Zero-Shot Emotion Recognition via Affective Structural Embedding

Anonymous ICCV submission

Paper ID 4618

Abstract

Image emotion recognition attracts much attention in recent years due to its wide applications. It aims to understand the emotional response of humans, where candidate emotion categories are generally defined by specific psychological theories. However, with the development of psychological theories, emotion categories become increasingly diverse, fine-grained, and difficult to collect samples. In this paper, we investigate zero-shot learning (ZSL) problem in the emotion recognition task, which aims to recognize the new unseen emotions. Specifically, we propose an affective structural embedding framework, utilizing mid-level semantic representation, i.e.*, adjective-noun pairs (ANP) features, to construct an intermediate embedding space. By doing this, the learned intermediate space can bridge the affective gap between low-level visual features and high-level semantics. In addition, we introduce an adversarial constraint to combine the visual and affective embeddings so as to retain the discriminative capacity of visual features and the affective structural information of semantic features during training process. Our method is evaluated on five widely-used affective datasets and the experimental results show the proposed algorithm outperforms the state-of-the-art approaches. Our source code and trained models will be released.*

1. Introduction

With the rapid development of social media, more and more people prefer to record their lives and express opinions via visual contents, *e.g.*, images and videos [45]. In particular, computational understanding emotions of online images has attracted increasing attention from academia and industry due to its various applications, *e.g.*, opinion mining [29], online advertisement [14] and social networks [17].

In the past few years, many methods [37, 38, 42] have made huge progress for image emotion recognition, which aims to classify emotions evoked by image content. Most existing methods follow the general view in psychology that a specific emotion can be recognized as a fixed number of basic emotions. For example, Peng *et al.* [27] train regression models to predict the probabilities of Ekman's six basic emotions [11, 10], including happiness, sadness, disgust, anger, fear, and surprise. There are also methods employing different psychological theory for emotion modeling. Yang *et al.* [39] jointly optimize emotion classification and distribution learning task according to Mikels' wheel [23], which replace happiness and surprise in Ekman's basic emotions with amusement, content, awe, and excitement. However, with the development of psychological theories, basic emotion categories become increasingly fine-grained. Traditional supervised learning methods can only recognize the seen classes, *e.g.*, happiness, anger and other four emotions on the manifold as shown in Fig. 1. Such recognition model trained on the pre-defined categories cannot recognize emotions dynamically, when new categories are explored according to different psychological theories. In addition, it is labor-intensive and time-consuming to collect samples for rare emotion categories.

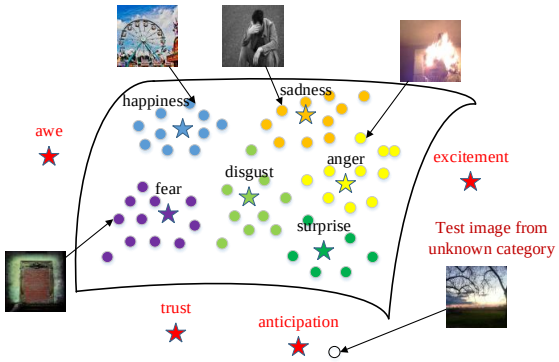


Figure 1. Overview of zero-shot emotion recognition problem. Each class has a prototype (denoted by a star). Training classes are on the manifold with different colors. Test unseen classes are in red and side information is provided to determine where an unseen classes locates. To classify an unseen image, we assign it a label that corresponds to the nearest unseen prototypes.

Zero-shot learning (ZSL) [36] aims to recognize new categories that are not exists in the training set, which has been widely used in various vision tasks [9, 33, 47]. The conven-

tional zero-shot learning methods usually build a common space based on the correspondence between the seen images and their class semantic representations. The space are also shared by both seen and unseen classes, which rely on the side information (*e.g*. attributes and Word2vec) about how unseen classes are semantically related to the seen classes. Then zero-shot learning can be simplified into a nearest neighbor search task, and test images will be assigned to the nearest unseen class in the common embedding space. Such ZSL paradigm relies on the cross-modality similarity between visual features and class semantic representations. There exists an affective gap between low-level image features and high-level emotional semantics [22] and directly computing similarities is hard to describe the similarity relationship correctly between them. Thus, zero-shot emotion recognition becomes more challenging.

In this paper, we propose an affective structural embedding framework using the mid-level semantic representations, *i.e*. adjective-noun pairs (ANP) [6] features, to construct an intermediate embedding space. Both visual and class semantic features are embedded into the learned embedding space and aligned with the affective structure of the ANP features. Therefore, our method can effectively bridge the affective gap and address the zero-shot and generalized zero-shot learning problems in emotion recognition. Note that in the zero-shot setting, training and test classes are disjoint and in the more realistic generalized zero-shot setting, training classes are present at test time. In the training process, both visual embedding and affective embedding are dynamically changing, and it is difficult to combine them directly and effectively. We further introduce an affective adversarial constraint to force the visual embedding to choose an embedding space that preserves the affective structural information.

