Journal of Intelligent & Fuzzy Systems 39 (2020) 7217–7231
DOI:10.3233/JIFS-200675
IOS Press

Visual sentiment analysis via deep multiple clustered instance learning

Wenjing Gao, Wenjun Zhang, Haiyan Gao and Yonghua Zhu*
Shanghai Film Academy, Shanghai University, Shanghai, China

Abstract. The increasing tendency of people expressing opinions via images online has motivated the development of automatic assessment of sentiment from visual contents. Based on the observation that visual sentiment is conveyed through many visual elements in images, we put forward to tackle visual sentiment analysis under multiple instance learning (MIL) formulation. We propose a deep multiple clustered instance learning formulation, under which a deep multiple clustered instance learning network (DMCILN) is constructed for visual sentiment analysis. Specifically, the input image is converted into a bag of instances through visual instance generation module, which is composed of a pre-trained convolutional neural network (CNN) and two adaptation layers. Then, a fuzzy c-means routing algorithm is introduced for generating clustered instances as semantic mid-level representation to bridge the instance-to-bag gap. To explore the relationships between clustered instances and bags, we construct an attention based MIL pooling layer for representing bag features. A multi-head mechanism is integrated to form MIL ensembles, which enables to weigh the contribution of each clustered instance in different subspaces for generating more robust bag representation. Finally, we conduct extensive experiments on several datasets, and the experimental results verify the feasibility of our proposed approach for visual sentiment analysis.

Keywords: Visual sentiment analysis, deep multiple clustered instance learning, fuzzy c-means routing, multi-head mechanism

1. Introduction

With the advent of social networks, people have been willing to express their opinions online via posting multimedia data. Among them, images are one of the most convenient and intuitive mediums for users to express ideas and convey moods. This gives rise to a great demand for an efficient approach to automatic visual semantics inference, which endeavors to recognize the content of an image and infers its high-level semantics. Image sentiment analysis is an important research direction in the field of image understanding, which studies the emotion response of humans on the images. The approaches developed for sentiment prediction on visual content, can be helpful to understand the users' behaviors and attitudes. They will further benefit social media communication and enable broad applications, e.g., affective image retrieval [1], opinion mining [2], comment assistant [3]. Therefore, how to infer the visual sentiment information has attracted increasing research attention.

However, visual sentiment analysis is more challenging than conventional recognition tasks due to the highly abstract nature of visual sentiment, which is originated from the semantic gap between low-level features and high-level semantics. Despite the challenges, various kinds of approaches have been proposed for visual sentiment analysis. Early studies on this issue explored hand-crafted features related to emotional expression, such as color, texture and shape. Inspired by the psychology and art theories, different groups of low-level features are manually designed to study the emotional reactions towards visual content [4]. However, the hand-crafted features are mostly effective on small datasets containing specific styles of images, such as artistic images.

*Corresponding author. Yonghua Zhu, Shanghai Film Academy, Shanghai University, Shanghai, China. E-mail: zyh_shu@vip.163.com.

ISSN 1064-1246/20/$35.00

Fig. 1. Some example images highlighted with visual elements contributing more to the evocation of sentiment.

In recent years, deep learning based models enable more robust feature representation than the traditional hand-crafted ones, making great changes in the research field of visual sentiment analysis. Some transfer learning based convolutional neural network (CNN) models [5, 6] have been employed for learning sentiment-level representation using training strategies and achieved significant advances. However, images in the same class of traditional cognitive-level recognition tasks mainly contain the same type of object. While in visual sentiment analysis, each class includes images with much more diverse contents, which results in large intra-class differences. It is difficult to discover discriminative features. To this end, it is necessary to take more cues into consideration for visual sentiment analysis. Many well-designed deep learning models have been developed to predict sentiment from images, which mainly fall into three branches according to the strategies they focused on. One branch is fusing textual information to visual sentiment analysis [7–9]. These studies focus on utilizing rich complementary information behind textual modality. Although, these approaches achieve improvement in visual sentiment analysis, they are insufficient to handle images without user-generated captions. Others attempt to bridge the semantic gap by learning mid-level feature representation of sentiment-related visual concepts [10–12]. However, most of them rely on pre-defined visual concepts, which may fail in tackling complex visual scenes since the same object may convey different sentiment information in different visual contexts. In addition, some researchers pay attention to the utilization of local information for sentiment analysis through discovering affective regions [13, 14]. However, an affective region is hard to define, and thus the detected regions may be different from real affective regions. Besides, the contextual information, which also reserves the characteristics of affective information, is underutilized in most existing local regions based methods.

In fact, the sentiment information in an image is delivered by multiple visual elements. Some visual elements contributing more to the evocation of sentiment are identified as affective local regions such as the visual elements highlighted with red bounding boxes in Fig. 1. Based on this observation, we convert the visual sentiment analysis into a multiple instance learning (MIL) problem. MIL is accomplished in handling the complex data in the form that each bag is associated with multiple instances. It can tackle problems through key instance selection and instance-to-bag relationships exploration. This enables a joint method combining the advantages of both mid-level representation learning and affective region discovery for tackling visual sentiment analysis. In details, the input image can be modeled as a bag consisting of multiple visual instances as mid-level representation. For sentiment classification, the bag can be re-represented by aggregating the instances based on their contributions to sentiment conveyance through a MIL pooling function.

To achieve this, we propose a deep multiple clustered instance learning network (DMCILN), which mainly contains three modules. A CNN model is first constructed to generate feature vectors for multiple visual instances from the input image and transform visual sentiment analysis into a MIL problem. Then, to bridge the semantic gap, we generate clustered instances by aggregating visual instances through a fuzzy c-means routing algorithm. Finally, a multi-head attention based MIL pooling layer is employed to determine the contribution of each clustered visual instances to sentiment evocation in different subspaces, based on which bag representation is produced for final sentiment classification.

Our main contributions to this field are summarized as follows.

First, according to the characteristics of visual sentiment conveyance, we investigate the problem of visual sentiment analysis by regarding it as a MIL problem. To achieve this, we propose an effective deep multiple clustered instance learning network (DMCILN), which can predict visual sentiment through key instance selection and instance-to-bag exploration. In this way, our method can brigde the

semantic gap and perceive affective regions. Various experiments have been conducted on several datasets to verify its effectiveness.

Second, within DMCILN, we design a fuzzy c-means routing algorithm to generate clustered instances through a joint optimization of feature representation learning and clustering. Without predefinition of mid-level concepts, our method can generate effective mid-level features in a weakly supervised way to reduce the research space of instance-to-bag relationships.

Third, a multi-head attention mechanism inspired MIL pooling function is constructed to produce discriminative bag representation by weighing the contribution of features to sentiment evocation. The automatic discovery of affective regions is realized by calculating the attentive weights of each local mid-level feature in different subspaces.

The remainders of this paper are organized as follows. Related works are reviewed in Section 2. Our approach to visual sentiment analysis is elaborated in Section 3. The experimental setup and analysis of the results are represented in Section 4. We finally conclude this work in Section 5.

2. Related work

2.1. Visual sentiment analysis

Learning a discriminative feature representation is crucial to visual sentiment analysis. Early works on visual sentiment analysis focused on designing a combination of low-level features inspired by psychology or art theories, color composition and SIFT-based shape descriptor included [15]. However, the problem of "semantic gap" cannot be well-solved by hand-crafted features. To tackle this, more advanced mid-level representation has been designed for visual sentiment analysis. SentiBank [10] and DeepSentiBank [11] were constructed to detect the existence of sentiment-related visual concepts (Adjectives Noun Pairs, ANPs) in the images as semantic representation. Similarly, Sentribute defined the mid-level representation based on scene-based attributes and eigenface for sentiment prediction [12].

