IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 19, 2020, accepted August 30, 2020, date of publication September 18, 2020, date of current version October 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024948

Visual Sentiment Analysis With Active Learning

JIE CHEN[1], **QIRONG MAO**[1,2], **(Member, IEEE), AND LUOYANG XUE**[1]
[1]School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
[2]Jiangsu Engineering Research Center of Big Data Ubiquitous Perception and Intelligent Agriculture Applications, Zhenjiang 212013, China

Corresponding author: Qirong Mao (mao_qr@ujs.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 61672267 and Project U1836220, in part by the Qinglan Talent Program of Jiangsu Province, and in part by the Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace.

ABSTRACT Visual Sentiment Analysis (VSA) has attracted wide attention since more and more people are willing to express their emotion and opinions via visual contents on social media. Meanwhile, extensive datasets drive the rapid development of deep neural networks for this task. However, the annotation of large-scale datasets is very expensive and time consuming. In this paper, we propose a novel active learning framework, which uses few labeled training samples to achieve an effective sentiment analysis model. First, we attach a new branch to the traditional Convolution Neural Network (CNN), which is named "texture module". The affective vector will be obtained by computing inner products of feature maps from different convolutional blocks in this branch. We will utilize this vector to distinguish affective images. Second, the query strategy is formed by the classification scores from both the traditional CNN and the texture module. Then, we can use samples obtained by utilizing the query strategy to train our model. Extensive experiments on four public affective datasets show that our approach uses few labeled training samples to achieve promising results for VSA.

INDEX TERMS Visual sentiment analysis, active learning, convolutional neural network, texture information.

I. INTRODUCTION

Visual Sentiment Analysis (VSA) is aimed to subjectively describe images and enable computers to detect and express emotion. Since the number of active Internet users who express their emotion and opinions on the web is rapidly increasing worldwide, the automatic assessment of image has become a fashion trend and widely been used in many fields (e.g., education, entertainment and advertisement) [1], [2]. In the early years, many methods of hand-crafted features have been proposed, which focus on the low-level visual features (e.g., color, texture and shape) to predict sentiment. Recently, the effectiveness of machine learning based on deep features has been demonstrated over hand-crafted features by automatically learning effective feature representations from large amounts of training samples. More and more researchers apply numerous deep approaches to VSA [3].

For image classification task, Deep Neural Network (DNN) has demonstrated the ability of learning representative features by many literatures [4]. It can get better classification accuracy when using more annotated samples. However, the labeled samples are not available and enough for annotation labor and cost of time [5], which is a challenge in the VSA. In addition, image labeling is more difficult than traditional vision tasks for that the emotions are not absolutely independent in the aspect of semantic. On the contrary, those concrete objects (e.g., cat, dog and bird) are more independent [6].

To solve the above problems, we introduce the active learning method [7], which overcomes the labeling bottleneck by effectively selecting few labeled training samples and applying them to achieve high accuracy. Active learning has been developing for decades. Reference [8] first proposed a kind of active learning method by looking for the most uncertain samples. At present, there are few literatures about the introduction of active learning into CNN. Specifically, most of methods use general uncertainty sampling strategies for unlabeled samples which are hard to obtain with traditional CNN [9]. Given an unlabeled data pool, active learner evaluates all instances in this pool and selectively queries useful samples from the pool. There have been three major selection criteria: uncertainty sampling approach, diversity-based approach and expected model change [5]. The simplest and most commonly used query framework is uncertainty sampling. Meanwhile, the simplest method of the uncertainty sampling

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqing Zhang.

approach is to utilize class posterior probabilities [10] to find uncertainty.

In this paper, we propose a novel framework and a new uncertainty sampling strategy for VSA with active learning. Specifically, we attach the texture module to a traditional DNN, which will obtain emotion information to effectively distinguish affective images. While, the traditional CNN model is the one branch of our framework and the texture module is the other branch. We joint training these two branches to find the class posterior probabilities to define uncertainty. Then, the further study is facilitated by selecting samples according to the uncertainty. The illustration of proposed branches is shown in Fig. 1, including the traditional CNN, texture module, classifiers A and B.

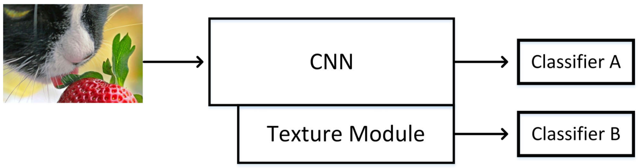


FIGURE 1. A new active learning method with a texture prediction module. The proposed texture prediction module is attached to the traditional CNN.

To our best knowledge, we are the first paper introducing active learning method to VSA. By the proposed method, we can achieve robust classification performance by using few labeled samples to reduce annotation burden. The main contributions of this paper are:

1) By attaching the texture module to the traditional CNN, we compute inner products of feature maps from multiple convolutional layers to obtain affective vectors. These vectors can effectively express emotion for our VSA task.
2) We design a new query strategy, which utilizes few labeled samples to train model. Specifically, the query strategy is formed by the classification scores from the traditional CNN and the texture module.

II. RELATED WORK

In this section, we review of related traditional methods for VSA, and then focus on active learning methods, as they are related to our work.

A. VISUAL SENTIMENT ANALYSIS

VSA is a challenging task, which makes its automated identification more tricky. In addition, because of the immense application potentials of VSA, numerous researchers focus on how to analysis the emotion in social media [11], [12]. At present, primary studies on VSA have mainly focused on visual cognition and machine learning.