Our contributions are summarized as follows: First, we propose an end-to-end affective structural embedding framework to learn an intermediate space and preserve emotion-related information, in which both visual and class semantic features are learned. To the best of our knowledge, this is the first zero-shot learning work on image emotion recognition. Second, we apply an affective adversarial constraint to retain the discriminative capacity of visual features and the affective structural information of semantic features during training process. The experimental results show the superiority of the proposed method over the state-of-the-art methods on five datasets.

2. Related Work

2.1. Image Emotion Recognition

Previous approaches for image emotion recognition mainly focus on the classification problem utilizing hand-crafted features or deep learning features. In the early years, many methods design hand-crafted features with different levels to recognize image emotion. For low-level features, Machajdik *et al.* [22] define a combination of hand-crafted features according to aesthetics and psychology theory, including color, texture, and composition. Zhao *et al.* [46] further investigate more robust visual features related to art principles as the mid-level representation. In another research [7], adjective-none pairs are regarded as mid-level semantic features, and a bank of visual sentiment classifiers (SentiBank) is proposed for image affective analysis.

Recently, Convolutional Neural Network (CNN) has been applied to image emotion recognition tasks and achieved satisfactory results. Inspired by the research [7], DeepSentiBank [8] adopts a deep CNN model to construct a detector for visual sentiment concept based on the adjective-noun pairs. You *et al.* [41] propose a novel progressive CNN architecture PCNN, to make use of large noisy web data for binary sentiment classification. Yang *et al.* [37] explore the relation between emotions via deep metric learning and employ a multi-task framework to optimize retrieval and classification simultaneously. Later several methods [40, 38] consider both global and local information for image emotion recognition.

All the above methods employ a supervised manner to learn the relationship between image visual content and emotions, which depends on the pre-defined psychological theories. In addition, many recent works [5, 21] suggest the types of emotions are much more various than previously assumed. Because of the diversity of emotional descriptions, it is difficult to assign an emotional image to an existing stereotypical label practically. The focus of our research is to classify a novel emotion class which does not appear in the training set.

2.2. Zero-Shot Learning

Zero-shot learning aims to classify classes without any training data. To cope with the challenge, most works [12, 44, 1, 34] utilize semantic attributes describing cross-class properties to transfer the semantic knowledge from the seen classes to the novel unseen classes. However, semantic attributes need manual definition and annotation, which limits the scalability of the above approaches. Several works [13, 2, 30] explore zero-shot learning using word vector representations [24], which is constructed by the large-scale text corpora in an unsupervised way. For the affective datasets without attribute annotations, we choose word vector representations as the class semantic features.

Many zero-shot learning works using the embedding-based method seek to measure the similarities between the visual features and class semantic features in different embedding space. DeViSE [13] directly learns a linear mapping from the image space to the semantic space using a ranking loss function. SJE [2] optimizes the structural SVM

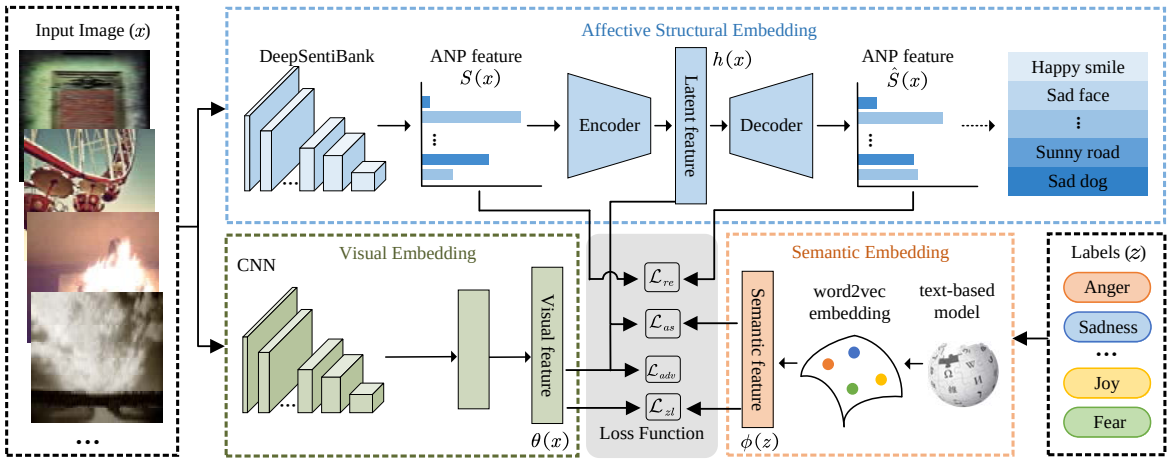


Figure 2. Pipeline of the proposed approach for zero-shot emotion recognition. Given the training image, we first extract the ANP features using the pre-trained DeepSentiBank detector and feed them into an auto-encoder to conduct the latent ANP space. Meanwhile, visual features are also embedded in the latent ANP space to align with the embedded semantic features and measure the similarities for zero-shot emotion recognition. The whole framework is trained by optimizing the multi-loss function in an end-to-end manner.

loss function to learn the bilinear compatibility between visual and semantic space. SAE [20] proposes a semantic autoencoder to regularize the model. It firstly projects the image features to the semantic space and further reconstructs them back to the visual space. DEM [43] chooses to embed the semantic features to the visual space. PSR [3] further considers the inter-class semantic relationships during the mapping process. Moreover, many zero-shot learning approaches [18, 19, 31] learn to embed both visual and semantic features into a latent intermediate space.