For images, features extracted from non-emotional regions may generate classification noise. Rao et al. [16] employed MIL to determine dominant visual elements for emotion evocation from segmented images in a weakly supervised way. Then, object detection based approaches were utilized to generate region proposals containing semantic visual objects which may evoke emotion. Sun et al. [13] selected affective regions by ranking each region proposal based on its objectness scores. Rao et al. [17] proposed a feature pyramid network combined with region proposal network to generate multi-level representation for multiple local regions. However, crisp region proposals tend to find foreground objects in an image which may neglect the contextual or global information. To make up this limitation, She et al. [18] detected soft sentiment map by class activation mapping in a weakly supervised manner. Other researcher preferred soft regions discovered by attention mechanism based deep models [19].

Some researches consider both visual feature learning and affective region discovery to enhance the performance. For example, Wu et al. used attention mechanism to discover features of the region of interest under the guidance of visual attributes [20]. Different from existing methods in the literature that rely on pre-defined visual concepts, our model learns mid-level representation under weakly supervised MIL formulation with only sentiment-level labels. As for affective region discovery, the above mentioned methods produce either crisp regions proposal or attentive features over the global feature map. In contrast, our proposed DMCILN weighs the contribution of each clustered instance to sentiment prediction, which can certainly reduce the research space.

2.2. Multiple instance learning

2.2.1. Multiple instance learning with neural network

In the early researches on MIL, instances are mostly precomputed by certain feature extraction algorithms, which are then classified by specific instance-level or bag-level classifiers. Attracted by the capability of automatically representing functions and learning features, some researchers begin to approach MIL problems using neural network. Multiple instance neural network is constructed to learn instance representation and estimate instance probabilities. Different pooling operators are then adopted to calculate bag probability upon all instance probabilities like log-sum-exp operator [21] and max pooling operator [22]. Wang et al. [23] confirmed that realizing MIL with fully-connected neural network can be beneficial to bag-level prediction.

With the raise of deep learning, many researches have made efforts in combining MIL with deep neural

models. Ilse et al. [24] provided a general procedure for MIL and parameterized all transformations using neural network. They also proposed a permutation-invariant aggregation operator, the weights of which were obtained through training an attention mechanism. Lin et al. then leveraged attention based deep multiple instance learning for fashion outfit recommendation [25]. In this paper, we also leverage the attention-based MIL pooling function [24] for capturing the affective local regions across multiple inputs. The novelty of our model lies in that a multi-head mechanism is introduced to form MIL ensembles by exploring various instance-to-bag relationships in different subspaces, which is more robust for discovering key instances.

2.2.2. *Clustering based multiple instance learning*

It is believed that the multiple instance representation follows some patterns that can contribute to discriminate bags. Therefore, some researches utilize unsupervised algorithms like clustering to find the inherent structure of a dataset. Zhou et al. [26] introduced constructive clustering ensemble (CCE) to encode the bag by a binary feature vector indicating that whether a bag has an instance in one cluster or not. However, CCE considers only the presence of the cluster member, which might be problematic for the MIL problems encoding threshold or count-based assumptions. Then, Xu et al. [27] put forward the multiple clustered instance learning formulation (MCIL) by embedding clustering concept to distinguish multiple cancer subtypes. They first classified the segmented regions into different caner subtypes and then predicted the bag upon instance-level prediction. Based on the structural similarity of the problems, we also integrate clustering algorithm to uncover the mid-level feature structure for visual sentiment analysis. Different from them, DMCILN parameterizes all the functions in clustering based MIL with deep neural network, which is more adaptive to bag-level representation.

3. Approach

3.1. Deep multiple clustered instance learning

As the sentiment-related elements often lie in local regions in the images, we represent each image as many local features and assume that each local feature contributes to sentiment evocation in some degree. This treatment allows us to cast visual sentiment analysis as a MIL problem. Specifically, we follow the generalized MI assumption where positive bags are unable to be identified by a single instance but by the distribution of all instances. Under this formulation, the whole image conveying certain sentiment is regarded as a bag consisting many local features called instances. To further bridge semantic gap, we propose the concept of clustered instances generated from aggregating instances as mid-level representation. In this way, we can predict visual sentiment by exploring the relationships between clustered instances and bags. To realize it, a DMCILN is constructed, where all the functions are parameterized in one deep neural network. Hence, we called our solution to visual sentiment analysis as deep multiple clustered instance learning.

3.1.1. *MIL formulation*

In this section, we give a brief introduction to the generalized MIL assumption and the procedures of tackling MIL with neural network. Besides, we focus on MIL pooling functions, which serve as the building blocks of our proposed DMCILN. In the classical binary supervised learning problems, one aims at finding a function that predicts a category y_i for an input sample X_i modeled as x_i. In the case of MIL problems, there are multiple instances $\{x_{i1}, \ldots, x_{iN}\}$ that exhibit neither dependency nor ordering among each other in a bag X_i labeled as y_i. MIL is preferable to the analysis of multimedia data which has the multiple instance structure. For example, an image can be segmented into multiple local regions, which are then represented by a set of feature vectors derived from each region. As implied in [24], how to design a symmetric function for modeling the bag probability $\mathrm{S}(X_i)$ is a core issue for solving MIL problems, which should be in the following form:

$$\mathrm{S}(X_i) = g\left(\sum_{\sigma} f(X_i)\right) \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ denote suitable transformation functions. There are mainly two utilities of transformation functions, called instance-level approach and embedding-level approach respectively. In this paper, we only focus on the embedding-level approach, which is preferable in terms of the bag-level classification performance compared to instance-level approach [23]. The embedding-level approach maps the jth instance to feature representation x_{ij} via function $f(\cdot)$. Then the bag embedding $\mathrm{h}(X_i)$ is generated

from the aggregation of instance representation through a permutation-invariant pooling function $\sigma(\cdot)$. Finally, the class probability is obtained via function $g(\cdot)$ over bag embedding.

For maximum pooling operator,

$$\mathrm{h}(\mathrm{X_i}) = \max_{\mathrm{j}=1,\ldots,\mathrm{N}} \left\{\mathrm{x_{ij}}\right\} \tag{2}$$

For mean pooling operator,

$$\mathrm{h}(\mathrm{X_i}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{x_{ij}} \tag{3}$$

These fixed MIL pooling operators have a clear disadvantage, namely, they are pre-defined and non-trainable. Hence, one flexible and adaptive MIL pooling is proposed to achieve better results by adjusting to a specific task, i.e., attention-based MIL pooling [24]. The attention based MIL pooling averages the weighted instances where weights are learned by training the neural network.

For attention-based pooling operators,

$$\mathrm{h}(\mathrm{X_i}) = \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{a_{ij}} \mathrm{x_{ij}} \tag{4}$$

$$\mathrm{a_{ij}} = \frac{\exp\left\{\mathrm{W^T tanh}\left(\mathrm{V x_{ij}^T}\right)\right\}}{\sum_{\mathrm{k}=1}^{\mathrm{K}} \exp\left\{\mathrm{W^T tanh}\left(\mathrm{V x_{ik}^T}\right)\right\}} \tag{5}$$

3.1.2. DMCIL formulation on visual sentiment analysis

In this paper, we approach visual sentiment analysis mainly by addressing two issues: the modeling of visual elements and the exploration of how visual elements contribute to the evocation of visual sentiment. To this end, visual sentiment analysis is modeled as a generalized multiple instance learning problem. Under generalized MIL assumption, each input image is in a form of a bag that contains a set of visual instances. For visual sentiment analysis, the goal is to predict the bag-level categorization by modeling the instances and exploring the interaction between visual instances and bags. In details, a permutation-invariant MIL pooling function is utilized to realize the transformation from instance space to bag space, by which various relationships between visual instances and sentiment are excavated. Specifically, the ith image is denoted as a bag X_i; the jth image patch sampled from an image corresponds to a visual instance representation x_{ij}. The bag-level label of the bag X_i is defined as y_i. The feature representation of a bag is obtained by the weighted-sum of the transformed instance as Equation (6), where a_{ij} denotes the contribution of the jth visual instance representation x_{ij} to the bag embedding $\mathrm{h}(X_i)$.

$$\mathrm{h}(\mathrm{X_i}) = \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{a_{ij}} \mathrm{x_{ij}} \tag{6}$$

However, it is complicated to explore the relationships between visual instances and the corresponding sentiment as the features of visual instances generated by CNN model are locally not semantically sampled. As the visual entities serve as the basic semantic units of an image, one good way is to model the mid-level representation with the feature vector of a specific visual entity. Traditional MIL method is not capable of obtaining such semantic mid-level representation without instance-level labels.