Sentiment analysis based on visual cognition maps the low-level visual features of images to the high-level emotional semantics by applying the knowledge of visual cognition and psychology. Colombo *et al.* [13] propose a visual sentiment method utilizing the theory of art painting for art images. Wang *et al.* [14] empirically develop a systematic method based on psychology and cinematography to distinguish the semantic orientation of movies. In [15], they extract low-level image features to represent affective content, while these features mainly apply theoretical and empirical concepts from psychology and art. In [16], they survey and discuss the problems in aesthetics and emotion inference from images. Zhao *et al.* [17] extract principles-of-art-based emotion features to mine the emotion of images later.

Sentiment analysis based on machine learning extracts low-level features (e.g., color, texture and shape) from images, which is followed by training the classifier. Lu *et al.* [18] study the computability of emotion by investigating how shape characteristics in images acting on emotion of human beings. Early works use a variety of hand-crafted features to represent emotion evoked by images. Yanulevskaya *et al.* [19] design a emotional classification system by assessing local image statistics. Zhang *et al.* [20] leverage on Tamura texture and other features giving rise to image emotion classification. In [21], Gabor texture feature, HSV color histogram feature and bag of words on SIFT descriptor are all used to analyze visual sentiment.

In recent years, deep learning technology has made a breakthrough in VSA. More and more researchers have implemented the deep learning model to analyse visual sentiment. Chen *et al.* [22] present a method of visual sentiment classification based on CNN. In [23], they use the results of the models that have been trained by large-scale dataset (using ImageNet [24]) as initialization parameters, which can obtain much better performance than training from visual sentiment dataset alone. Next, fine-tuning the CNN for VSA, which is a current trend [25]–[27]. You *et al.* [28] utilize large-scale web samples to progressively learn CNN. Then they further utilize texts and images for multi-modality sentiment analysis on the Flickr and Instagram dataset [29]. The work in [30] applies weight multi-modal conditional probability neural networks to solve VSA task that is thought as a probability distribution problem. Wu *et al.* [31] propose an augmented CNN to improve sentiment classification result.

B. ACTIVE LEARNING

In supervised learning of a CNN, a large number of labeled samples are necessary, or the model will face generalization issues [32]. Recently, with the implementation using GPUs, deep learning approaches have made a huge success on VSA in large-scale annotated training samples [33]. The problem is that manual annotation of sentiment labels takes time and manpower. Meanwhile, the budget for annotation is limited [5]. Most of the existing emotion datasets that are no more than 1,000 images [2]. For the training CNN, this is far from the required scale. Because a limited financial budget is existing, or a few experts are available, techniques to reduce labeling effort become important. The goal of such algorithm is to find few labeled training samples to reach high classification performance. So we use the framework of active learning to solve this problem.

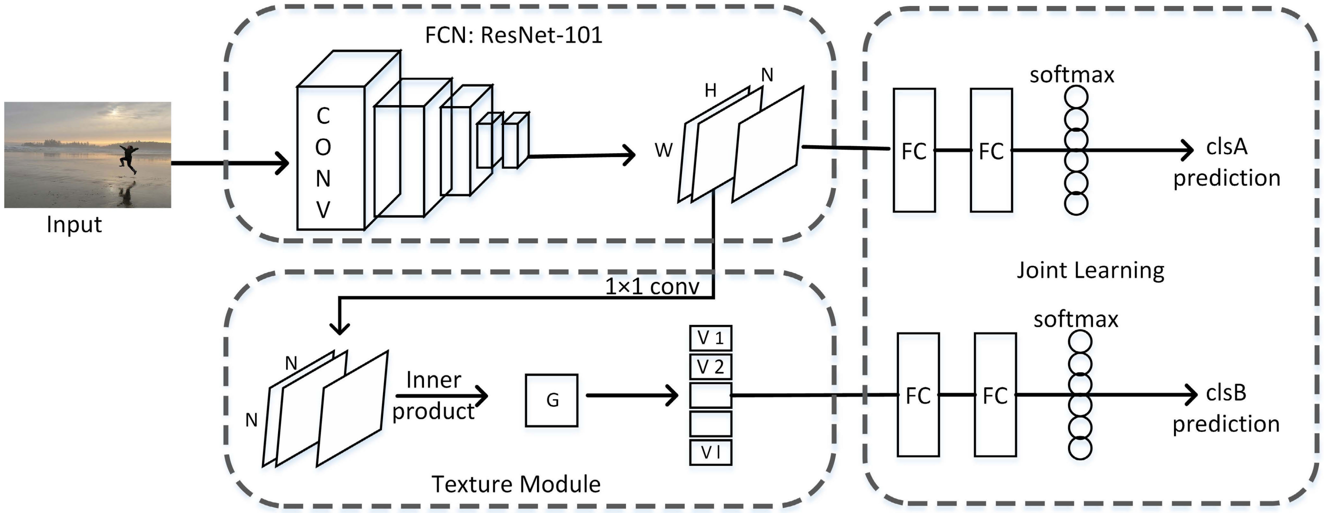


FIGURE 2. Illustration of the proposed algorithm. We first feed the input image into the ResNet-101, the response feature maps are then delivered into two classifiers. The texture module extracts texture information by the inner product operation. Our framework simultaneously optimizes the two classification losses (e.g., softmax).

Active learning is a subfield of machine learning, and it has been developing for decades. In [34], they combine Markov random field and active learning to study hyperspectral image classification. After that, Zhou *et al.* [35] compute entropy and diversity locally on a small number of patches in the unlabeled pool as evaluation criteria to determine which unlabeled images are worth to label. It effectively demonstrates that the cost of annotation can be reduced by at least half to solve the problem of lacking annotated datasets in biomedical imaging. Yang *et al.* [36] use a novel active selection strategy of uncertainty estimation and similarity estimation in the field of medical image segmentation. But there is currently little literature on active learning for CNN [9]. To the best of our knowledge, [10] first incorporates CNN into active learning framework and selects some uncertain images to label.

