Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network

Jufeng Yang, Ming Sun, Xiaoxiao Sun
College of Computer and Control Engineering, Nankai University
Tianjin, China

Abstract

Visual sentiment analysis is raising more and more attention with the increasing tendency to express emotions through images. While most existing works assign a single dominant emotion to each image, we address the sentiment ambiguity by label distribution learning (LDL), which is motivated by the fact that image usually evokes multiple emotions. Two new algorithms are developed based on conditional probability neural network (CPNN). First, we propose BCPNN which encodes image label into a binary representation to replace the signless integers used in CPNN, and employ it as a part of input for the neural network. Then, we train our ACPNN model by adding noises to ground truth label and augmenting affective distributions. Since current datasets are mostly annotated for single-label learning, we build two new datasets, one of which is relabeled on the popular Flickr dataset and the other is collected from Twitter. These datasets contain 20,745 images with multiple affective labels, which are over ten times larger than the existing ones. Experimental results show that the proposed methods outperform the state-of-the-art works on our large-scale datasets and other publicly available benchmarks.

Introduction

In recent years, lots of attention has been paid to affective image classification (Jou et al. 2015; Joshi et al. 2011; Chen et al. 2015). Most of these works are conducted by psychological studies (Lang 1979; Lang, Bradley, and Cuthbert 1998), and focus on manual design of features and classifiers (You et al. 2015a). As defined as a single-label learning (SLL) problem which assigns a single emotional label to each image, previous works (You et al. 2016; Sun et al. 2016) have performed promising results.

However, image sentiment may be the mixture of all components from different regions rather than a single representative emotion. Meanwhile, different people may have different emotional reactions to the same image, which is caused by a variety of elements like the different culture background and various recognitions from unique experiences (Peng et al. 2015). Furthermore, even a single viewer may have multiple reactions to one image. Figure 1 shows examples from a widely used dataset, i.e. Abstract Paintings (Machajdik and Hanbury 2010), which provides detailed annotations by 14 users. Existing works use this dataset for SLL task, that is, for each image the category with the most votes is selected as the ground truth. Surprisingly, the 14 users can not reach an agreement on any of the total 228 samples, which indicates using the dominant votes as sentiment label misses the diversity of viewer emotions. This fact encourages us to explore multiple affective labels in images.

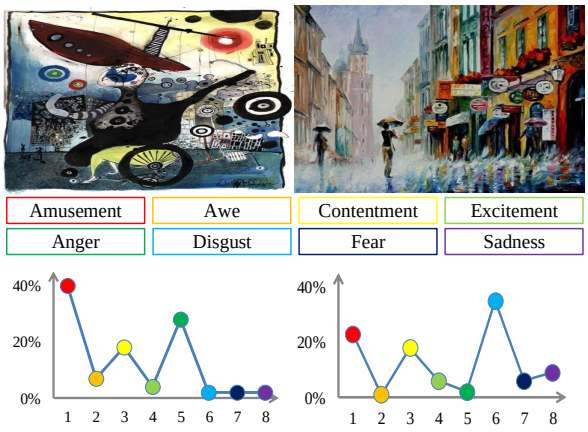


Figure 1: For each image, its sentiment label distribution is shown at the bottom. Both images are selected from the Abstract Paintings, and annotated by 14 users. Different colors indicate different sentiments, e.g. red is for amusement, and yellow is for contentment.

Multi-label learning (MLL) studies the problem where one instance is associated with a number of class labels (Zhou and Zhang 2006). Since MLL does not fit some real applications well where the overall distribution of the importance of the labels matters, label distribution learning (LDL) is proposed to cover a certain number of labels, representing the degree to which each label describes the instance (Geng 2016). A state-of-the-art algorithm is conditional probability neural network (CPNN) (Geng, Yin, and Zhou 2013). However, CPNN uses signless integers as label representation and inputs them into the network for computation. It's unreasonable doing this in sentiment prediction systems, be-

cause it is meaningless to add two sentiment labels or subtract one label from another. Besides, CPNN aims to predict probability density of various classes which needs abundant samples in training phase, while labeling image sentiments is subjective and time-consuming.