To tackle this, we embed the concept of clustering into the MIL setting and propose deep multiple clustered instance learning which assumes that there is implicit semantic information between instances and bags that can be captured by clustering similar patches. Particularly, as illustrated in Equation (7), the instances of all the bags are clustered into K groups through function $p(\cdot)$ and the center representation of the kth group is regarded as clustered instance representation v_{ik}. By this way, the mid-level representation is obtained, and then MIL pooling $\sigma(\cdot)$ is performed to explore the interaction between clustered instances and sentiment and aggregate mid-level representation. In this way, each bag is re-represented by one feature vector $\mathrm{h}(X_i)$ so that single-instance classifiers can be used to distinguish different classes of bags.

$$\mathrm{v_{ik}} = \mathrm{p}(\mathrm{x_{i1}}, \ldots, \mathrm{x_{iN}}) \tag{7}$$

$$\mathrm{h}(\mathrm{X_i}) = \sum_{\sigma} \mathrm{a_{ik}} \mathrm{v_{ik}} \tag{8}$$

The details of clustering function $p(\cdot)$ and obtaining bag embedding $\mathrm{h}(X_i)$ are stated in the following sections. The major differences among classical supervised learning, MIL and MCIL are illustrated as Fig. 2. In this paper we parameterize the transformation $f(\cdot)$, $p(\cdot)$ and $g(\cdot)$ as well as the pooling function $\sigma(\cdot)$ using deep neural networks for more flexibility and end-to-end optimization.

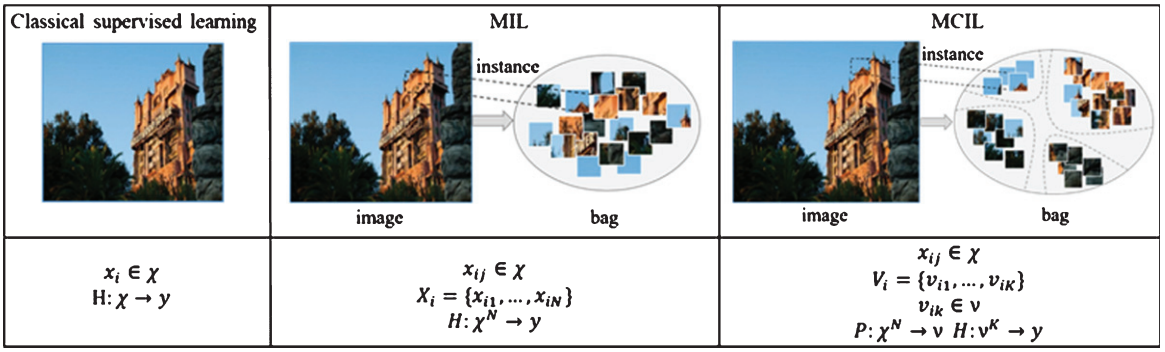

Fig. 2. Distinct problem formulations and learning goals among classical supervised learning, MIL, and MCIL.

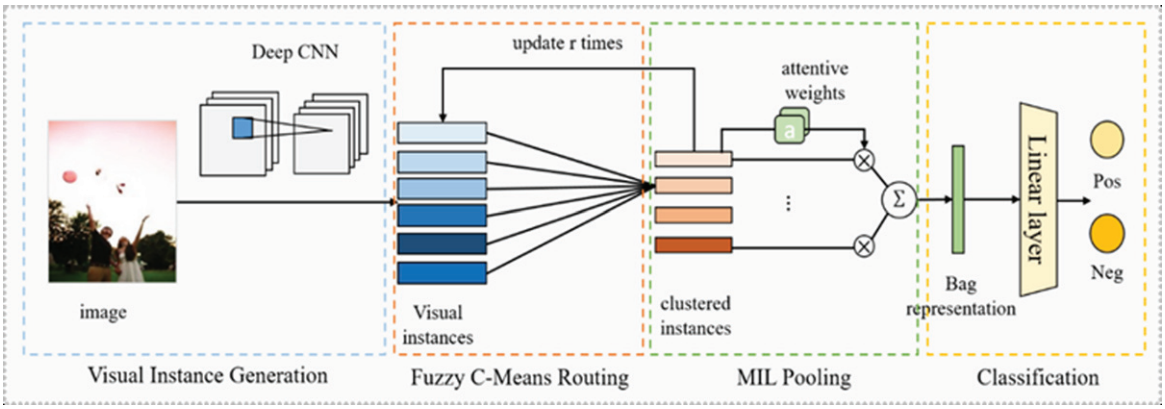

Fig. 3. The framework of our proposed DMCILN for visual sentiment analysis.

3.2. *The architecture of DMCILN*

In order to realize visual sentiment analysis under the formulation of DMCIL, we propose a deep multiple clustered instance learning network. Our proposed DMCILN can bridge the semantic gap by learning a set of mid-level representation in a weakly supervised way, over which affective regions can be discovered to improve the performance of visual sentiment analysis. Figure 3 demonstrates the framework of our proposed DMCILN, which consists of four parts: visual instance generation, fuzzy c-means routing, MIL pooling and sentiment classification.

Visual instance generation part is a specific-designed CNN model built to extract a set of feature maps from the holistic image information. By adding two adaption layers, the pre-trained CNN can be better adapted to model multiple instances. We regard one image as a bag and the concatenation of feature maps across all the channels as multiple instances. By this way, visual sentiment analysis can be transformed into a MIL problem.

For reducing the gap between instances and bags, we design a fuzzy c-means based dynamic routing layer to generate clustered instances. Different from traditional clustering based MIL methods, we introduce fuzzy c-means into deep neural network to generate clustered instances by jointly optimizing feature representation learning of instances and clustering in an end-to-end manner. The clustered instances are capable of keeping the major semantic features of instances in a robust way.

In MIL pooling and sentiment classification part, the bag representation is generated from all the clustered instance representation through an attention-based MIL pooling function. We innovatively embed the multi-head mechanism into MIL pooling layer to form ensembles. MIL ensembles can explore various combinations of visual elements for generating a specific bag in different subspaces. Finally, a sentiment classification layer is built on top of the bag representation for visual sentiment prediction.

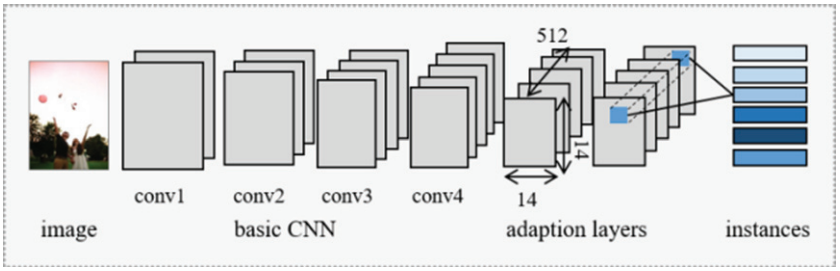

Fig. 4. The architecture of visual instance generation module.