For active learning, the uncertainty-based selection [37] as one of the most common query strategy measures the uncertainties of unlabeled data. Inspired by the idea of many large collections of unlabeled samples can be gathered at once in the real world, the pool-based sampling [38] is proposed to assume that there is a small set of labeled data L and a large pool of unlabeled data U available. Then they select samples from the pool. There have been many proposed methods of how to formulate query strategies in the literature.

To our knowledge, we are the first to integrate active learning into VSA task. Our work focuses on achieving good classification performance by using few labeled samples.

III. VISUAL SENTIMENT ANALYSIS WITH ACTIVE LEARNING

In this section, we present our new method using active learning in VSA. The architecture is shown in Fig. 2, which has an input layer, one fully convolutional layer, one texture synthesis module, four fully connected layers and two classifiers. The goal is to select worthy samples to join the training process, and then achieve good classification performance. Specifically, the proposed framework learns the classification task jointly with two network branches. We use the traditional CNN branch and texture module branch to extract the high-level emotional semantics and low-level texture information respectively. Then, these information will be used to help the classifiers achieve better performance. Finally, we apply the classification scores through the two branches to form a new uncertainty calculation method.

A. TEXTURE MODULE FOR VISUAL FEATURE EXTRACTION

The texture information as one of the important low-level visual features plays an important role for image affective classification [15]. While, our proposed texture module is trained to generate the texture information from a given image.

The affective image was passed through the traditional CNN. Then, we will obtain a set of feature maps for each layer in the network. Taking l^{th} layer as an example, the output feature maps will be with pixel size H_l and W_l, and N_l channels. These feature maps are stored in the matrix $F^l \in R^{N_l \times H_l \times W_l}$, where F^l_{kj} is the activation value of the j^{th} feature map at the position k of layer l. Before stored, we apply the 1×1 convolutional layer to shrink the number of parameters [6] after the previous convolution operation. We mainly use layer-to-layer correlation to represent textures which stored in the Gram matrix $G^l \in R^{N_l \times N_l}$, where $G^l_{ij} = \sum_{k=1}^{H_l \times W_l} F^l_{ki} F^l_{kj}$ represents the inner product between filter i and j in layer l. Because of the symmetry of the matrix, the size of the Gram matrix can be shrunk again, where $V^l = \{G^l_{1,1}, G^l_{2,1}, G^l_{2,2}, \ldots, G^l_{N_l,1}, \ldots, G^l_{N_l,N_l}\}$ is the sentiment vector of layer l [6].

These vectors of each layer of the texture synthesis module are then stacked into one feature vector as

$V = \{V^1, V^2, \ldots, V^l\}$, which provides a stationary description of the texture.

B. UNCERTAINTY SAMPLING FOR AFFECTIVE SAMPLES SELECTION

There have been many types of common actively learning criteria can be used in our framework. In particular, the three uncertainty sampling criteria are:

Entropy: Entropy is a concept used to measure information content [39]. The higher entropy value, the more informative the sample will has, which sample can be as the uncertain sample. According to the entropy value in descending, K most uncertain samples that contain the most abundant information will be picked adding to labeled pool L. Given a set of samples $\{(x_i, y_i)\}_{i=1}^N$, here y_i is the corresponding sentiment label for i^{th} affective image x_i. The entropy is defined as:

$$En_i = -\sum_j p(y_i = j|x_i; \theta) \log q(y_i = j|x_i; \theta). \tag{1}$$

Least confidence: If the class label with the highest posterior probability is low [10], which class label for affective sample is uncertain for the classifier. $p(y_i = j|x_i; \theta)$ is the output of the network for the i^{th} example, which responses mapping relationship from input space to class confidence scores. The ranking of all the unlabeled samples is obtained via:

$$Lc_i = \max_j p(y_i = j|x_i; \theta). \tag{2}$$

Learn Loss: This study takes advantage of recent method to obtain uncertain samples from CNN: Loss Prediction (LP) module. The LP module is attached to the traditional CNN, which uses the representation of multi-level layer of the traditional CNN. After each of the active learning cycle, all the affective samples are evaluated in the unlabeled pool by the LP module to obtain data-loss pairs $\{(x_i, Ll_i)|x_i \in U\}$, then human oracles manually annotate the samples of the K-highest losses [5]:

$$Ll_i = \Theta loss(h), \tag{3}$$

where the feature set h will be obtained to compute the Ll_i. h can be obtained from all intermediate layers of the traditional CNN.

Our solution is to provide a more convenient uncertainty sampling strategy. The detailed steps are shown in Fig. 3 and Algorithm 1.

Given training sample x_i, we consider the classification probability $p(y_i = j|x_i; \theta)$ as the output of classifier A. This probability p is a ground-truth target of the proposed texture module. At the same time, we obtain another classification probability $q(y_i = j|x_i; \theta)$ through classifier B, which is defined as predicted probability. After getting the ground-truth and the predicted probability, we do not use the mean square error (MSE) which may be easy to consider. However, we propose a new choice that involves the concept of information content. Now, $-logq(y_i = j|x_i; \theta)$ represents the information content which is more emotional because of the texture module. Our new sampling strategy noted as "Ground-truth and Prediction" (GP) finds most emotional and informative samples by:

Algorithm 1 The Training Process of Our Proposed Framework

```
Input: Given the small labeled pool $L$, the large unlabeled
    pool $U$, uncertain samples number $K$, maximum training
    iteration number $T$, fine-tuning interval $t$.
while not reach maximum training iteration $T$ do
    Compute the classification probabilities via the traditional CNN and the texture module.
    Pick $K$ uncertainty samples from $U$ into $L$ in Eq.(4).
    In every $t$ iterations:
    Fine-tuning the network and update CNN parameter $\omega$.
end while
```

