instance will be assigned with an emotion and it will generate emotional descriptions. In this section, we set the λ with a set of values to test to which degree the image can be recognized as neutral. The automatic evaluation metrics are adopted as well. The set of values for the parameter λ is [0.2, 0.4, 0.5, 0.6, 0.7, 0.9] and the corresponding performance are shown in Fig. 5. Despite the trivial decreasing results under the CIDEr metrics, EmoCap achieves the best performance when $\lambda = 0.6$. The automatic evaluations when $\lambda = 0.6$ are reported in Table 3 as well. These performances demonstrate that the judgement of emotional or neutral images help to provide better results in the captioning tasks.

5.4.4. Human preferences

Human recognition is the clearest evaluation for the quality of descriptions [56]. We conduct the human voting scheme following [72]. Details of the human-sourcing evaluations have been described in the Section 5.2.2. Participants are asked to rate the descriptiveness and attractiveness separately. Apart from the flat descriptions generated by NIC, we also compare the emotional descriptions with the annotations from SentiCap [39]. As each image is annotated with six sentences in SentiCap, we group them into six groups for a clear comparison. Table 5 shows human preferences on diverse descriptions. The flat captions generated by NIC show somewhat comparable results in descriptiveness but less-than-ideal performances in attractiveness due to the limitation on emotional expressions. The EmoCap takes the highest proportion on rate 'Excellent' and 'Good' in human preferences. The performance in the attractiveness is five points higher then the SentiCap, indicating the description with emotions improves the understanding of images and attracts human attention.

5.5. Application

In the experiments of EmoCap, we analyze emotional descriptions for various emotion category with judgments on image emotions. Inspired by the polarized annotation from the SentiCap, we wonder if it is possible to generate sentences for neutral images with totally different emotions. Though we argue that such annotations are fuzzy for images with obvious images, it would be funny and useful if we can translate a neutral image with positive/negative emotions. We propose to transfer the descriptions of neutral images with the adjustment on λ. We extract images whose highest probability in the emotion vector is below 0.6. We select these images and generate descriptions with positive/negative emotions. For example, we attempt to describe a neutral image with amuse concepts We set the corresponding dimension of amuse as 1 and then send normalization of the vector into RNN and generate novel descriptions. Some examples are presented in Fig. 6. For each of the neutral images, we adjust their emotional probability vectors and generate sentences with different sentiments. While the EmoCap tries to annotate emotions conveyed by the image, we further develop the interchangeable descriptions of neutral images.

6. Conclusion

In this paper, we aim to generate image descriptions with the recognition of different emotions. To achieve this goal, we develop a framework of emotional captioning via transferring knowledge from the large-scale sentiment recognition benchmark. To quantify the results of emotional descriptions, we adopt both automatic and human-sourcing metric for evaluation. Based on the generation results of EmoCap, we judge the necessity of emotional descriptions. We extract a proper threshold for the examinations of emotions and then apply the proposed method to generate positive/negative descriptions for neutral images. Experiments indicate that the EmoCap can generate sentences with desired emotions for neutral images, which provides changeable understanding for a given image.

Acknowledgement

This work was supported by NSFC (NO. 61876094), Natural Science Foundation of Tianjin, China (NO. 18JCYBJC15400), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neucom.2018.03.078

References

[1] X. Chen, C. Lawrence Zitnick, Mind's eye: a recurrent visual representation for image caption generation, in: Proceedings of the CVPR, 2015.

[2] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, C.L. Zitnick, Microsoft coco: common objects in context, in: Proceedings of the ECCV, 2014.

[3] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, J. Artif. Intell. Res. 47 (2013) 853–899.

[4] J. Machajdik, A. Hanbury, Affective image classification using features inspired by psychology and art theory, in: Proceedings of the ACM MM, 2010.

[5] J. Liu, Q. Miao, Y. Sun, J. Song, Y. Quan, Fast structural ensemble for one-class classification, Pattern Recognit. Lett. 80 (2016) 179–187.

[6] Q. Miao, Y. Cao, G. Xia, M. Gong, J. Liu, J. Song, Rboost: label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners, IEEE Trans. Neural Netw. Learn. Syst. 27 (11) (2016) 2216–2228.

[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[8] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proceedings of the CVPR, 2015.

[9] P. Kuznetsova, V. Ordonez, T. Berg, Y. Choi, Treetalk: composition and compression of trees for image descriptions, Trans. Assoc. Comput. Linguist. 2 (1) (2014) 351–362.

[10] L. Zheng, S. Wang, J. Wang, Q. Tian, Accurate image search with multi-scale contextual evidences, Int. J. Comput. Vis. 120 (1) (2016) 1–13.

[11] 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.

[12] L. Camras, Emotion: A Psychoevolutionary Synthesis by Robert Plutchik, Harper & Row, 1980.

[13] Q. Miao, T. Liu, J. Song, M. Gong, Y. Yang, Guided superpixel method for topographic map processing, IEEE Trans. Geosci. Remote Sens. 54 (11) (2016) 6265–6279.

[14] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: generating sentences from images, in: Proceedings of the ECCV, 2010.