The details of these four parts are discussed in Section 3.2.1, Section 3.2.2, Section 3.2.3 and Section 3.2.4, respectively.

3.2.1. Visual instance generation

The deep multiple clustered instance learning assumption in visual sentiment analysis assumes that an image can be decomposed into many visual instances. Thus, each image can be represented as a multi-instance bag. The first step towards visual sentiment analysis under the framework of the DMCIL is multiple instance modelling, which aims to transform each image into a multi-instance bag. Since MIL problems assume that there is neither dependency nor ordering relationship among all the instances in a bag, we propose to utilize convolutional neural network for extracting local features representing each instances. In this section, we describe the generation of visual instances from images by exploiting the architecture of convolutional neural network as shown in Fig. 4.

To map an input image X_i into a bag of instances $\{x_{ij}\}$, we design a CNN model $f(\cdot)$ to extract a collection of feature maps x_i representing instances as shown in Equations (9) and (10). To guarantee the quality of the learned instance representation, the pre-trained model weights on ImageNet dataset are employed to initialize our CNN model. However, the original task of image classification treats each image as one instance, which mixes the visual information corresponding to different visual entities and brings difficulty in adapting to learning multiple instance representation. Hence, in this paper, we first extract feature representation from the input image on the basis of a pre-trained CNN, which is obtained by removing the last two convolutional layers and the following pooling and FC layers of VGGNet [28]. Then, two adaptation layers are added on top of the pre-trained CNN to better adapt to model multiple instances [29]. Two adaptation layers both contain 512 convolutional filters of size 3×3 and are followed by a ReLU [30] activation function and a dropout layer.

$$x_i = f(X_i) \tag{9}$$

$$\mathrm{x_i} = \left[\mathrm{x_{i1}}, \ldots, \mathrm{x_{ij}}, \ldots, \mathrm{x_{iN}}\right], \mathrm{x_{ij}} \in \mathbb{R}^{\mathrm{D}} \tag{10}$$

The feature vectors generated at the bottom of the CNN model are regarded as the overview of visual appearance. Specifically, the final layer generates feature maps in size of $14 \times 14 \times 512$. We follow [31] to generate the representation for each instance by concatenating feature values at the same location across different channels of the feature map. By this way, each of the 14×14 feature vector with dimension of $\mathrm{D} = 512$ can be treated as one instance representation of the input image. Here, each convolutional feature map corresponds to one specific local region in the input image and thus can be utilized as an instance.

3.2.2. Clustered instance generation

The feature maps generated from CNN model are scanned channel-wise to produce the so-called visual instances. By this way, an image is modeled with multiple instances. However, the visual instances generated from CNN model fail to be semantically meaningful since an instance corresponds to only a small part of local region in the input image. To further bridge the instance and bag feature space, our target is to learn an effective mid-level representation which can be served as semantic visual elements.

As we all know, the probabilistic based clustering algorithms (e.g. K-means) are efficient in discovering hidden visual patterns and keeping the major features of similar patches by producing cluster centroids. To this end, we attempt to utilize clustering for maintaining the important information of visual instances. However, introducing the traditional unsupervised clustering to generate mid-level representation in

DMCIL formulation may confront with some problems. To name a few, the data allocation of each cluster is done in a crisp manner, and thus the quality of mid-level representation relies highly on the clustering performance, which is lack of robustness. Besides, they are incapable of jointly optimizing feature representation learning of instances and clustering in an end-to-end manner.

To tackle these challenges, a fuzzy c-means routing algorithm is proposed to obtain clustered instances, where the connection between low-level instances and mid-level representation are determined based on fuzzy c-means clustering. Fuzzy c-means routing algorithm works in two aspects: first, we utilize a soft data allocation strategy for generating more robust features, which relies less on the clustering performance; second, it supports feature representation learning by training the network end-to-end.

We first briefly review the basic fuzzy c-means algorithm and realize it by a dynamic routing approach [32]. Given a set of feature vectors of instances in an image $\{x_{i1}, \ldots, x_{iN}\}$, the number of clustered instances K, and fuzzification parameter $\theta(\theta > 1)$, the feature vectors of clustered instance are identified using fuzzy c-means algorithm by optimizing the following objective function, where function $d(\cdot)$ computes the distance between each instance and one cluster center using Euclidean distance $|x_{ij} - v_{ik}^2|$ and the $\phi_{ij,k}$ represents the fuzzy membership indicating the degree that the jth sample belongs to the kth cluster:

$$\mathrm{L}(\Phi, \mathrm{v}) = \sum_{\mathrm{j}=1}^{\mathrm{N}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \left(\phi_{\mathrm{ij,k}}\right)^{\theta} \mathrm{d}\left(\mathrm{x_{ij}}, \mathrm{v_{ik}}\right) \tag{11}$$

Traditional fuzzy c-means iteratively updates fuzzy membership $\phi_{ij,k}$ and cluster center v_{ik} with a random initial estimate of cluster centers and terminates when objective loss is below a specified tolerance. In this paper, we introduce a routing process to realize fuzzy c-means algorithm in the end-to-end network. The fuzzy c-means routing is a routing process between instances and clustered instances, which iteratively updates the representation of clustered instance by aggregating the instance features. Firstly, the instance space is transformed into clustered instance space by a transformation matrix $W_{ij,k}$ [33]. The initial K clustered instances are set by a weighted sum of transformed instances as Equation (12). Then, we iteratively determine the contribution of each instance to cluster centers and update the representation of clustered center.

$$\mathrm{v_{ik}^0} = \frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{W_{ij,k}} \mathrm{x_{ij}} \tag{12}$$

Particularly, we compute the fuzzy membership $\phi_{ij,k}$ of the jth transformed instance belonging to the kth cluster. The representation of cluster centers is updated by evaluating the influence of the jth component on the kth cluster center with their fuzzy membership $\phi_{ij,k}$ in each iteration. The larger the fuzzy membership $\phi_{ij,k}$ is, the higher impact of the jth instances have on the kth clustered instance. By aggregating the transformed instance based on their fuzzy memberships, the clustered instances are capable of keeping the major semantic features of instances in a robust way. The whole procedures are summarized on algorithm 1.

ALGORITHM 1: Fuzzy c-means based dynamic routing

Input: the collection of visual instances $\left\{x_{ij}\right\}$, and iteration time r

Initialize $v_{ik}^0 \leftarrow \frac{1}{K} \sum_{i=1}^{n} W_{ij,k} x_{ij}$

for r iteration **do**

  for all j and k **do**

$$\phi_{ij,k} \leftarrow \frac{d\left(W_{ij,k} x_{ij}, v_k\right)^{1/1-\theta}}{\sum_{c=1}^{K} d\left(W_{ij,k} x_{ij}, v_c\right)^{1/1-\theta}}$$

  for all k **do**

$$v_{ik}^r \leftarrow \frac{\sum_{j=1}^{N} \phi_{ij,k}^{\theta} W_{ij,k} x_{ij}}{\sum_{j=1}^{N} \phi_{ij,k}^{\theta}}$$

end

for all k **do**

$$v_{ik} \leftarrow ReLU\left(v_{ik}\right)$$

end

Output: the collection of clustered instances $\{v_{ik}\}$

Through fuzzy c-means routing, we obtain K centroids in the clustered instance space. Each clustered instance can be regarded as a mid-level representation obtained from aggregating local features, which reflects the affective information to a certain degree. The final clustered instances are obtained after ReLU activation.