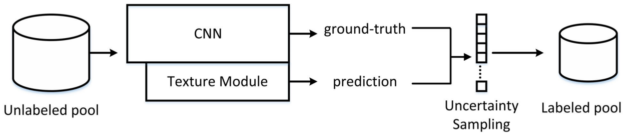


FIGURE 3. The proposed uncertainty sampling method. Given an input, the traditional CNN outputs the target prediction and the texture module outputs predicted value. The target prediction and the predicted value are jointly calculated to form a new uncertainty sampling method. According to the uncertainty sampling rule, we choose top-K samples to label and add them to the labeled pool.

$$Gp_i = -\sum_j p(y_i = j|x_i; \theta) \log q(y_i = j|x_i; \theta), \tag{4}$$

where probabilities of all class labels will be taken to measure the uncertainty. The higher G_{p_i} value, the more emotional informative the sample will has, which sample can be as the uncertain sample.

C. JOINT TRAINING PROCESS FOR VISUAL FEATURE LEARNING

In this section, we provide a detailed description of how to train the proposed framework, pick the most informative samples and ask human oracles to annotate them for the next stage of active learning.

We all know that there are a mass of CNN models developed in deep learning filed. The exact forms could vary, but the major components and computations are similar [40]. For our task, the input vectors of the two classifiers are different, which leads our framework to produce two outputs. In particular, classifier A can effectively learn the high-level semantic feature that is extracted from input sample step by step. Meanwhile, classifier B learns the specific texture information. Given the affective images and the labels, we mainly train the deep model by minimizing the joint loss to obtain the correct class as far as possible. Our loss function is integrated

(a) EmotionROI dataset

(b) Twitter I dataset

FIGURE 4. The first(a) and second(b) line respectively show partial images from Twitter I and EmotionROI dataset. They are collected from the social websites like Flickr and Twitter. Meanwhile, they also cover multiple domains such as art, life and abstract.

with two losses:

$$L = L_{clsA} + \lambda L_{clsB}, \tag{5}$$

where λ is a parameter taking the balance between L_{clsA} and L_{clsB}. The network parameters are learned by using adaptive moment estimation (Adam) to minimize the loss function.

D. THE PROCESS OF SENTIMENT ANALYSIS WITH ACTIVE LEARNING

Recently, active learning has been incorporated into deep CNN [10], which offers a promising way to get a good classification performance by using the least number of annotated training samples. The purpose of our framework for VSA is gradually selecting the most useful samples for model updating in each of the iterative process.

All the affective training examples constitute a large unlabeled samples U that are easily obtained. We denote unlabeled pool $U = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{1, \ldots, C\}$ is the corresponding sentiment label of the affective image x_i. The subscript N means the number of affective samples. At first, we randomly pick K affective examples from the unlabeled pool and annotate them, which will constitute an initial small labeled pool L. Then the unlabeled pool will became U_{N-K}. After that, we enter the next iteration: evaluating all unlabeled samples in U_{N-K} according to our new uncertainty calculation method. Those most uncertain samples will be manually annotated and added to L. This process is repeated. Then, the labeled dataset L will be consistently growing. Once the labeled pool is obtained, we retrain the model with the selected and the initial labeled samples, until a satisfactory model performance is reached or the budget of annotation is exhausted.

IV. EXPERIMENTS

In this section, we apply our method to VSA. Compared with several baseline methods, our framework achieves better performance.

We have implemented our method for VSA with PyTorch. Then datasets, implementation details, experimental settings and performance evaluation are described in the following sections.

A. DATASETS

We perform our active learning framework on four popular affective datasets, including Twitter I [28], EmotionROI [41], the Flickr and Instagram (FI) [29] and Instagram [42] dataset.

Twitter I: The Twitter I dataset is the most popular image sentiment benchmark with 1,269 images collected from tweets [43]. Each image was labeled with a sentiment label (e.g., positive and negative) by five Amazon Mechanical Turk (AMT) works. There are three different opinions, including "Five agree", "At least four agree" and "At least three agree". In particular, "Five agree" means that all the five AMT participants give the same sentiment label for a image [44]. We will use this annotation result in the next experiment.

EmotionROI: The EmotionROI dataset consists of 1,980 images with six sentiment categories (e.g., anger, disgust, fear, joy, sadness and surprise). Fig. 4 shows examples from Twitter I and EmotionROI dataset.

FI: The FI dataset is a public dataset with 23,308 affective images crawled from social websites. There are 225 AMT workers asked to generate sentiment label (i.e. anger, amusement, awe, contentment, disgust, excitement, fear and sadness). In our paper, we just get partially valid 21,829 images.

Instagram: The Instagram dataset contains 42,856 images from Instagram, including two sentiment categories (i.e. positive and negative).

B. IMPLEMENTATION DETAILS

Target model: We employ ResNet-101 [45] as the traditional CNN, which has 101-layer residual network. Meanwhile, we apply the pre-trained weights of the large-scale hierarchical image database (e.g., ImageNet) [46] to initialize our network. The softmax layer is added on the top of the convolutional layers which changed to the corresponding number of categories.