To address the problems, we introduce two new algorithms named BCPNN and ACPNN, respectively. In BCPNN, the integer labels are replaced by an m-dimensional binary code, where m is the number of categories in affective datasets. Thus, each emotional class owns its weight in network instead of sharing with others, which further boosts performance. Then, we build ACPNN based on BCPNN, aiming to augment sentiment labels based on the observation that image sentiments are usually unbalanced. We add noises to the ground truth sentiment labels to generate more roughly labeled distributions and samples. These augmented labels have similar properties with the ground truth while varying in concrete degrees, which make our model more robust.

As aforementioned, most previous works employ a dominant emotion as image label. For example, a weakly labeled dataset (Borth et al. 2013) from Flickr is generated by adjective noun pairs to distinguish positive and negative emotions (You et al. 2015b). To evaluate our proposed methods, we relabel a subset of the Flickr dataset with eight emotions {*Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, Sadness*}, and name it as **Flickr_LDL**. Then, we download 30,000 images from Twitter, labeling them in the same eight emotions space. We provide an extra option to label an image as neutral when no obvious emotions can be found. As a result, about two-thirds of images are removed in this process. The rest images with multiple labels are named **Twitter_LDL**. We evaluate our proposed methods on both of these large-scale datasets, as well as other two benchmark datasets, i.e. Abstract Paintings (Machajdik and Hanbury 2010) and Emotion6 (Peng et al. 2015). The results show ACPNN is superior to BCPNN and both of them outperform the state-of-the-art works.

Our contributions are threefold. First, we introduce MLL and LDL into image sentiment prediction, while previous works usually treat it as a SLL problem. Second, we propose two new models, BCPNN and ACPNN, to address the problem, which take advantage of binary label representation and augment affective labels, respectively. Last, two large-scale affective datasets with multiple labels are collected. These new datasets contain over ten times images than the available benchmarks. We make the datasets publicly available to peer researchers at `http://cv.nankai.edu.cn/projects/SentiLDL`, which will be beneficial to further researches in this field.

Related Work

Emotion Modeling

Existing approaches can be grouped into two aspects: dimensional spaces and categorical states. Researches on dimensional approaches map affective representations into valence-arousal space (Nicolaou, Gunes, and Pantic 2011; Xu et al. 2008) or activity-weight-heat space (Solli and Lenz 2009). In contrast, categorical approaches (Chen et al. 2014b; Zhao et al. 2014) classify emotions into representative categories. Compared to the former, categorical models make it easier for people to understand. Typical categorical methods try to solve the problem using low-level features and mid-level representations, while several recent works exploit deep features and achieve significant progress (Chen et al. 2014a). You (You et al. 2015b; You et al. 2016) proposes a novel progressive CNN architecture to make use of noisy data, and further performs benchmarking analysis on a massive scale well-labeled dataset.

Although visual sentiment has been studied from various perspectives, the destination of all these works is to predict one most descriptive word from the label set. However, as emotions evoked by images are affected by various factors, choosing one single emotion to represent the whole image is insufficient and unreliable. According to Plutchik's wheel of emotion theory (Plutchik 1980), only small amounts of emotions are the basic ones, based on which occurring the other emotions as the combination results. Each affective image, on the contrary to the single emotion assumption of most existing methods, usually reflects a mixture of basic emotions with different intensities.

Label Distribution Learning

Learning with ambiguity has been a popular topic of machine learning for years, and MLL is successfully applied to many computer visual tasks. In the framework of MLL, each instance is represented by a single feature vector simultaneously associated with multiple class labels (Zhang and Wu 2015; Bengio, Weston, and Grangier 2010). However, MLL can hardly deal with the problem of describing the exact role of each label, in which it is unlikely that multiple affective labels happen to be equally relevant to the image. Therefore, this work represents sentiment via a distribution constituted by degrees to basic emotions and employs LDL for prediction.