However, all the above ZSL methods fail to capture the specific emotion information for emotion recognition problem. The visual and class semantic features are located in different structural spaces, both of which are independent of emotions. Our model utilizes the mid-level semantic representations to construct an intermediate space. It can reserve emotion-related information and effectively bridge the affective gap.

2.3. Adversarial Methods for ZSL

There are a few recent works that tackle the zero-shot learning problem utilizing adversarial learning methods and generative adversarial networks (GAN). GAZSL [48] leverages GANs to imagine the visual features given the noisy textual descriptions from Wikipedia. CVAE-ZSL [26] proposes to use conditional variational autoencoder to generate samples for unseen classes. f-CLSWGAN [35] applies GAN to generate image features conditional on class attributes. The idea of GAN and adversarial learning methods are to train a generator that can fool a discriminator to confuse the distributions of the generated and true samples. The max-min training procedure can lead the generator to model the data distribution. Our method is similar to the GAN applied in the feature level. In this paper, we employ adversarial learning to bridge the gap between visual and affective features.

3. Methodology

In this section, we first formalize the zero-shot emotion task and then introduce the proposed affective structural embedding model. As shown in Fig. 2, we propose an independent affective structural embedding with traditional visual-semantic embedding. Specifically, the extracted ANP features are fed into an auto-encoder to learn the latent ANP space, and then both visual and class semantic features are embedded into the learned ANP space so as to effectively bridge the affective gap. In addition, we introduce an affective adversarial constraint to effectively combine visual and ANP features so as to retain the discriminative capacity and the affective structural information.

3.1. Problem Definition

Following conventional zero-shot learning problem, we split the affective dataset with s seen classes and u unseen classes. The training set is defined as $D_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, where $x_i^s \in \mathcal{X}_S$ denotes the i-th image of the seen class and $y_i^s \in \mathcal{Y}_S$ is the corresponding class label. On the other hand, we define the test set as $D_U = \{(x_j^u, y_j^u)\}_{j=1}^{n_u}$, where $x_j^u \in \mathcal{X}_\mathcal{U}$ denotes the j-th unseen image and $y_j^u \in \mathcal{Y}_\mathcal{U}$ is the class label. The seen and unseen classes are disjoint, *i.e.* $\mathcal{Y}_\mathcal{S} \cap \mathcal{Y}_\mathcal{U} = \emptyset$. Additionally, we choose the word vector z_i^s and z_j^u obtained by the NLP model [28] as the class semantic features. Note that during the training stage, only seen class images are used to learn the classifier model with the assistance of semantic information z^s. Given a test image x^u and the semantic feature z^u, we aim to predict the

corresponding class y^u.

3.2. Affective Structural Alignment

ZSL problems are usually addressed by measuring similarities between the visual and class semantic features. Since affective datasets, *e.g.*, the FI dataset, lack the attribute annotations as class semantic features, we use a text-based model as an alternative, such as Word2vec [25]. Word2vec is learned from a large-scale text corpus in an unsupervised manner, which requires little or no human labor to annotate. However, they only capture the weakly semantic relationship between different classes and are not discriminative enough to classify. What's more, visual features directly extracted from pre-trained CNN model are also limited by the affective gap. In order to align visual and class semantic features with more emotional structure in the latent intermediate space, we introduce an independent affective structural embedding.

First, we utilize the mid-level semantic representations ANP features to construct an intermediate latent space. Given a training image x, we choose the pre-trained ANP detector DeepSentiBank [8] $S(\cdot)$ to extract the ANP feature $S(x) \in \mathbb{R}^d$. To learn an effective latent space for compact affective representation of original affective features $S(x)$, we adopt an auto-encoder model. Suppose the input of the auto-encoder as the d-dimension ANP feature $S(x) \in \mathbb{R}^{d \times n}$ which contains n samples. The encoding part of the auto-encoder embeds the input into the l-dimension latent space $h(x)$ using FC layers, which can be defined as: $h(x) = f(W_1 S(x))$. Similarly, the decoder aims to reconstruct the input as $\hat{S(x)} \in \mathbb{R}^{d \times n}$: $\hat{S(x)} = f(W_2 h(x))$. W_1 and W_2 are the weight matrices of the FC layers and $f(\cdot)$ is the activation function. To learn the auto-encoder parameters, the input and output of the auto-encoder should be close enough by optimizing the following loss:

$$\mathcal{L}_{re} = \|\hat{S(x)} - S(x)\|_2^2. \tag{1}$$

Meanwhile, the class semantic feature z corresponding to the training image x is also projected into the learned latent ANP space by the non-linear embedding $\phi(\cdot)$. In other words, we hope to minimize the distance between embedded class semantic features $\phi(z)$ and learned latent ANP features $h(x)$. So the first part of loss function is defined as $\|h(x) - \phi(z)\|_2^2$. On the other hand, class semantic space and learned latent space have different inter-class structures. As the auto-encoder is used to reconstruct the ANP features and make the latent features preserving emotion-relevant information, we seek to match the structures of class semantic features and the ANP features in the learned latent space. Inspired by [15, 32], we project the class semantic features to the mean of the ANP features of the corresponding classes. Thus the second part of loss function is defined as $\|\mathcal{C}^h - \phi(z)\|_2^2$, where $\mathcal{C}^h$ denotes the mean vector of the latent ANP features $h(x)$ for each class. By optimizing all the above constraints, embedded class semantic features could learn emotion-relevant class representations, which are better associated with the latent emotional concepts. The affective structural alignment loss is formulated as:

$$\mathcal{L}_{as} = \|h(x) - \phi(z)\|_2^2 + \|\mathcal{C}^h - \phi(z)\|_2^2. \tag{2}$$