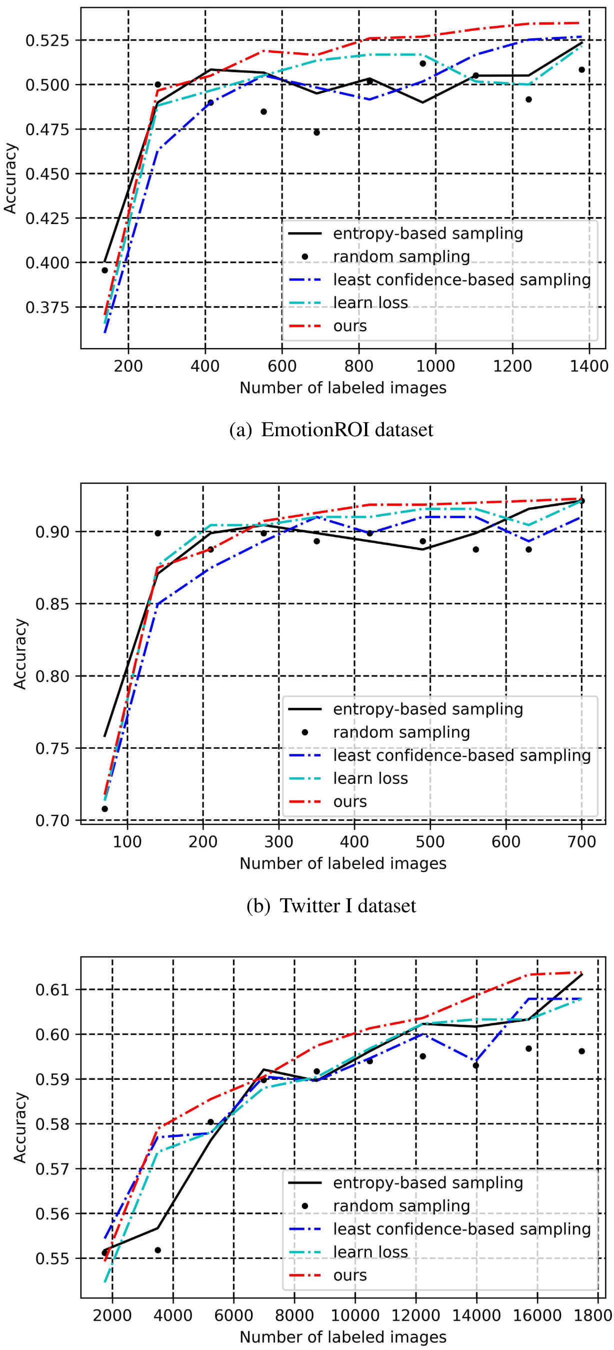
Texture module: There are 4 basic blocks for ResNet-101 after the first convolution layer. Each block respectively comprises three, four, twenty three and three convolution layers. Starting from the first convolutional block, the number of channels in the outputs of successive blocks gradually increase as 256 → 512 → 1024 → 2048 [47]. This design makes the net to learn richer and high-level semantic feature. The proposed texture module is connected to each of the basic blocks to utilize the four rich features from the blocks [5]. The feature maps are calculated via inner product operation to learn low-level texture feature. Then, we use the sentiment vector to denote the correlation between each pair of feature maps. Taking the output of conv5-3 layer as an example, after using 16 filters with kernel size of 1×1, the sentiment vector is 136. Finally, by stacking vectors of each layer of the texture module, we can obtain a total of 680-dimensional features.

C. LEARNING

For training, the size of original images is too large to compute efficiently, and hence all samples are resized to 224×224 [48]. Meanwhile, we utilize the channel mean and standard deviation vectors estimated to normalize the image. For every active learning cycle, we jointly learn with the two branches (the traditional CNN and texture module). The learning rate is initialized as 10^{-4}. A weight decay of 5×10^{-4} with a momentum of 0.9 is used in our method. Due to the limit of GPU memory and the large image size, it is necessary to reduce the batch size for the iterations [49]. We fine-tune all layers by Adam through the whole net to a mini-batch size of 8 in each iteration. The total number of iterations is 30 epochs. The datasets are randomly split into 80% training and 20% testing sets. Among the training set, we randomly select 10% samples to initialize the network and the rest are the unlabeled datasets for the subsequent learning process. We average 6 trial results as the final result, and this help us get rid of the influence of randomness that will lead to unreliable experimental results.

D. EVALUATION SETTING

We compare our method with random sampling (RS), entropy-based sampling (ES), least confidence-based sampling (LCS) and learn loss (LL) method [5] to prove the effectiveness of our framework with few labeled samples for VSA. For the RS method, we randomly select affective samples to be annotated which will be used to fine-tune the deep network architecture. This method do not need any active learning techniques and can be thought as the lower bound [10]. For the ES method, we calculate the entropy values of an image which from softmax activity function outputs. The simple ES method works very effectively for the classification task which is typically learned to minimize cross-entropy between predictions and target labels. For LCS method, we mainly consider the sample that the classification probability of the most probable class is low, then this sample will be annotated. Moreover, we also evaluate our method against the LL method. For LL, to the best of our knowledge, it is the recent state-of-the-art approach for active learning. During the experiment, we use the same hyper-parameters.



(a) EmotionROI dataset

(b) Twitter I dataset

(c) FI dataset

FIGURE 5. Comparison between different uncertainty sampling strategy on (a) EmotionROI, (b) Twitter I and (c) FI dataset.

E. PERFORMANCE EVALUATION

We evaluate the proposed method on four public emotional datasets, including Twitter I, EmotionROI, FI and Instagram dataset.

1) ACCURACY ON PUBLIC EMOTIONAL DATABASES

We set the hyper-parameter $\lambda = 0.5$ in the proposed method, which is the weight to control the tradeoff between the losses

TABLE 1. Classification accuracy (%) on the test set of four datasets: EmotionROI, Twitter I, FI and Instagram. Here, '√' denotes using different CNNs structure (e.g., GAP+FC and texture synthesis).