In a recent work, LDL is used to transform a single label into a Gaussian function (Geng, Yin, and Zhou 2013). Then, CPNN is proposed to predict label distribution. Besides, existing methods in computer vision are improved through different strategies, such as problem transformation (PT), algorithm adaption (AA), and specialized algorithms design(SA), to deal with LDL problem. The structure of CPNN is similar to a neural network having only one hidden layer except for two differences. One is that CPNN takes both features and labels as input, the other is it outputs label distribution while neural network outputs predicted probability of single or multiple labels.

Building Datasets for LDL

Some datasets have been designed for single label affective image classification, including IAPSa (Lang, Bradley, and Cuthbert 1999), Artphoto, Abstract Paintings (Machajdik and Hanbury 2010), F&I (You et al. 2016), and Flickr (Borth et al. 2013). People also collect two datasets from Twitter (Borth et al. 2013; You et al. 2015b) and label the images with binary emotions. For convenience, they are named

Table 1: Sentiment datasets in different systems. "F&I" is the abbreviation of Flickr and Instagram, and "Abstract" denotes the Abstract Paintings dataset.

Label Type	Dataset	Classes#	Images#
Single	IAPSa	8	395
	Artphoto	8	882
	Twitter I	2	596
	Twitter II	2	1,269
	F & I	8	23,308
	Flickr	2	484,222
Multiple	Abstract	8	228
	Emotion6	7	1,980
	Flickr_LDL	8	10,700
	Twitter_LDL	8	10,045

as Twitter I and Twitter II in this paper, respectively. All of aforementioned datasets are used for SLL tasks, except for Emotion6 which has multiple affective labels (Peng et al. 2015). However, this dataset doesn't use the eight emotions space, and it is mainly built for transfering emotions among images. Table 1 briefly summarizes current datasets, as well as our proposed Flickr_LDL and Twitter_LDL. More details are discussed in the supplemental material.

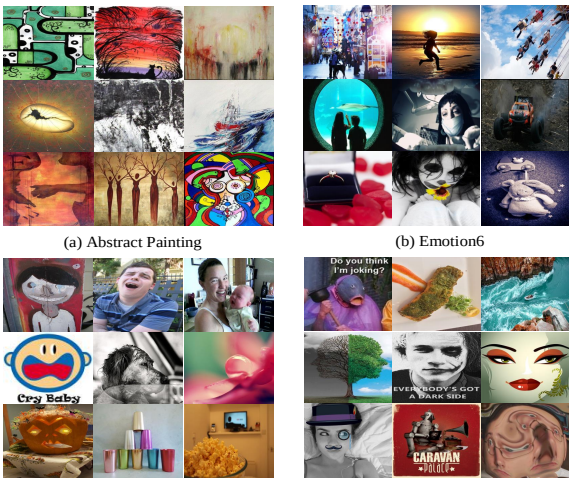
In this paper, we establish two new large-scale datasets for learning visual sentiment distributions. First, we extract a subset from the Flickr dataset (Borth et al. 2013). Unlike the other datasets that use the name of emotions to search images, the Flickr dataset is gathered upon 1,200 adjective noun pairs, reaching half million images. We hire 11 viewers to label the subset wth the eight commonly used emotions. Finally, Flickr_LDL containing 10,700 images is generated, in which the numbers of each class are roughly equal.

Twitter is another valuable source of affective images. Existing Twitter datasets contain only several hundreds of images and are labeled with binary emotions. To build a large-scale LDL dataset from Twitter, we collect about 30,000 images by searching various sentiment key words. For example, "sad", "heart-broken" and "grieved" are used to search sad images. To avoid duplication, the images with high similarity are filtered by several rounds of hand operation, before 8 viewers are hired to label this dataset. Finally, there are 10,045 images in the so-called Twitter_LDL dataset. In both datasets, the votes from the workers are integrated to generate the ground truth label distribution for each image.