3.2.3. MIL pooling layer

Given all the clustered instances, the next step is to generate bag representation for sentiment classification. Our DMCIL formulation assumes that the sentiment is conveyed through the interaction among collections of visual instances in the image. All the clustered instances are generated from aggregating local features, which exhibits neither dependency nor

ordering among each other. Some clustered instances evoking stronger sentiment than others are identified as key clustered instances. To explore how various combinations of clustered instances form a specific sentiment-level bag, we construct an attention based MIL pooling layer to transform the clustered instance space to bag space. Specifically, the attention-based MIL pooling layer can be expressed as a weighted sum pooling over clustered instances where each instance weight a_{ik} is determined by the attention mechanism. The function for generating bag representation $\mathrm{h}(X_i)$ from clustered instance $\{v_{ik}\}$ can be expressed as Equation (13).

$$\mathrm{h}(\mathrm{X_i}) = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{a_{ik}} \mathrm{v_{ik}} \tag{13}$$

The clustered instance weight a_{ik} is a scalar describing the contribution of the kth clustered instance to its bag representation $\mathrm{h}(x_i)$. Based on these weights, all the clustered instances $\{v_{ik}\}$ are aggregated into bag representation in a weighted-sum pooling fashion. The clustered instance weights, are determined using Equation (14) as [24]:

$$\mathrm{a_{ik}} = \frac{\exp\left\{\mathrm{W^T tanh}\left(\mathrm{V v_{ik}^T}\right)\right\}}{\sum_{\mathrm{k=1}}^{\mathrm{K}} \exp\left\{\mathrm{W^T tanh}\left(\mathrm{V v_{ik}^T}\right)\right\}} \tag{14}$$

However, based on the evocation mechanism of visual sentiment, there might be various combinations of visual elements for generating a specific bag, among which the key instances may vary. Since there is no criterion available for judging which kind of combination result is the best for generating bag representation, a possible solution is to produce many different combinations and then combine their results. This practice can be regarded as one kind of MIL ensemble strategies which are known to be much more robust for prediction than one MIL function. Therefore, we propose to design MIL ensembles, the goal of which is obtaining a strong bag representation from a set of individual learners to improve classification performance.

To obtain MIL ensembles for the exploration of various clustered instances-to-bag relationships, we train an ensemble of multiple instance pooling layers, where multiple attention maps are created to select key clustered instances in different aspects. In particular, we project the clustered instance representation into M lower-dimensional subspaces, for each of which the above-mentioned attention function is performed as shown in the Equation (15). Each attentive weight a_{ik}^m represents the importance of the kth clustered instance to bag representation computed in the mth subspace.

$$\mathrm{a_{ik}^m} = \frac{\exp\left\{\left(\mathrm{W_2^m}\right)^{\mathrm{T}} \tanh\left(\mathrm{W_1^m v_{ik}^T}\right)\right\}}{\sum_{\mathrm{k=1}}^{\mathrm{K}} \exp\left\{\left(\mathrm{W_2^m}\right)^{\mathrm{T}} \tanh\left(\mathrm{W_1^m v_{ik}^T}\right)\right\}} \tag{15}$$

When fusing the output of each subspace, we consider two fusion operators for producing the integrated bag representation: average fusion and concatenation fusion. The average fusion is operated on multiple attention maps in different subspaces to obtain the final attentive feature map as shown in Equation (16). While the concatenation fusion computes the attended bag representation in each subspaces and concatenates them to obtain the integrated representation. Both two operations can explore the various transformation of clustered instances to bag representation by integrating the results of all the subspaces.

$$\mathrm{a_{ik}} = \frac{1}{\mathrm{M}} \sum_{\mathrm{m=1}}^{\mathrm{M}} \mathrm{a_{ik}^m} \tag{16}$$

3.2.4. Sentiment classification layer

After obtaining bag representation from the weighted aggregation of clustered instance, the multi-instance learning problem is converted into single instance learning, which can be tackled by a sentiment classification layer. In DMCIL, the sentiment classification layer is set to a fully-connected layer with non-linear activation function so that the bag representation is transferred to a vector d^{c} with length of C, where C is the number of sentiment categories. In visual sentiment analysis, the sentiment categories can be either positive or negative.

$$\mathrm{d^c} = \tanh\left(\mathrm{W_c h}\left(\mathrm{X_i}\right) + \mathrm{b_c}\right) \tag{17}$$

Then the probability distribution p_c over the sentiment categories is computed using softmax function as Equation (18), where $\mathrm{d_c^C}$ denotes the value of the cth category in d^{c}:

$$\mathrm{p_c} = \frac{\exp\left(\mathrm{d_c^C}\right)}{\sum_{\mathrm{k=1}}^{\mathrm{C}} \exp\left(\mathrm{d_k^C}\right)} \tag{18}$$

To train the whole DMCILN, a loss function L for guaranteeing sentiment classification results of the input image is defined as follows. In this work, we use the cross-entropy error between gold sentiment distribution p_C^g and predicted sentiment distribution

p_C as the loss function:

$$\mathrm{L} = -\sum_{c=1}^{\mathrm{C}} p_C^g \cdot \log\left(p_C\right) \tag{19}$$

4. Experiment

4.1. Experiment settings

4.1.1. Dataset

We evaluate our proposed method on 4 public datasets, including Flickr and Instagram(FI) [34], Flickr [35], EmotionROI [36] and Twitter [5].

Flickr and Instagram (FI) is a well-labeled affective dataset crawled from Flickr and Instagram, which is collected by querying with eight emotion categories as keywords i.e., anger, amusement, awe, contentment, disgust, excitement, fear, sadness. The crawled images are further labeled by AMT workers. We reserve the labeled images receiving at least three agreements which are totally 23,308.

Flickr is constructed from the images provided by [35]. The images in the dataset are labeled with sentiment labels via crowd-sourcing. There are totally 60,745 images labeled with positive or negative.

Emotion ROI consists of 6 emotion categories corresponding with Ekman's 6 basic emotions (anger, disgust, joy, fear, sadness and surprise), with 330 images per category. The total number of images in this dataset is 1980.

Twitter I is built from those tweets containing images. It contains 1269 images totally, which are labeled with sentiment labels by AMT workers. We reserve those images for which at least three AMT workers give the same sentiment label.

Twitter II contains 603 images collected from the Twitter website. The ground-truth labels are given by AMT workers, and 470 samples are labeled positive.

For binary sentiment classification, multi-emotion labels are mapped into two categories: positive and negative. For dataset FI, we divided the eight emotion into binary sentiment categories. The labels of amusement, contentment, excitement, and awe are mapped to positive category, and images identified as sadness, anger, fear, and disgust are labeled as negative. For the six categories in EmotionROI, images with labels of anger, disgust, fear, and sadness are labeled as negative, and those with joy and surprise are positive.

4.1.2. Baseline models

To demonstrate the effectiveness of our proposed DMCILN for visual sentiment analysis, we evaluate our model against the following baselines including methods based on hand-crafted features, mid-level representation, transfer learning, affective regions.

For methods using hand-crafted features, we compare with the principle-of-art features proposed in Zhao et al. [4]. For mid-level representation based approaches, we utilize SentiBank [10] and pre-trained DeepSentiBank [11] to discover ANP concepts as feature vectors. SentiBank exploits 1200 dimensional features detected by a concept detector library. While the pre-trained DeepSentiBank is adopted to extract features followed by a fully-connected layer for sentiment classification. For transfer learning based deep models, the fully-connected features are extracted from the VGGNet trained on ImageNet and are classified by LIBSVM for visual sentiment classification. We also evaluate the results of VGGNet with16 layers, which adopts the pre-trained weights on ImageNet dataset and is fine-tuned on the experimental datasets. A progressive CNN model proposed in You et al. [5] is also compared with our method. We fine-tune the model in the noisy labeled dataset with VGGNet architecture for visual sentiment analysis. For methods utilizing affective regions, we use Sun's method [13] to select top-1 crisp region from off-the-shelf tools and combine the holistic features with the local features. In addition, we design a variant of our model called DMILN to show the feasibility of the generation of clustered instances by fuzzy c-means routing. This network is composed of only visual instance generation, multi-head attention-based MIL pooling and sentiment classification.