Dataset	+GAP+FC	+ Texture Synthesis	Acc.(%)
EmotionROI	√		53.03
		√	**55.72**
Twitter I	√		92.70
		√	**93.26**
FI	√		60.06
		√	**61.50**
Instagram	√		83.15
		√	**84.02**

from the two classifiers. Specifically, λ indicates the weight of texture module in the optimization objective function. In order to inspect its influence on the classification accuracy, we test different threshold values, which are performed in the range of $[0.1, 0.9]$ on four datasets. We greedily search the parameter values according to the classification performance through an incremental iterative process [50], [51]. Our method achieves the best performance with $\lambda = 0.5$, and it becomes worse when the threshold is larger or smaller. We can conclude that a well-designed threshold can effectively mine minority informative samples and improve the performance for our task. Table 1 shows the classification accuracy when we trained with different network structures. Compared with using the traditional classification framework (eg., +GAP+FC), results of using texture synthesis module are more effective for the emotional classification. Moreover, the result of our proposed framework using texture synthesis outperforms the traditional CNN model by 2.69% on EmotionROI dataset and achieves 55.72%. Texture synthesis method for FI dataset reaches 61.50%, making improvement over the traditional framework (+GAP+FC) by 1.44%. On Instagram dataset and Twitter I dataset, our method outperforms the traditional CNN model by 0.97% and 0.56%, respectively. This illustrates that utilizing the texture information from the CNN can be more discriminative, and the texture information will be as one of the important low-level visual features for visual sentiment categorization.

2) ACTIVE LEARNING RESULTS FOR VSA

As shown in Table 2, Table 3 and Table 4, our proposed method is better adapted to VSA. It exceeds the compared methods to performing the highest classification accuracy when given the same percentage of annotated samples in almost every active learning cycle. To justify the effectiveness and the consistently good performance of this method, we implement 6 trials for every compared methods. Each result that shown in Table 2, Table 3 and Table 4 is an average of 6 trials. We compare our method with RS, ES, LCS and LL method [5] as uncertainty sampling strategy to assess the effectiveness of ours. In particular, most of the compared methods have better results than the RS strategy [5]. In the last cycle, RS, ES, LCS and LL method show 50.84%, 52.35%, 52.69% and 52.17% respectively, while our method shows 53.46% on EmotionROI dataset. This is 0.77% higher than the best compared result. From the aspect of the sample annotation amount, to achieve 52.60% recognition accuracy by using 60% labeled samples for ours, ES, LCS and LL method require almost 100% labeled samples. Meanwhile, for Twitter I dataset, to achieve the 92.13% accuracy, RS, ES and LL method require 100% labeled samples. Ours needs only 90% labeled samples. On FI dataset, the results demonstrate that giving the same percentage of labeled samples and compared with RS, ES, LCS and LL method, our method obtains better performance. This justifies that our proposed sampling strategy can effectively get promising performance with few labeled samples.

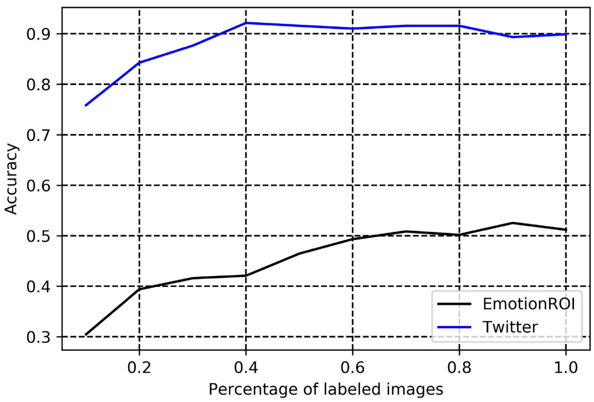


FIGURE 6. Accuracy of the texture module.

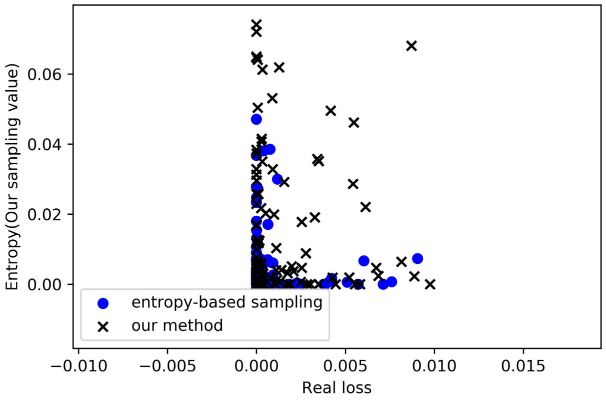


FIGURE 7. Data visualization on the testing set of the Twitter I. The blue and black colors represent different sentiment samples. As we can see, our method can obtain higher real loss values.

As illustrated in Fig. 5, these curves demonstrate the results of classification under different percentages of annotated samples of different uncertainty sampling strategies (e.g., RS, ES, LCS and LL method) on EmotionROI [41], Twitter I and FI dataset in the whole training process. Giving the same percentage of labeled samples, our method shows the best results in almost all active learning cycles. This validates that our proposed sample strategy can further improve the classification accuracy and select informative samples for feature learning in every steps.

Fig. 6 shows the accuracy of the texture module for the test set. As more labeled samples are added to the training

TABLE 2. Classification performances under different percentages of annotated samples on the EmotionROI dataset. The recognition accuracy is reported in %..

Percentage of labeled samples	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
random sampling	39.56	**50.00**	48.99	48.48	47.31	50.17	51.19	50.51	49.17	50.84
entropy-based sampling	**40.07**	48.99	**50.84**	50.67	49.50	50.34	48.99	50.50	50.51	52.35
least confidence-based sampling	36.03	46.30	48.99	50.51	49.83	49.16	50.17	51.68	52.52	52.69
learn loss	36.56	48.82	49.66	50.51	51.35	51.68	51.68	50.17	50.00	52.17
ours	37.04	49.66	50.50	**51.89**	**51.66**	**52.60**	**52.69**	**53.11**	**53.42**	**53.46**

TABLE 3. Classification performances under different percentages of annotated samples on the Twitter I dataset. The recognition accuracy is reported in %.