Methods

We first formulate the state-of-the-art LDL algorithm CPNN (Geng, Yin, and Zhou 2013), then propose our BCPNN and ACPNN successively.

For an image $\mathbf{x}$, the description degree $d_{\mathbf{x}}^{y} \in [0,1]$ is assigned to each affective label y, representing the degree to which y describes $\mathbf{x}$. $d_{\mathbf{x}}^{y}$ satisfies the constrains $\sum_y d_{\mathbf{x}}^{y} = 1$, which can be represented by the conditional probability $p(y|\mathbf{x})$. Special attention should be paid to the meaning of $d_{\mathbf{x}}^{y}$, which is not the probability that y correctly labels $\mathbf{x}$, but the proportion that y accounts for in a full class description



(a) Abstract Painting (b) Emotion6

(c) Flickr_LDL (d) Twitter_LDL

Figure 2: Images for learning visual sentiment distributions.

of $\mathbf{x}$. Thus, all the affective labels with a non-zero description degree are actually the 'correct' labels to describe the image, but just with different importance measured by $d_{\mathbf{x}}^{y}$.

Let n denote the number of images and m denote the number of affective classes. Given a training set $S = \{(\mathbf{x}_1, D_1), (\mathbf{x}_2, D_2), \ldots, (\mathbf{x}_n, D_n)\}$, where $\mathbf{x}_i$ is the i-th image and $D_i = \{d_{\mathbf{x}_i}^{y_1}, d_{\mathbf{x}_i}^{y_2}, \ldots, d_{\mathbf{x}_i}^{y_m}\}$ is the label distribution associated with $\mathbf{x}_i$, our goal is to learn a conditional probability mass function $p(y|\mathbf{x})$ from S. Suppose $p(y|\mathbf{x})$ is a parametric model $p(y|\mathbf{x};\mathbf{w})$, where $\mathbf{w}$ is the vector of the model parameters. Then, we further find the $\mathbf{w}$ that can generate a distribution similar to D_i given the instance $\mathbf{x}_i$. If the Kullback-Leibler divergence is used as the measurement of the similarity between two distributions, then the best model parameter vector $\mathbf{w}^*$ is determined by

$$\begin{aligned}\mathbf{w}^* &= \operatorname*{argmin}_{\mathbf{w}} \sum_i \sum_j (d_{\mathbf{x}_i}^{y_j} \ln \frac{d_{\mathbf{x}_i}^{y_j}}{p(y_j|\mathbf{x}_i;\mathbf{w})}) \\ &= \operatorname*{argmax}_{\mathbf{w}} \sum_i \sum_j d_{\mathbf{x}_i}^{y_j} \ln p(y_j|\mathbf{x}_i;\mathbf{w}).\end{aligned} \tag{1}$$

CPNN

CPNN (Geng, Yin, and Zhou 2013) models $p(y|\mathbf{x};\mathbf{w})$ by a three layer neural network, which is shown in Figure 3. The input of CPNN includes both $\mathbf{x}$ and a discrete y, and the output of the network is

$$p(y|\mathbf{x};\mathbf{w}) = \exp(b(\mathbf{x};\mathbf{w}) + f(\mathbf{x},y;\mathbf{w})), \tag{2}$$

where bias $b(\mathbf{x};\mathbf{w})$ ensures that $\int p(\mathbf{x})d\mathbf{x} = 1$. $b(\mathbf{x};\mathbf{w})$ is defined as

$$b(\mathbf{x};\mathbf{w}) = -\ln(\sum_y \exp(f(\mathbf{x},y;\mathbf{w}))). \tag{3}$$

The net activation of output $f(\mathbf{x},y;\mathbf{w})$ is

$$f(\mathbf{x},y;\mathbf{w}) = \sum_{v=1}^{M_2} \mathbf{w}_{31m} G(\sum_{k=0}^{M_1} \mathbf{w}_{2vk}(\mathbf{x}_k, y_k)). \tag{4}$$