The total affective structural embedding is optimized by the combination of reconstruction loss and the affective structural alignment loss:

$$\mathcal{L}_{ae} = \mathcal{L}_{re} + \mathcal{L}_{as}. \tag{3}$$

3.3. Affective Adversarial Constraint

To address the zero-shot learning problem, we also embed visual and class semantic features to construct visual-semantic embedding model and measure similarities. Suppose $\theta(\cdot)$ denotes the visual features embedding process, the loss to align the visual and semantic features is defined as:

$$\mathcal{L}_{zl} = \|\theta(x) - \phi(z)\|_2^2. \tag{4}$$

Currently, both the traditional visual-semantic embedding and proposed affective structural embedding contribute to recognizing unknown emotions. Visual features have a better discriminative capacity while ANP features contain some affective structural information, which bridges the affective gap and useful for the zero-shot emotion learning. However, it is difficult to combine both embeddings effectively during training process as both visual and affective structural embedding are dynamically changing. Our goal is to retain the discriminative capacity of visual features $\theta(x)$ and combine the rich affective structural information preserved in $h(x)$. To this end, we apply an adversarial constraint, which try to fool a discriminator network D so that the output visual features are as similar as embedded ANP features:

$$\mathcal{L}_{adv} = \mathbb{E}_x \left(\log D(h(x))\right) + \mathbb{E}_x \left(\log \left[1 - D(\theta(x))\right]\right), \tag{5}$$

where $\theta(\cdot)$ tries to minimize $\mathcal{L}_{adv}$ against D that tries to maximize it. Considering this kind of adversarial learning is tricky to optimize, in order to obtain better train stability, we adopt the strategy of WGAN [4]. Please refer to [4] for more details.

Combining all above mentioned constraints, the whole model is trained by the following loss function:

$$\mathcal{L} = \mathcal{L}_{zl} + \mathcal{L}_{ae} + \mathcal{L}_{adv}. \tag{6}$$

3.4. ZSL Prediction

Given the test image x and the set of class semantic features $\mathbf{Z}$ of candidate emotion classes, we can classify the

unseen emotion class via the simple nearest-neighbor research. More specifically, the test image and candidate class semantic features are fed into the visual and semantic embedding branch separately to get $\theta(x)$ and $\phi(\mathbf{Z})$. Then the test image is recognized by calculating its distance to the class semantic embedding features in the latent space:

$$\hat{y}^t = \min_{y \in \mathcal{Y}_\mathcal{U}} \|\theta(x) - \phi(\mathbf{Z}^y)\|_2^2, \tag{7}$$

where $\mathbf{Z}$ denotes the semantic features associated with the emotion label y. For the generalized ZSL setting, we only need to modify the candidate space of labels as $y \in \mathcal{Y}_\mathcal{U} \cup \mathcal{Y}_\mathcal{S}$.

4. Experiments

In this section, we first introduce the detailed experimental setup, including the datasets, implementation details, and evaluation metrics. And then we compare with the state-of-the-art approaches and analyze the results. More results are available in the supplementary material.

4.1. Datasets

We perform our experiments on five datasets, including Flickr and Instagram (FI) [42], IAPSa [23], ArtPhoto [22] ,Emotion6 [27] and Abstract Paintings [22]. The FI dataset is collected from 3 million weakly labeled web images Flickr and Instagram by labeling with Mikels eight emotion categories. A group of 225 Amazon Mechanical Turk workers was employed to label the images. In total, 23,308 images receiving at least three agreements between workers are included in the FI dataset. The International Affective Picture System (IAPS) is widely used in visual sentiment analysis research, which contains 395 pictures from IAPS and is also labeled with Mikels eight emotion categories. ArtPhoto includes 806 art photographs from photo sharing sites and the owner of each image provides the ground truth labels. Abstract Paintings includes 228 abstract paintings consisting of texture and color. Emotion6 is collected from Flickr for the sentiment prediction, which contains 1980 images and is annotated by seven emotional categories.

4.2. Implementation Details

We employ the ResNet-50 model [16] as the backbone CNN network and initialize our framework with the weights from the pre-trained model on ImageNet. In addition, we apply the pre-trained DeepSentiBank [8] detector to extract 2089-dimension ANP features. For the auto-encoder, the dimension of the latent ANP feature is fixed to 1024. We utilize a fully connected layer before the ReLU layer in order to embed both the visual and semantic features to the latent ANP space. The discriminator D is composed of two fully connected layers and a ReLU layer, and takes 1024-d features as input. The learning rate of stochastic gradient descent (SGD) is 1e-4 and the weight decay is 1e-3. The momentum is set as 0.9. We implement our model with Pytorch and run all experiments on an NVIDIA GTX 1080Ti GPU.