[15] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Babytalk: understanding and generating simple image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2891–2903.

[16] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele, Translating video content to natural language descriptions, in: Proceedings of the ICCV, 2013.

[17] I. Nwogu, Y. Zhou, C. Brown, Disco: describing images using scene contexts and objects, in: Proceedings of the AAAI, 2011.

[18] B.Z. Yao, X. Yang, L. Lin, M.W. Lee, S.-C. Zhu, I2t: Image parsing to text description, Proc. IEEE 98 (8) (2010) 1485–1508.

[19] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, Y. Choi, Collective generation of natural image descriptions, in: Proceedings of the ACL, 2012.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the ICCV, 2015.

[21] T. Mikolov, A. Deoras, D. Povey, L. Burget, Strategies for training large scale neural network language models, in: Proceedings of the ASRU, 2012.

[22] A. Karpathy, F.F. Li, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the CVPR, 2015.

[23] L. Zheng, Y. Yang, Q. Tian, Sift meets CNN: a decade survey of instance retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 14 (8) (2016) 1–18.

[24] C. Lynch, K. Aryafar, J. Attenberg, Unifying visual-semantic embeddings with multimodal neural language models, in: Proceedings of the TACL, 2015.

[25] J. Mao, X. Wei, Y. Yang, J. Wang, Learning like a child: fast novel visual concept learning from sentence descriptions of images, in: Proceedings of the ICCV, 2015a.

[26] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks (m-rnn), in: Proceedings of the ICLR, 2015b.

[27] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, K. Saenko, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the CVPR, 2015.

[28] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, T. Darrell, Natural language object retrieval, in: Proceedings of the CVPR, 2016.

[29] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the CVPR, 2016.

[30] S. Kazemzadeh, V. Ordonez, M. Matten, T. Berg, Referitgame: referring to objects in photographs of natural scenes, in: Proceedings of the EMNLP, 2014.

[31] N. FitzGerald, Y. Artzi, L.S. Zettlemoyer, Learning distributions over logical forms for referring expression generation, in: Proceedings of the EMNLP, 2013.

[32] K.V. Deemter, I.V.D. Sluis, A. Gatt, Building a semantically transparent corpus for the generation of referring expressions, in: Proceedings of the INLG, 2006.

[33] J. Johnson, A. Karpathy, F.F. Li, Densecap: fully convolutional localization networks for dense captioning, in: Proceedings of the CVPR, 2016.

[34] J. Viethen, R. Dale, The use of spatial relations in referring expression generation, in: Proceedings of the INLG, 2010.

[35] M. Mitchell, K.V. Deemter, E. Reiter, Natural reference to objects in a visual domain, in: Proceedings of the INLG, 2010.

[36] D. Golland, P. Liang, K. Dan, A game-theoretic approach to generating spatial descriptions, in: Proceedings of the EMNLP, 2010.

[37] Q. Miao, P. Xu, X. Li, J. Song, W. Li, Y. Yang, The recognition of the point symbols in the scanned topographic maps, IEEE Trans. Image Process. 26 (6) (2017) 2751–2766.

[38] L.A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, T. Darrell, Deep compositional captioning: describing novel object categories without paired training data, in: Proceedings of the CVPR, 2016.

[39] A. Mathews, L. Xie, X. He, Senticap: generating image descriptions with sentiments, in: Proceedings of the AAAI, 2016.

[40] L. Zheng, S. Wang, Q. Tian, Coupled binary embedding for large-scale image retrieval, IEEE Trans. Image Process. 23 (8) (2014) 3368–3380.

[41] Q.G. Miao, C. Shi, P.F. Xu, M. Yang, Y.B. Shi, A novel algorithm of image fusion using shearlets, Opt. Commun. 284 (6) (2011) 1540–1547.

[42] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, J. Song, Large-scale gesture recognition with a fusion of RGB-D data based on the c3d model, in: Proceedings of the ICPR, 2016.

[43] S. Zhao, Y. Gao, G. Ding, T.S. Chua, Real-time multimedia social event detection in microblog, IEEE Trans. Cybern. (2017).

[44] L. Zheng, S. Wang, Z. Liu, Q. Tian, Fast image retrieval: query pruning and early termination, IEEE Trans. Multimed. 17 (5) (2015) 648–659.

[45] L. Zheng, S. Wang, Q. Tian, L(p) -norm IDF for scalable image retrieval, IEEE Trans. Image Process. 23 (8) (2014) 3604–3617.

[46] W. Wang, Q. He, A survey on emotional semantic image retrieval, in: Proceedings of the ICIP, 2008.

[47] 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.

[48] 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, 2014a.

[49] Y.Y. Chen, T. Chen, W.H. Hsu, H.Y.M. Liao, S.F. Chang, Predicting viewer affective comments based on image content in social media, in: Proceedings of the ICMR, 2014b.

[50] 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.

[51] Q. Miao, P. Xu, T. Liu, Y. Yang, J. Zhang, W. Li, Linear feature separation from topographic maps using energy density and the shear transform, IEEE Trans. Image Process. 22 (4) (2013) 1548–1558.