4.1.3. Implementation details

In this section, we describe how our proposed DMCILN is trained for visual sentiment analysis. For data augmentation, we apply random horizontal flips to the original images and randomly crop into a 224×224 sub-image to get more images, followed by a normalization process. Following transfer learning, the pre-trained layers of the CNN model in visual instance generation are initialized with the weights trained on ImageNet dataset. The output of the conv4_3 layer is utilized as the input of the following adaption layers. Particularly, we only compute the gradients of the first iteration in fuzzy c-means routing to make sure the effectiveness of the iterative process. During training, the SGD optimizer is

Table 1
Sentiment classification accuracy (%) on three datasets, including FI, Flickr, EmotionROI and Twitter. The best performances of all models are represented in bold

Methods	Datasets				
	FI	Flickr	EmotionROI	Twitter I	Twitter II
Zhao et al. [4]			75.24	67.92	67.51
SentiBank [10]			66.18	66.63	65.93
DeepSenti Bank [11]	61.54	70.16	70.11	71.25	70.23
VGG16	70.64	73.14	72.25	75.49	72.61
Fine-tuned VGG16	83.05	78.14	77.02	76.75	76.99
You et al. [5]	75.34	75.67	73.58	76.36	77.64
Sun et al. [13]		79.85		**81.06**	**80.84**
DMILN	82.45	78.35	77.69	77.63	77.01
DMCILN	**85.12**	**81.83**	**80.53**	80.80	80.12

adopted with the mini-batch size set to 10, and the learning rate for pre-trained convolutional layers and the remaining parts are set as 0.00001, 0.0001 respectively. The total iterations are 20 epochs, while the learning rate drops by a factor of 10 every 10 epochs. The FI dataset is split randomly into 80% for training, 5% for validation and 15% for testing. For Flickr dataset, we randomly sample 12606 images and split it into 90% training set and 10% for testing set. The rest of the datasets are all randomly split into 80% for training and 20% for testing.

4.2. Experiment results

4.2.1. Comparison with the baselines

We compare the classification performance of our model with the above-mentioned baselines. We set the fuzzy parameter $\theta = 1.5$, cluster number $K = 12$, multi-head number $M = 2$ and concatenation fusion as the default setting. Table 1 reports the performance of the baselines along with our proposed DMCILN. The hand-crafted features and SentiBank perform worse than deep neural models, which verifies the feasibility of deep representation.

As we can see, compared to SentiBank and DeepSentiBank, which detects the pre-defined visual concepts as mid-level representation, our model makes an obvious improvement in all datasets. The reasons lie in two aspects. First, our model learns mid-level representation in a weakly supervised way rather than detecting concrete visual concepts, which is proven more effective and robust for semantic gap reduction. Second, our model weighs the importance of local features instead of treating them equally, which can suppress noisy patches. Our method also shows an advantage over You's method and fine-tuned VGGNet, which demonstrates that our model is capable of learning more discriminative features under the multiple clustered instance learning formulation than the transfer learning based CNN models with different training strategies. As for affective region based method, our method outperforms Sun's method in Flickr while lags behind them in small-scale datasets Twitter I and Twitter II. This indicates that our architecture need more training images to achieve more satisfied results. The advantages of our model lies in that it can learn discriminative representation for visual sentiment analysis in an end-to-end weakly supervised way rather than using off-the-shelf tools. Besides, it takes no trouble to fuse localized and holistic representations. Finally, we compare with the variant DMILN without the generation of mid-level representation. From Table 1 we can see that the performance of DMILN is similar to that of fine-tuned VGGNet in that DMILN directly attend to each instance without the valid bridge between low-level features and high level semantics. While DMCILN achieves the improvement of about 3% accuracy over DMILN and fine-tuned VGGNet. This proves the effectiveness of the fuzzy c-means routing module by introducing mid-level representation between instances and bags.

4.2.2. Comparison of different pooling functions

Under the multiple clustered instance learning assumption, each image is viewed as a bag, and its local features are regarded as clustered instances. The DMCILN generates bag representation by

aggregating the clustered instances through a kind of pooling function, which converts multi-instance learning problem to single-instance classification that can be solved by a classification layer. In this case, the choice of pooling function is an important decision. In order to verify the effectiveness of our proposed multi-head attention based MIL pooling function, we compare the following types of pooling functions both theoretically and experimentally.

1) Max-pooling function simply takes the largest value in the feature dimension across all the instance representation. 2) Average-pooling function assigns an equal weight to each instance and takes an average of the feature representation of all instances. The first two pooling functions are pre-defined and non-trainable. 3) Attention based pooling function [24] learns the weights of each instance representation by optimizing the network, which is a flexible and adaptive MIL pooling for achieving better results by adjusting to a task and its dataset. 4) Average operation based multi-head attention pooling function (average+multi-head) computes the attentive weights in two subspaces and then takes an average of them to obtain the final attentive weights. 5) Concatenation operation based multi-head attention pooling function (concatenation+multi-head) generates the attended bag embedding in two subspaces and then concatenates them to get the final bag representation.

To evaluate the effect of different MIL pooling functions on visual sentiment analysis, we propose three variants of our model with different pooling functions. The results are shown in Fig. 5. We can see that the average-pooling obtains the lowest accuracy on both FI and Flickr datasets as it considers all the clustered instances equally. Max-pooling function performs a little better than average pooling function. This can be explained by that this mechanism focuses on the most important information rather than treating them equally.

The performance of both attention-based pooling function and multi-head attention-based pooling functions are comparatively more satisfying than the fixed pooling functions, i.e., max-pooling and average-pooling. This demonstrates that the weights computed by training the network allow dynamically selecting features for more effective sentiment classification.

As for the strategy of MIL ensembles, the multi-head mechanism in attention-based pooling gives a 0.2–0.5 percent improvement compared to attention based pooling function. From this, we may draw the conclusion that the bag representation obtained by

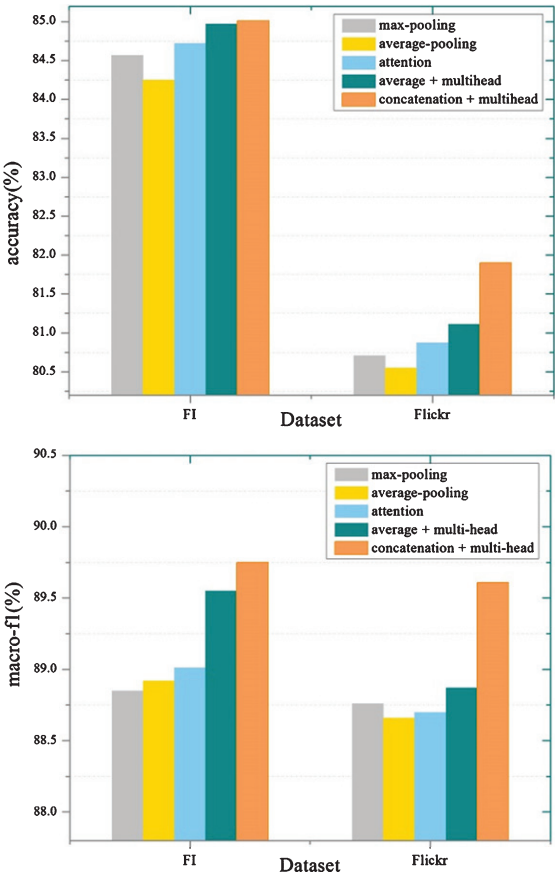

Fig. 5. Accuracy and macro-f1 of different pooling functions on the testing set of the FI and Flickr datasets.

aggregating attentive weights in different subspaces is more robust than single MIL pooling layer. Besides, Fig. 4 demonstrates the comparison of two fusion operations, which illustrates that concatenation is the most effective way since it retains all the information.