For the class semantic feature, we choose to use Word2vec [25], where each instance is represented by a 300-dimensional vector. The features are constructed automatically from large unlabeled text corpora without additional manual annotation.

4.3. Evaluation Metrics

Following previous ZSL works [36], we employ the average per-class accuracy as the evaluation metric. For the generalized ZSL setting [36], we compute the average per-class accuracy on unseen classes ($A_{U\to T}$) and average per-class accuracy on seen classes ($A_{S\to T}$) when the prediction label set is the union of seen and unseen classes. We also compute the harmonic mean (H) on seen classes and unseen classes, *i.e.*, $H = 2 * (A_{U\to T} * A_{S\to T})/(A_{U\to T} + A_{S\to T})$.

4.4. Results and Analysis

To evaluate the effectiveness of our model for zero-shot emotion recognition, we compare with a variety of ZSL methods, including common ZSL methods (*i.e.* LATEM [34], SSE [44], SAE [20] and DEM [43]) and the recent ZSL methods (*i.e.* LAD [19], CDL [18] and RN [31]). Since the ANP features extracted by the DeepSentiBank detector can be used as visual features in image emotion recognition, we also report the zero-shot recognition results using extracted ANP features.

We evaluate the performance of the proposed zero-shot emotion recognition method on five affective datasets. As shown in Table 1, we conduct the experiment on each dataset with two kinds of split strategies in order to prove the effectiveness and robustness of the proposed method. In detail, for the datasets with 8 emotions, the splits of training classes and testing classes are 6:2 and 4:4; for the Emotion6 dataset with 6 emotions, the splits of training classes and testing classes are 4:2 and 3:3. For the FI dataset with $6:2$ split setting, our method attains 68.87%, which is slightly higher than the state-of-the-art reported by CDL (67.07%). Among the comparison methods, CDL gets the best performance followed by DEM (65.49%) while SSE gets the worst performance. Our method and CDL both consider the structural information of different space during the embedding process, so the performance is much better than other methods. Besides, our method further considers the specific affective structural information which is relevant to the emotion recognition. What's more, we could utilize the affective adversarial constraint to automatically find the optimal solution of combining the two features when both features are changing during the training process. Thus, our method could obtain the best performance. For the $4:4$

Table 1. Zero shot emotion recognition accuracy (%) of all methods on the FI, ArtPhoto, Abstract, IAPSa and Emotion6 datasets. We evaluate the proposed model with several baseline zero-shot learning methods. S denotes the DeepSentiBank features, while D denotes the CNN-based features, and M denotes the concatenation of the DeepSentiBank and CNN-based features.

		FI		ArtPhoto		Abstract		IAPSa		Emotion6	
Setting		6 : 2	4 : 4	6 : 2	4 : 4	6 : 2	4 : 4	6 : 2	4 : 4	4 : 2	3 : 3
LATEM [34]	D	53.32	32.73	38.57	27.66	56.71	24.64	49.81	25.57	52.88	32.42
	S	57.79	38.59	42.98	25.26	49.04	22.39	43.74	25.26	55.91	29.89
	M	58.25	37.57	44.35	24.17	42.37	25.68	41.24	29.35	55.12	34.82
SSE [44]	D	42.67	33.61	45.57	26.55	47.36	21.41	41.34	22.97	42.12	28.99
	S	43.37	21.04	42.85	26.83	44.73	17.64	50.73	27.47	53.18	27.98
	M	42.02	33.55	40.63	20.51	49.64	22.33	44.51	30.62	52.85	31.21
SAE [20]	D	61.12	37.34	49.66	23.45	60.53	29.41	49.04	29.73	44.24	38.89
	S	54.66	31.60	51.70	22.88	42.11	25.49	53.85	32.43	45.45	37.88
	M	57.82	26.57	45.58	23.45	52.63	28.43	48.08	29.28	54.55	35.86
LAD [19]	D	51.44	34.94	41.27	20.53	43.21	13.75	38.42	27.66	45.36	33.65
	S	43.97	29.36	43.51	22.34	45.44	15.24	35.71	22.68	40.04	28.34
	M	44.18	26.53	42.18	22.03	42.11	23.53	50.96	18.83	47.27	32.93
DEM [43]	D	65.49	49.37	48.30	29.66	63.42	32.35	47.12	34.69	50.61	33.74
	S	62.06	34.70	48.97	31.35	60.53	32.35	46.15	30.18	51.66	36.36
	M	64.73	50.96	50.33	27.97	60.71	30.72	44.52	31.38	50.43	31.28
RN [31]	D	64.97	47.83	40.45	28.86	57.85	26.69	40.57	32.43	56.51	39.49
	S	50.34	31.57	44.23	23.79	42.51	16.55	43.31	30.76	52.45	29.21
	M	63.99	49.31	47.14	25.32	56.16	29.52	43.98	33.62	49.85	33.64
CDL [18]	D	67.03	41.28	50.52	30.46	52.11	24.11	52.88	30.56	51.67	36.36
	S	61.44	36.88	48.57	25.81	53.64	22.62	51.00	35.74	56.97	40.10
	M	67.07	41.37	48.35	29.22	55.19	27.42	48.73	33.92	53.86	37.91
Ours		**68.87**	**54.73**	**53.22**	**35.58**	**64.71**	**34.45**	**57.82**	**38.30**	**59.94**	**42.83**