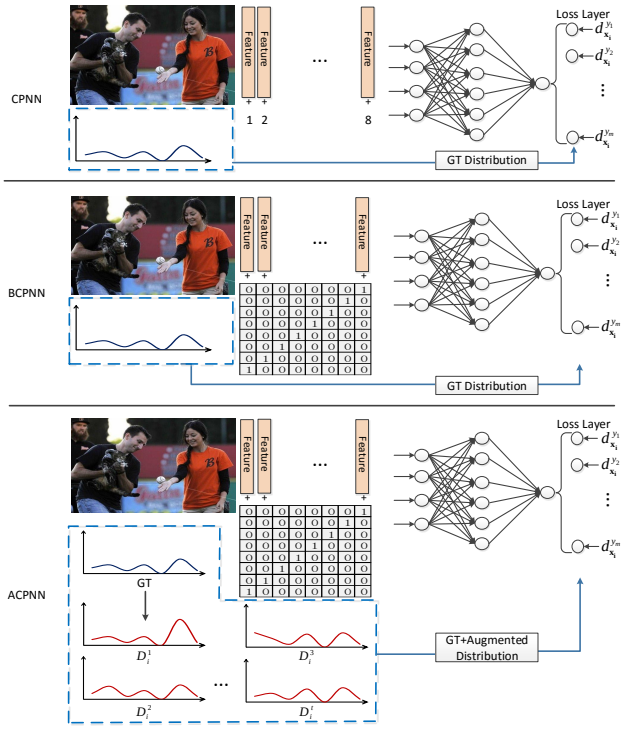


Figure 3: CPNN, BCPNN and ACPNN share the similar structures, which have three layers, take feature vector and label as input, and output the predicted distributions. Binary coding and augmented label distributions are introduced into BCPNN and ACPNN, respectively. "GT" indicates ground truth.

Here, G is the sigmoid activation function, M_l is the number of units on the l-th layer, and $\mathbf{w}_{lvk}$ is the weight of the v-th unit on the l-th layer associated with the output of the k-th unit on the $(l-1)$-th layer. Recall Equation 1, then the target function to minimize is

$$\begin{aligned} T(\mathbf{w}) &= -\sum_i \sum_j d_{\mathbf{x}_i}^{y_j} \ln p(y_j|\mathbf{x}_i;\mathbf{w}) \\ &= -\sum_i \sum_j d_{\mathbf{x}_i}^{y_j} (b(\mathbf{x}_i;\mathbf{w}) + (f(\mathbf{x}_i, y_j;\mathbf{w}))). \end{aligned} \tag{5}$$

The gradient of Equation 5 w.r.t. $\mathbf{w}$ is

$$\frac{\partial T(\mathbf{w})}{\partial \mathbf{w}} = -\sum_i \sum_j d_{\mathbf{x}_i}^{y_j} \left(\frac{\partial b(\mathbf{x}_i;\mathbf{w})}{\partial \mathbf{w}} + \frac{\partial (f(\mathbf{x}_i, y_j;\mathbf{w}))}{\partial \mathbf{w}}\right). \tag{6}$$

The partial derivative in Equation 6 can be calculated by backpropagation (Modha and Fainman 1994), and finally the weights are updated by the RPROP algorithm (Riedmiller and Braun 1993).

BCPNN

CPNN has several disadvantages when applied to visual sentiment prediction. For example, the target function of CPNN in Equation 5 has two input terms. $\mathbf{x}$ is a feature vector, in which each value ranges from 0 to 1. y_j is the j-th affective category evoked by the image ranging from 1 to m, where m is the number of categories. When they are input into network simultaneously, it's hard to balance $\mathbf{x}$ and y_j. Even y_j is normalized, it's meaningless adding two sentiment labels or subtracting one label from another.