[52] T. Chen, D. Borth, T. Darrell, S.F. Chang, Deepsentibank: visual sentiment concept classification with deep convolutional neural networks, in: Proceedings of the CVPR, 2014.

[53] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the CVPR, 2014.

[54] 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.

[55] N. Morina, E. Leibold, T. Ehring, Vividness of general mental imagery is associated with the occurrence of intrusive memories, J. Behav. Therapy Exp. Psychiatry 44 (2) (2013) 221–226.

[56] C. Gan, Z. Gan, X. He, J. Gao, L. Deng, Stylenet: generating attractive visual captions with styles, in: Proceedings of the CVPR, 2017.

[57] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, X. Gao, Discriminative multi-instance multitask learning for 3d action recognition, IEEE Trans. Multimed. 19 (3) (2017) 519–529.

[58] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, X. Gao, Latent max-margin multitask learning with skelets for 3-d action recognition, IEEE Trans. Cybern. 47 (2) (2016) 439–448.

[59] 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.

[60] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.

[61] C. Fellbaum, G. Miller, Wordnet : an electronic lexical database, Lib. Q. Inf. Commun. Pol. 25 (2) (1998) 292–296.

[62] E. Loper, S. Bird, Nltk: the natural language toolkit, in: Proceedings of the ACL, 2002.

[63] 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.

[64] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the EMNLP, 2014.

[65] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[66] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the ICLR, 2015.

[67] X. Peng, H. Tang, L. Zhang, Z. Yi, S. Xiao, A unified framework for representation-based subspace clustering of out-of-sample and large-scale data, IEEE Trans. Neural Netw. Learn. Syst. 27 (12) (2016) 2499–2512.

[68] J. Yang, D. She, M. Sun, Joint image emotion classification and distribution learning via deep convolutional neural network, in: Proceedings of the IJCAI, 2017.

[69] X. Peng, C. Lu, Y. Zhang, H. Tang, Connections between nuclear-norm and Frobenius-norm-based representations, IEEE Trans. Neural Netw. Learn. Syst. 29 (1) (2018) 218–224.

[70] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemomet. Intell. Lab. Syst. 2 (1) (1987) 37–52.

[71] X. Chen, H. Fang, T.Y. Lin, R. Vedantam, S. Gupta, P. Dollar, C.L. Zitnick, Microsoft coco captions: data collection and evaluation server, in: Proceedings of the CVPR, 2015.

[72] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, C. Sienkiewicz, Rich image captioning in the wild, in: Proceedings of the CVPR workshops, 2016.

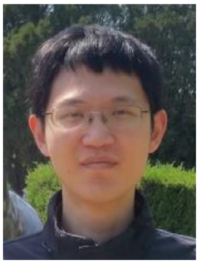
[73] P.J. Lang, M.M. Bradley, B.N. Cuthbert, International Affective Picture System (IAPS): Technical Manual and Affective Ratings, NIMH Center for the Study of Emotion and Attention, 1997, pp. 39–58.

[74] S. Zhao, H. Yao, Y. Gao, R. Ji, G. Ding, Continuous probability distribution prediction of image emotions via multitask shared sparse regression, IEEE Trans. Multimed. 19 (3) (2017) 632–645.

Jufeng Yang received the Ph.D. degree in control theory and engineering from Nankai University in 2009. He is currently an associate professor in the Department of Computer Science, Nankai University and was a visiting scholar with University of California, Merced. His research interests include computer vision, machine learning and multimedia computing.

Yan Sun received the B.S. degree from Nankai University and is a Ph.D. student in the same school. Her research intersts include computer vision and machine learning.

Jie Liang received the B.S. degree from the Ocean University of China and is a Ph.D. student in Nankai University. His research intersts include computer vision and machine learning.



Bo Ren received the B.S. and Ph.D. degrees from Tsinghua University in 2010 and 2015 respectively. He is currently a Lecturer at CCCE&CS, Nankai University. His main research interests are in Physically-based simulation and rendering, scene geometry reconstruction and analysis.

Shang-Hong Lai received the B.S. and M.S. degrees in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, and the Ph.D. degree in electrical and computer engineering from University of Florida, Gainesville, U.S.A., in 1986, 1988 and 1995, respectively. He joined Siemens Corporate Research in Princeton, New Jersey, as a member of technical staff in 1995. Since 1999, he became a faculty member in the Department of Computer Science, National Tsing Hua University, Taiwan. He is currently a professor and the department head in the same department. In 2004, he was a visiting scholar with Princeton University. Dr. Lais research interests include computer vision, visual computing, pattern recognition, medical imaging, and multimedia signal processing. He has authored more than 200 papers published in the related international journals and conferences. Dr. Lai has served as an area chair or a program committee member for a number of international conferences, including CVPR, ICCV, ECCV, ACCV, ICPR, PSIVT and ICME. He is also a program co-chair for ACCV'16 and several international workshops. Moreover, he has served as an associate editor for Journal of Signal Processing Systems and a guest editor for special issues in Journal of Visual Communication and Image Representation as well as Journal of Signal Processing Systems.