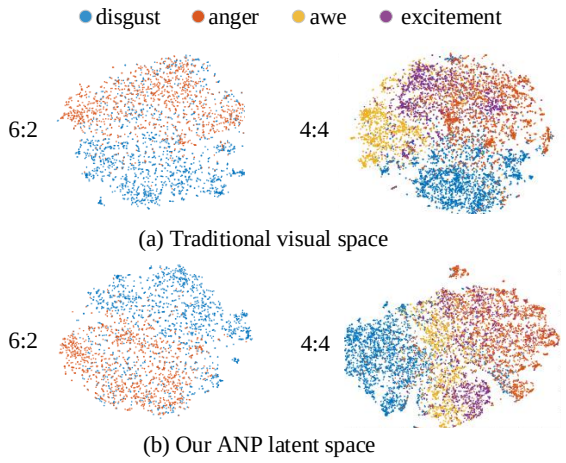
Table 2. Generalized ZSL recognition accuracy on the FI dataset following $6:2$ split setting and cross dataset recognition accuracy of all methods between the FI dataset and Emotion6 datase. $A_{U\to T}$ denotes the recognition accuracy on unseen classes, while $A_{S\to T}$ denotes the average recognition accuracy on seen classes. H denotes the harmonic mean. 'FI→Emo6' denotes using the images of FI dataset belong to the common categories as training set and the images of Emotion6 dataset belong to the other categories as the test set. 'Emo6→FI' denotes the similar setting but change the dataset.

Method	$A_{U\to T}$	$A_{S\to T}$	H	FI→Emo6	Emo6→FI
LATEM [34]	1.82	55.31	3.54	51.21	26.43
RN [31]	3.23	62.56	6.14	59.85	29.22
SSE [44]	7.51	53.57	13.17	51.32	29.40
DEM [43]	13.43	56.25	21.68	56.52	22.36
LAD [19]	20.83	59.46	30.85	52.27	21.03
SAE [20]	24.25	65.59	35.42	54.70	30.03
CDL [18]	26.48	54.87	35.72	55.15	32.34
Ours	**28.12**	**66.57**	**39.54**	**61.94**	**34.48**

split setting, our method has achieved gains up to 3.77% than DEM (50.96%). For the other small datasets, our method can still achieve 1.29% $\sim$ 3.97% improvements.

We also observe that the ANP features (S) have a similar performance with the deep features (D) extracted by deep learning models pre-trained on ImageNet. For example, on the Emotion6 dataset, the results of using ANP features are generally better than using deep features, *e.g.* LATEM, SSE, SAE, DEM, and CDL. For other datasets such as the FI dataset, the performance of using ANP features is slightly lower in most cases. This demonstrates that the ANP features contain some affective structural information, which may be not discriminative for classification tasks, but useful for the zero-shot learning problem. Our method utilizes deep features with the ANP features as a supplement and shows the effectiveness of considering both discriminative visual features and affective structural information. We have also validated all the comparison methods using both ANP and deep features and reported the results in the rows marked with M in Table 1. That also shows that applying both two features directly induces no performance gain, while our method effectively improves zero-shot emotion recognition performance.

We further report the generalized ZSL recognition accuracy of all methods on the FI dataset following $6:2$ split setting in Table 2. For the accuracy on unseen classes $A_{U\to T}$, CDL obtains the best performance 26.48% among the comparison methods, while SAE achieves 65.59% when it comes to the accuracy on seen classes $A_{S\to T}$, which is much better than CDL. For $A_{U\to T}$ and $A_{S\to T}$, our method obtains 28.12% and 66.57%. Compared with the most competitive CDL, our method improves the harmonic mean by 3.82% on the FI dataset. Our method outperforms all com-



(a) Traditional visual space

(b) Our ANP latent space

Figure 3. t-SNE plots of the distribution of unseen class visual samples from the FI dataset in traditional visual space and our ANP latent space. "6:2" and "4:4" denote the two splits for zero-shot emotion recognition.

pared methods in all three cases.

4.5. Ablation Study

Table 3. Ablation experiment on the FI dataset. The baseline is the basic visual-semantic embedding to conduct zero-shot emotion recognition. $\mathcal{L}_{as}$ and $\mathcal{L}_{re}$ denote two parts of affective structural embedding. $\mathcal{L}^{*}_{as}$ denotes the structural alignment loss without the second part. $\mathcal{L}_{adv}$ denotes using the affective adversarial constraint.