4.2.3. Comparison of different K values

As stated above, our model generates clustered instances as semantic mid-level representation by grouping the instances into K clusters using fuzzy c-means routing. In terms of unsupervised clustering, the value of K will have impact on the quality of the learned mid-level representation. Under the multiple clustered instance learning formulation, we treat each clustered instances as the aggregation of local features depend on their similarity. To this end, we assume that the value of K can be regarded as the average number of semantic visual entities in the input images. If K is set higher, there will be less impact on reducing semantic gap. Otherwise, if K is set lower, it is difficult for clustering to give a good

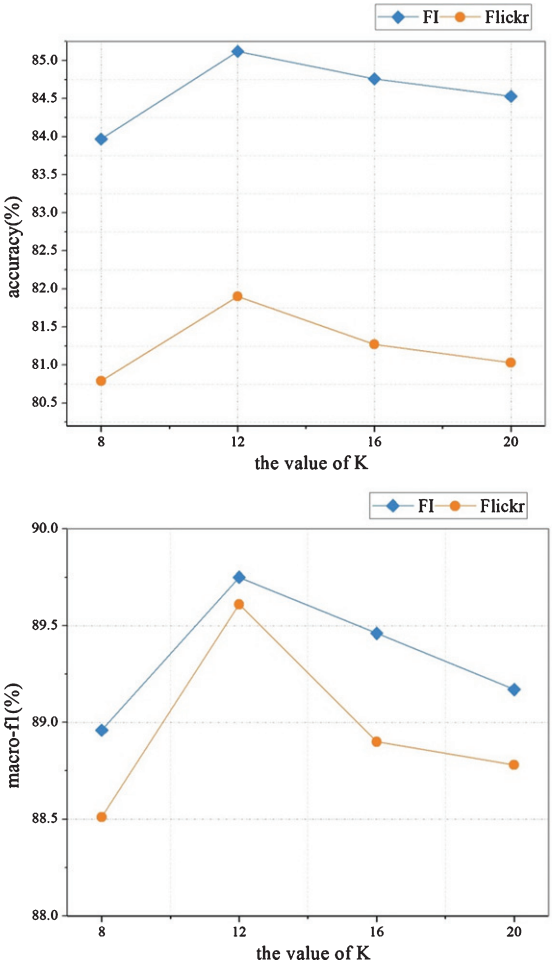

Fig. 6. Accuracy and macro-f1 of different K values on the testing set of the FI and Flickr datasets.

performance as well as some visual concepts may be neglected. Therefore, we report the classification accuracy and macro-f1 of our model with different K values on two large scale datasets, i.e., FI and Flickr. Intuitively, we set the possible values of K to 8, 12, 16 and 20. The results are shown in Fig. 6. From Fig. 6, we can see that both accuracy and f1-macro first increase when the value of K increases and then decreases. The peak value reaches for accuracy and f1-macro when K = 12.

4.2.4. Visualization of fuzzy c-means routing

In this section, we'd like to explore the semantics of the clustered instances generated from fuzzy c-means routing algorithm. Parts of the clustering results are visualized to help us understand the potential meaning of mid-level representation generated from DMCILN.

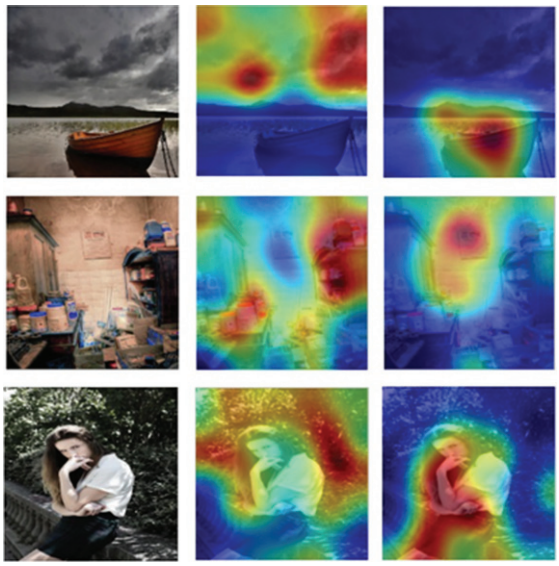

Fig. 7. The visualization of some key clustered instances generated from our DMCILN. The sample images are selected from FI dataset.

Under the framework of deep multiple clustered instance learning, the input image is modeled with multiple visual instances by visual instance generation module. Thus, each instance corresponds to a multi-channel feature vector sampled from a 14×14 local region. According to the fuzzy c-means routing algorithm, each clustered instance is obtained by weighted sum of all the instances, where the weight is defined as the fuzzy membership of each instance to the specific cluster. Inspired by the visualization of visual attention, we visualize the weight matrixes of all the fuzzy memberships. Specifically, the size of weight matrix for a cluster is 14×14 while the original image is 224×224. Hence, we upsample the weight matrix to the same size of the original image by a factor of 16 and then filtered by gaussian filter to generate heatmaps. One heatmap represents the fuzzy memberships of all instances in one cluster which can reflect the visualization of clustering results. From the sample images shown in Fig. 6, we can observe that, this kind of fuzzy clustering can capture the mid-level visual concepts by grouping similar patches. For example, as shown in the first row of Fig. 7, two visual concepts including "ship" and "sky" from local regions are mainly aggregated as mid-level representation after fuzzy clustering. This verifies the assumption of our proposed DMCIL where mid-level visual concepts can serve as semantic information between instances and bags.

5. Conclusions

In this paper, we present a deep multiple clustered instance learning network, which models the input image as a multi-instance bag and addresses the visual sentiment analysis under the formulation of deep multiple clustered instance learning. This end-to-end deep neural network realizes visual sentiment analysis by a joint work of semantic mid-level representation learning and affective regions discovery. In particular, a fuzzy c-means based routing algorithm is designed to generate clustered instances, which is able to learn both instance representation and fuzzy clustering. Besides, we also introduce a multi-head attention based MIL pooling layer for weighing the contribution of each feature representation over mid-level representation in different subspaces. The results of experiments on several datasets demonstrate that the DMCIL formulation inside our model has a distinct improvement in results over baseline models when performing visual sentiment analysis.

Acknowledgments

This work is supported by the National Key Research and Development Plan of China (No. 2017YFD0400101)

References

[1] L. Pang, S. Zhu and C.-W. Ngo, Deep Multimodal Learning for Affective Analysis and Retrieval, *IEEE Trans Multimedia* **17** (2015), 2008–2020. https://doi.org/10.1109/TMM.2015.2482228.

[2] Q.-T. Truong and H.W. Lauw, Visual Sentiment Analysis for Review Images with Item-Oriented and User-Oriented CNN. In *Proceedings of the 25th ACM international conference on Multimedia (MM 2017)*, 1274–1282. https://doi.org/10.1145/3123266.3123374.

[3] Y.-Y. Chen, T. Chen, T. Liu, H.-Y.M. Liao and S.-F. Chang, Assistive Image Comment Robot—A Novel Mid-Level Concept-Based Representation, *IEEE Trans Affective Comput* **6** (2015), 298–311. https://doi.org/10.1109/TAFFC.2014.2388370.