Percentage of labeled samples	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
random sampling	70.79	**89.89**	88.76	89.89	89.33	89.89	89.33	88.76	88.76	92.13
entropy-based sampling	**75.84**	87.07	89.89	90.45	89.89	89.33	88.76	89.89	91.57	92.13
least confidence-based sampling	71.35	84.97	87.48	89.33	91.01	89.89	91.01	91.01	89.33	91.01
learn loss	71.35	87.64	**90.45**	90.45	91.01	91.01	91.57	91.57	90.45	92.13
ours	71.77	87.50	88.78	**90.73**	**91.30**	**91.86**	**91.86**	**92.00**	**92.13**	**92.28**

TABLE 4. Classification performances under different percentages of annotated samples on the FI dataset. The recognition accuracy is reported in %..

Percentage of labeled samples	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
random sampling	55.12	55.18	58.04	58.98	59.17	59.40	59.51	59.30	59.68	59.62
entropy-based sampling	55.18	55.67	57.63	**59.21**	58.96	59.62	60.23	60.17	60.33	61.33
least confidence-based sampling	**55.44**	57.70	57.79	59.05	58.96	59.46	60.00	59.40	60.79	60.79
learn loss	54.46	57.37	57.81	58.80	59.04	59.68	60.23	60.33	60.33	60.79
ours	54.93	**57.89**	**58.55**	59.05	**59.74**	**60.13**	**60.36**	**60.86**	**61.63**	**61.38**

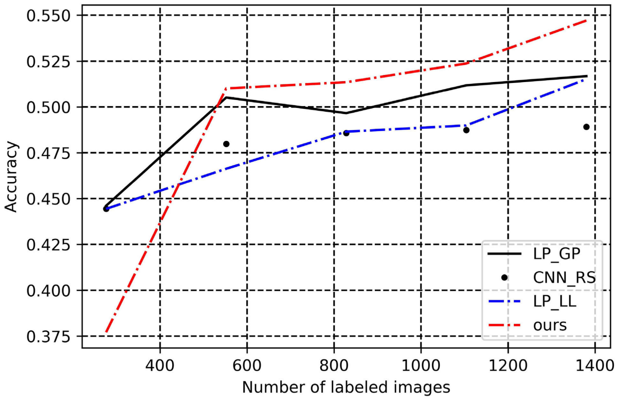


FIGURE 8. Image classification results under different percentages of annotated samples of the whole training process on EmotionROI. Our proposed method performs better than the compared ones.

process, the accurate of texture module gradually increases, which effectively proves that the texture module can get the classification scores like the traditional CNN. Then, these scores will help us form the novel sampling strategy.

In Fig. 7, all of the points are selected by our method or entropy from the last active learning cycle. Our method has higher real loss values than the ES method, which effectively illustrates that our method can obtain more informative samples to further learning emotional features for VSA task.

3) COMPARISON ANALYSIS

To comprehensively demonstrates that our proposed framework and uncertainty sampling strategy can improve the classification performance with few labeled samples, we compare our method with CNN_RS, LP_LL and LP_GP. For CNN_RS, it is the baseline method that we randomly select samples to be annotated to fine-tune the CNN. LP_LL is a new state-of-the-art active learning approach and has overcome the state-of-the-art methods with much less annotations, which using the state-of-the-art deep network model that researchers attach the LP module to the traditional CNN, and new sampling strategy that called LL. LP_GP is applied with the LP framework and GP that we proposed. Fig. 8 shows the results. Our method needs only 57% (nearly 800) labeled samples to reach the better performance for LP_LL and LP_GP. This justifies the effectiveness of our framework and the proposed sampling method.

V. CONCLUSION

In this paper, we incorporate deep CNN into our proposed new active learning method for VSA. The proposed novel method employs a new sample strategy: progressively select part of the most informative affective data by calculating the classification probability values of the outputs of two classifiers. Experimental results justify the effectiveness of our method on four public affective datasets. In future works, we will apply our framework on more challenging sentiment classification tasks, and continue this research to structure a better architecture and uncertainty sampling method to increase the classification accuracy.

REFERENCES

[1] Y. Zhao, B. Qin, T. Liu, and D. Tang, "Social sentiment sensor: A visualization system for topic detection and topic sentiment analysis on microblog," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 8843–8860, Aug. 2016.

[2] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7584–7592.

[3] K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 860–868.

[4] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3034–3043.

[5] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 93–102.

[6] J. Yang, D. She, Y. Lai, and M. Yang, "Retrieving and classifying affective images via deep metric learning," in *Proc. 32nd AAAI Conf. Artif. Intell., (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LI, USA, Feb. 2018, pp. 491–498.

[7] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 2372–2379.

[8] L. E. Atlas, D. A. Cohn, and R. E. Ladner, "Training connectionist networks with queries and selective sampling," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Denver, CO, USA, Nov. 1989, pp. 566–573.

[9] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler, "The power of ensembles for active learning in image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 9368–9377.

[10] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[11] R. Zheng, W. Li, and Y. Wang, "Visual sentiment analysis by leveraging local regions and human faces," in *Proc. 26th Int. Conf. MultiMedia Modeling (MMM)*, in Lecture Notes in Computer Science, Daejeon, South Korea, vol. 11961, Y. M. Ro, W. Cheng, J. Kim, W. Chu, P. Cui, J. Choi, M. Hu, and W. D. Neve, Eds. Springer, Jan. 2020, pp. 303–314.

[12] D. Cao, R. Ji, D. Lin, and S. Li, "Visual sentiment topic model based microblog image sentiment analysis," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 8955–8968, Aug. 2016.

[13] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in visual information retrieval," *IEEE Multimedia Mag.*, vol. 6, no. 3, pp. 38–53, Jul./Sep. 1999.