Base	$\mathcal{L}_{adv}$	$\mathcal{L}^{*}_{as}$	$\mathcal{L}_{as}$	$\mathcal{L}_{re}$	FI
√					65.12
√		√			65.46
√			√		66.41
√	√		√		67.53
√			√	√	67.28
√	√		√	√	68.87

We conduct ablation experiments to illustrate the effectiveness of the affective structural alignment and the affective adversarial constraint in Table 3. In particular, "Base" denotes the basic visual-semantic embedding to conduct zero-shot emotion recognition, where visual and class semantic features are directly embedded into the common space and measured similarities. From the results, we can clearly see that the affective adversarial constraint plays a significant role in improving the zero-shot recognition accuracy by 1.12% on the FI dataset. The results validate that adversarial learning combining two embedded visual and ANP features is superior to those direct combination or in a manual combination rule.

Affective structural embedding optimized by $\mathcal{L}_{as}$ and both $\mathcal{L}_{as}$ and $\mathcal{L}_{re}$ further boost the ZSL accuracy by 1.29% and 2.16%. Among the affective structural embedding, aligning the class semantic features with the center of the latent ANP features also results in 0.95% performance improvement. The results demonstrate that considering emotional structural information to align the visual and semantic features contributes a lot to zero-shot emotion recognition. Fig. 3 illustrates the distribution of unseen class visual samples in traditional visual space and our ANP latent space. It clearly shows that the embedded visual features are more separated from other classes in the ANP space. Finally, we obtain the best performance by training all the components, which shows the complementarity of these contributions.

Table 4. Zero-shot emotion recognition accuracy (%) on the FI dataset with different choices of testing classes. Note that $*$ means that we choose different testing classes under the same train/test radio corresponding to Table 1. Here, we take "excitement" and "sadness" as test classes for the 6:2 split setting and all the negative emotions as test classes for the 4:4 split setting.

Setting	LATEM	SSE	SAE	LAD	CDL	DEM	RN	Ours
$6:2^{*}$	70.47	63.73	67.78	61.84	68.70	71.98	72.41	**74.03**
$4:4^{*}$	27.09	37.94	42.24	38.30	41.95	37.84	39.87	**42.92**

We also explore the influence of different testing classes in Table 4 under the same train/test radio in Table 1. For the $6:2^{*}$ split setting, we take the "excitement" and "sadness" as test classes and others as training classes while in Table 1 we choose two negative emotions as test classes. For the $4:4^{*}$ split setting, we take all the negative emotions as test classes (*i.e.* "fear", "sad", "disgust" and "anger") and all the positive emotions as training classes while in Table 1 training and test classes are both two positive emotions and two negative emotions. We can see that, the performance of almost all the compared methods is improved when test emotions are in two polarities and is decreased when test emotions are all negative and close to each other. Our method achieves the state-of-the-art results consistently with different emotion prediction configurations.

4.6. Cross Datasets Recognition

To better evaluate the performance of zero-shot learning on image emotion recognition tasks, we further conduct the cross datasets experiment for emotion recognition. Conventional zero-shot learning methods assume the testing images are sampled from the same distribution with the training images, which is inconsistent with the real situation. Thus we try to recognize unseen emotion categories in different datasets. The FI and Emotion6 datasets share the four emotion categories including sadness, disgust, fear and anger. Specifically, we take the four emotion categories from one dataset as the training set and other categories from another dataset as the test set. Ideally, we focus on the image emotion recognition, which means the more emotion-related features and embedding, the higher the performance.

Table 2 shows the results of the cross datasets recognition. The proposed method obtains the best results compared with the other zero-shot methods in both two experiments. In particular, the proposed method achieves

Figure 4. Qualitative results of CDL and our method on FI (a) and Emotion6 (b). Predictions of unseen class labels for both two datasets are listed from top to bottom in probability and the most probable prediction is at the top. The ground truth labels are in red. "ex, aw, di, an, jo, sa and su" denote excitement, awe, disgust, anger, joy, sadness and surprise respectively.

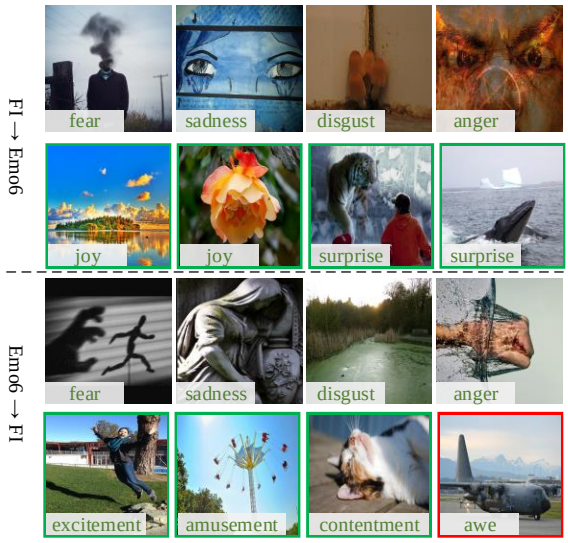


Figure 5. Qualitative results for cross dataset recognition on the FI and Emotion6 dataset. Misclassified images are marked with red bounding boxes.

2.08% improvements under FI→Emo6 setting and improves the recognition accuracy by 2.14% improvements under Emo6→FI setting. Affective structural information provided by ANP features are shared by different datasets. Considering both the affective structural information and affective enhanced visual features ensure the ability of our method to cross datasets recognition.