[4] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua and X. Sun, Exploring Principles-of-Art Features For Image Emotion Recognition, in: *Proceedings of the ACM International Conference on Multimedia - MM '14, ACM Press, Orlando*, Florida, USA, **2014** pp. 47–56. https://doi.org/10.1145/2647868.2654930.

[5] Q. You, J. Luo, H. Jin and J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI conference on artificial intelligence*. (2015).

[6] V. Campos, B. Jou and X. Giró-i-Nieto, From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction, *Image and Vision Computing* **65** (2017), 15–22. https://doi.org/10.1016/j.imavis.2017.01.011.

[7] Q. You, J. Luo, H. Jin and J. Yang, Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia, in: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16, ACM Press, San Francisco*, California, USA, **2016** pp. 13–22. https://doi.org/10.1145/2835776.2835779.

[8] F. Chen, R. Ji, J. Su, D. Cao and Y. Gao, Predicting Microblog Sentiments via Weakly Supervised Multimodal Deep Learning, *IEEE Trans Multimedia* **20** (2018), 997–1007. https://doi.org/10.1109/TMM.2017.2757769.

[9] F. Huang, X. Zhang, Z. Zhao, J. Xu and Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, *Knowledge-Based Systems* **167** (2019), 26–37. https://doi.org/10.1016/j.knosys.2019.01.019.

[10] D. Borth, R. Ji, T. Chen, T. Breuel and S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: *Proceedings of the 21st ACM International Conference on Multimedia - MM '13, ACM Press*, Barcelona, Spain, **2013** pp. 223–232. https://doi.org/10.1145/2502081.2502282.

[11] T. Chen, D. Borth, T. Darrell and S.F. Change, Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586* (2014).

[12] J. Yuan, S. Mcdonough, Q. You and J. Luo, Sentribute: image sentiment analysis from a mid-level perspective, in: *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13, ACM Press, Chicago, Illinois*, **2013** pp. 1–8. https://doi.org/10.1145/2502069.2502079.

[13] M. Sun, J. Yang, K. Wang and H. Shen, Discovering affective regions in deep convolutional neural networks for visual sentiment prediction, in: *2016 IEEE International Conference on Multimedia and Expo (ICME), IEEE, Seattle, WA*, USA, **2016** 1–6. https://doi.org/10.1109/ICME.2016.7552961.

[14] J. Yang, D. She, M. Sun, M.-M. Cheng, P.L. Rosin and L. Wang, Visual Sentiment Prediction Based on Automatic Discovery of Affective Regions, *IEEE Trans. Multimedia* **20** (2018), 2513–2525. https://doi.org/10.1109/TMM.2018.2803520.

[15] E. Ko, C. Yoon and E.-Y. Kim, Discovering visual features for recognizing user's sentiments in social images, in: *2016 International Conference on Big Data and Smart Computing (BigComp)*, IEEE, Hong Kong, China, 2016: pp. 378-381. https://doi.org/10.1109/BIGCOMP.2016.7425952.

[16] T. Rao, M. Xu, H. Liu, J. Wang and I. Burnett, Multi-scale blocks based image emotion classification using multiple instance learning, in: *2016 IEEE International Conference on Image Processing (ICIP), IEEE, Phoenix*, AZ, USA, **2016** 634–638. https://doi.org/10.1109/ICIP.2016.7532434.

[17] T. Rao, X. Li, H. Zhang and M. Xu, Multi-level region-based Convolutional Neural Network for image emotion classification, *Neurocomputing* **333** (2019), 429–439. https://doi.org/10.1016/j.neucom.2018.12.053.

[18] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P.L. Rosin and L. Wang, WSCNet: Weakly Supervised Coupled Networks for Visual Sentiment Classification and Detection, *IEEE Trans Multimedia* (2019) 1–1. https://doi.org/10.1109/TMM.2019.2939744.

[19] X. He, H. Zhang, N. Li, L. Feng and F. Zheng, A Multi-Attentive Pyramidal Model for Visual Sentiment Analysis, in: *2019 International Joint Conference on Neural Networks (IJCNN), IEEE, Budapest, Hungary*, **2019** pp. 1–8. https://doi.org/10.1109/IJCNN.2019.8852317.

[20] Z. Wu, M. Meng and J. Wu, Visual Sentiment Prediction with Attribute Augmentation and Multi-attention Mechanism, *Neural Process Lett* **51** (2020), 2403–2416. https://doi.org/10.1007/s11063-020-10201-2.

[21] J. Ramon and L. De Raedt, Multi instance neural networks. In: *Proceedings of the ICML-2000 workshop on attribute-value and relational learning* (2000), p. 53–60.

[22] Z.H. Zhou and M.L. Zhang, Neural networks for multi-instance learning. In: *Proceedings of the International Conference on Intelligent Information Technology*, Beijing, China. (2002), p. 455–459.

[23] X. Wang, Y. Yan, P. Tang, X. Bai and W. Liu, Revisiting multiple instance neural networks, *Pattern Recognition* **74** (2018), 15–24. https://doi.org/10.1016/j.patcog.2017.08.026.

[24] M. Ilse, J.M. Tomczak and M. Welling, Attention-based deep multiple instance learning, *arXiv preprint arXiv:1802.04712*, (2018).

[25] Y. Lin, M. Moosaei and H. Yang, OutfitNet: Fashion Outfit Recommendation with Attention-Based Multiple Instance Learning, in: *Proceedings of The Web Conference 2020, ACM, Taipei Taiwan*, **2020** 77–87. https://doi.org/10.1145/3366423.3380096.

[26] Z.-H. Zhou and M.-L. Zhang, Solving multi-instance problems with classifier ensemble based on constructive clustering, *Knowl Inf Syst* **11** (2007), 155–170. https://doi.org/10.1007/s10115-006-0029-3.

[27] Y. Xu, J.-Y. Zhu, E.I.-C. Chang, M. Lai and Z. Tu, Weakly supervised histopathology cancer image segmentation and classification, *Medical Image Analysis* **18** (2014), 591–604. https://doi.org/10.1016/j.media.2014.01.010.

[28] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*, (2014).

[29] L. Song, J. Liu, B. Qian, M. Sun, K. Yang, M. Sun and S. Abbas, A Deep Multi-Modal CNN for Multi-Instance Multi-Label Image Classification, *IEEE Trans on Image Process* **27** (2018), 6025–6038. https://doi.org/10.1109/TIP.2018.2864920.

[30] X. Glorot, A. Bordes and Y. Bengio, Deep sparse rectifier neural networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence And Statistics* (2011), p. 315–323.

[31] J. Feng and Z.H. Zhou, Deep MIML network. In: *Thirty-First AAAI Conference on Artificial Intelligence*. (2017).

[32] S. Sabour, N. Frosst and G.E. Hinton, Dynamic routing between capsules. In: *Advances in Neural Information Processing Systems* (2017), 3856–3866.

[33] H. Ren and H. Lu, Compositional coding capsule network with k-means routing for text classification, *arXiv preprint arXiv:1810.09177*, (2018).

[34] Q. You, J. Luo, H. Jin and J. Yang, Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In: *Thirtieth AAAI Conference on Artificial Intelligence*. (2016).

[35] M. Katsurai and S. Satoh, Image sentiment analysis using latent correlations among visual, textual, and sentiment views, in: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Shanghai, **2016** 2837–2841. https://doi.org/10.1109/ICASSP.2016.7472195.

[36] K.-C. Peng, A. Sadovnik, A. Gallagher and T. Chen, Where do emotions come from? Predicting the Emotion Stimuli Map, in: *2016 IEEE International Conference on Image Processing (ICIP), IEEE, Phoenix, AZ*, USA, **2016** pp. 614–618. https://doi.org/10.1109/ICIP.2016.7532430.