[14] H. Lin Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[15] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. Int. Conf. Multimedia (MM)*, Firenze, Italy, 2010, pp. 83–92.

[16] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, Sep. 2011.

[17] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring Principles-of-Art features for image emotion recognition," in *Proc. ACM Int. Conf. Multimedia (MM)*, Orlando, FL, USA, 2014, pp. 47–56.

[18] X. Lu, P. Suryanarayan, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, Nara, Japan, 2012, pp. 229–238.

[19] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek, "Emotional valence categorization using holistic image features," in *Proc. 15th IEEE Int. Conf. Image Process.*, San Diego, CA, USA, Oct. 2008, pp. 101–104.

[20] H. Zhang, E. Augilius, T. Honkela, J. Laaksonen, H. Gamper, and H. Alene, "Analyzing emotional semantics of abstract art using low-level image features," in *Proc. 10th Int. Symp. Adv. Intell. Data Anal. (IDA)*, in Lecture Notes in Computer Science, Porto, Portugal, vol. 7014, J. Gama, E. Bradley, and J. Hollmén, Eds. Springer, Oct. 2011, pp. 413–423.

[21] B. Li, S. Feng, W. Xiong, and W. Hu, "Scaring or pleasing: Exploit emotional impact of an image," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, Nara, Japan, Oct./Nov. 2012, pp. 1365–1366.

[22] T. Chen, D. Borth, T. Darrell, and S. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," 2014, *arXiv:1410.8586*. [Online]. Available: https://arxiv.org/abs/1410.8586

[23] Q. You, J. Luo, H. Jin, and J. Yang, "Joint visual-textual sentiment analysis with deep neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, Brisbane, NSW, Australia, 2015, pp. 1071–1074.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, in Lecture Notes in Computer Science, Munich, Germany, vol. 9351, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds. Springer, Oct. 2015, pp. 234–241.

[25] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 91–99.

[26] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.

[28] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, Austin, Tx, USA, Jan. 2015, pp. 381–388.

[29] 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 *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 308–314.

[30] S. Zhao, G. Ding, Y. Gao, and J. Han, "Learning visual emotion distributions via multi-modal features fusion," in *Proc. ACM Multimedia Conf. (MM)*, Mountain View, CA, USA, 2017, pp. 369–377.

[31] L. Wu, S. Liu, M. Jian, J. Luo, X. Zhang, and M. Qi, "Reducing noisy labels in weakly labeled data for visual sentiment analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 1322–1326.

[32] S. Chen, C. Zhang, and M. Dong, "Coupled End-to-End transfer learning with generalized Fisher information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4329–4338.

[33] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 742–751.

[34] S. Sun, P. Zhong, H. Xiao, and R. Wang, "An MRF model-based active learning framework for the spectral-spatial classification of hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1074–1088, Sep. 2015.

[35] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4761–4772.

[36] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Proc. 20th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, in Lecture Notes in Computer Science, Quebec City, QC, Canada, vol. 10435, M. Descoteaux, L. Maier-Hein, A. M. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Springer, Sep. 2017, pp. 399–407.

[37] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1995, vol. 29, no. 2, pp. 13–19.

[38] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Dublin, Ireland, Jul. 1994, pp. 3–12.

[39] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[40] S. Chen, C. Zhang, and M. Dong, "Deep age estimation: From classification to ranking," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2209–2222, Aug. 2018.

[41] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? Predicting the emotion stimuli map," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 614–618.

[42] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 2837–2841.

[43] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, Oct. 2018.

[44] 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*, vol. 20, no. 9, pp. 2513–2525, Sep. 2018.

[45] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ''ImageNet: A large-scale hierarchical image database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, Florida, USA, Jun. 2009, pp. 248–255.

[47] W. Wang and J. Shen, ''Deep visual attention prediction,'' *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[48] S. Zeng, J. Gou, and L. Deng, ''An antinoise sparse representation method for robust face recognition via joint l_1 and l_2 regularization,'' *Expert Syst. Appl.*, vol. 82, pp. 1–9, Oct. 2017.

[49] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, ''ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3462–3471.

[50] J. Gou, Z. Yi, D. Zhang, Y. Zhan, X. Shen, and L. Du, ''Sparsity and geometry preserving graph embedding for dimensionality reduction,'' *IEEE Access*, vol. 6, pp. 75748–75766, 2018.

[51] X.-J. Shen, S.-X. Liu, B.-K. Bao, C.-H. Pan, Z.-J. Zha, and J. Fan, ''A generalized least-squares approach regularized with graph embedding for dimensionality reduction,'' *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107023.

JIE CHEN received the B.S. degree in digital media technology from Yuncheng University, Yuncheng, China, in 2018. She is currently pursuing the M.S. degree in computer application technology with the School of Computer Science and Communication Engineering, Jiangsu University. Her research interest includes computer vision.

QIRONG MAO (Member, IEEE) received the M.S. and Ph.D. degrees in computer application technology from Jiangsu University, Zhenjiang, China, in 2002 and 2009, respectively. She is currently a Professor with the School of Computer Science and Communication Engineering, Jiangsu University. She has published over 40 technical articles, some of them in premium journals and conferences, such as the IEEE TRANSACTION ON MULTIMEDIA, ACM Multimedia, IEEE CVPR, and the IEEE TRANSACTION ON IMAGE PROCESSING. Her research interests include affective computing, pattern recognition, and multimedia analysis. Her research is supported by the key project of the National Science Foundation of China (NSFC), Jiangsu province, and the Education Department of Jiangsu Province.

LUOYANG XUE received the B.S. degree in computer science and technology from Jiangsu University, Zhenjiang, China, in 2017. He is currently pursuing the M.S. degree in computer application technology with the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include affect computing and pattern recognition.