4.7. Qualitative Results

We present some qualitative analysis for our proposed algorithm and CDL on FI and Emotion6 dataset in Fig. 4. As shown in Fig. 4 (a), the image of the class "excitement and "anger" can be correctly predicted among the four unseen classes. While the example image of the class "disgust" is predicted to the class "anger". Both emotion classes are similar in the emotion theories and the same image may evoke two emotions in different situations. Actually, the ground truth "disgust" comes in second just following "disgust" in the predictions. For the last examples in the Emotion6 dataset Fig. 4 (b), our model predictions confuse the class "sadness" and "surprise". These results further prove the ambiguity of emotions that even people will feel difficult to distinguish such similar emotions. Compared with the most competitive method, we can also see that our method significantly outperforms CDL in these examples. Although CDL still gives some correct results in cases, the predictions do not show the relationship between emotions. On the other hand, our method can not only output the correct prediction, but also output higher prediction probability for the emotions that are close to the ground truth, while the probability of the opposite emotion is lower.

Fig. 5 shows some results of cross datasets recognition. When unseen classes belong to the Emotion6 dataset, our model distinguishes the "joy" and "surprise" classes successfully. For the last examples, when unseen classes belong to the FI dataset, our model predicts the "awe" class to the "excitement". This is probably because such military images could evoke different emotions among different people. Interestingly, we train the model with images of four negative emotions and our model could recognize different positive emotions in the FI and Emotion6 dataset.

5. Conclusion

In this paper, we propose a novel affective structural embedding framework for the zero-shot emotion recognition problem. By utilizing ANP features to construct an affective embedding space, the affective gap between visual and semantic features can be effectively bridged. In addition, we introduce an affective adversarial constraint to force the visual embedding to choose an embedding space that preserves the affective structural information and retains the discriminative capacity simultaneously. Experiments on five widely-used affective datasets show that our method significantly outperforms the state-of-the-art approaches for zero-shot emotion recognition.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2016. 2

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 2

[3] Y. Annadani and S. Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, 2018. 3

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 4

[5] L. F. Barrett. Are emotions natural kinds? *Perspectives on psychological science*, 1(1):28–58, 2006. 2

[6] D. Borth, T. Chen, R. Ji, and S.-F. Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM MM*, 2013. 2

[7] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013. 2

[8] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014. 2, 4, 5

[9] M. L. Chuang Gan, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic interclass relationships (sir) for zero-shot action recognition. In *AAAI*, 2015. 1

[10] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 1

[11] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987. 1

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2

[13] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2

[14] H. Gauba, P. Kumar, P. P. Roy, P. Singh, D. P. Dogra, and B. Raman. Prediction of advertisement preference by fusing eeg response and sentiment analysis. *Neural Networks*, 92:77–88, 2017. 1

[15] O. Gune, B. Banerjee, and S. Chaudhuri. Structure aligning discriminative latent embedding for zero-shot learning. In *BMVC*, 2018. 4

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[17] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. Can we understand van gogh's mood? learning to infer affects from images in social networks. In *ACM MM*, 2012. 1

[18] H. Jiang, R. Wang, S. Shan, and X. Chen. Learning class prototypes via structure alignment for zero-shot recognition. *ECCV*, 2018. 3, 5, 6

[19] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen. Learning discriminative latent attributes for zero-shot classification. In *ICCV*, 2017. 3, 5, 6

[20] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017. 3, 5, 6

[21] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett. The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences*, 35(3):121–143, 2012. 2

[22] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010. 2, 5

[23] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626–630, 2005. 1, 5

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 4, 5

[26] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *CVPR*, 2018. 3

[27] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 2015. 1, 5

[28] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3

[29] S. Qian, T. Zhang, and C. Xu. Multi-modal multi-view topic-opinion mining for social event analysis. In *ACM MM*, 2016. 1

[30] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 2

[31] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: relation network for few-shot learning. In *CVPR*, 2018. 3, 5, 6

[32] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 4

[33] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014. 1

[34] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 2, 5, 6

[35] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 3

[36] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, 2017. 1, 5

[37] J. Yang, D. She, Y. Lai, and M.-H. Yang. Retrieving and classifying affective images via deep metric learning. In *AAAI*, 2018. 1, 2

[38] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018. 1, 2

[39] J. Yang, D. She, and M. Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017. 1

[40] Q. You, H. Jin, and J. Luo. Visual sentiment analysis by attending on local image regions. In *AAAI*, 2017. 2

[41] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015. 2

[42] 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 *AAAI*, 2016. 1, 5

[43] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 3, 5, 6

[44] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. 2, 5, 6

[45] S. Zhao, Y. Gao, G. Ding, and T.-S. Chua. Real-time multimedia social event detection in microblog. *IEEE Transactions on Cybernetics*, (99):1–14, 2017. 1

[46] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 2014. 2

[47] P. Zhu, H. Wang, T. Bolukbasi, and V. Saligrama. Zero-shot detection. *arXiv preprint arXiv:1803.07113*, 2018. 1

[48] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 3