Thus, we replace the typical signless integer y in CPNN with a binary vector $\hat{\mathbf{y}} = [b(y_1), b(y_2), \ldots, b(y_m)]$, where $b(\cdot)$ is a binary coding function. For the j-th affective category, the binary value $b(y_j)$ is 1 and the other binary values are 0. We input $\hat{\mathbf{y}}$ into the network as well as the image representations. Note our binary coding improves the length of input label from 1 to m. It's reasonable for computing because compared to the dimensions of the feature vector, m is small enough. For example, we employ mid-level features and deep features in our experiments, whose lengths are 1,200 and 4,096, respectively. However, m is very small in a visual sentiment problem, the typical value is 2, 7, or 8. As with CPNN, the description degree $d_{\mathbf{x}}^{y_j}$ is used as the output of the network, which can be represented by the following conditional probability

$$p(\hat{\mathbf{y}}|\mathbf{x};\mathbf{w}) = \exp(b(\mathbf{x};\mathbf{w}) + f(\mathbf{x}, \hat{\mathbf{y}};\mathbf{w})). \tag{7}$$

As shown in Figure 3, the conditional probability neural network with binary code is defined as BCPNN, which is regarded as a preliminary form of augmented conditional probability neural network (ACPNN). Our new target function is

$$T'(\mathbf{w}) = -\sum_i \sum_j d_{\mathbf{x}_i}^{\hat{\mathbf{y}}_j} (b(\mathbf{x}_i;\mathbf{w}) + (f(\mathbf{x}_i, \hat{\mathbf{y}}_j;\mathbf{w}))), \tag{8}$$

where $\hat{\mathbf{y}}_j$ is the binary vector corresponding to the j-th affective label. We use the similar approaches as CPNN to calculate the partial derivative and update the weights.

ACPNN

Moreover, training a robust LDL model needs a number of affective images, while labeling image sentiment is subjective and time-consuming. We propose our ACPNN based on the previous binary conditional probability neural network. Our target is to generate an extended training set $S' = S + \{(\mathbf{x}_i, D_i^v)|i = 1, 2, \ldots, n; v = 1, 2, \ldots\}$ from S, where D_i^v is the v-th augmented label distribution of image $\mathbf{x}_i$. In our experiments, the max value of v is set to 5. As defined before, $D_i = \{d_{\mathbf{x}_i}^{y_1}, d_{\mathbf{x}_i}^{y_2}, \ldots, d_{\mathbf{x}_i}^{y_m}\}$ is the ground truth label distribution. Thus, the augmented distribution is calculated by

$$D_i^v = D_i \times (I_{m\times m} + \frac{1}{v} \times diag(0, \cdots, \rho_k, \cdots, 0)) \tag{9}$$

where $I_{m\times m}$ is the identity matrix, $diag()$ is a diagonal matrix in the same size with $I_{m\times m}$, and has only one non-zero element ρ_k at the position (k, k). ρ_k is defined as

$$\rho_k = \begin{cases} \frac{y_i^k}{\sum_{j=1}^m y_i^j} & \text{if } y_i^k \geq Threshold \\ 0 & \text{else} \end{cases} \tag{10}$$

where $Threshold = E(d_{\mathbf{x}_i}^{y_j})$ is represented by the expected value of $d_{\mathbf{x}_i}^{y_j}$. We train ACPNN using both ground truth and augmented distributions. Finally, the output of our model is a predicted affective distribution for the testing image.

Table 2: Experimental results of CPNN and our proposed BCPNN and ACPNN.

Features		SentiBank						Deep features from VGGNet					
Criterion		Cheb ↓	Clark ↓	Canber ↓	KLdiv ↓	Cosine↑	Intersec ↑	Cheb ↓	Clark ↓	Canber ↓	KLdiv ↓	Cosine ↑	Intersec ↑
Abstract	CPNN	0.2933	1.7612	4.1347	0.6377	0.6804	0.5692	0.2809	**1.7434**	**4.1096**	0.5839	0.7124	0.5839
	BCPNN	0.2686	1.7607	4.1606	0.5611	0.7254	0.5877	0.2630	1.8058	4.3003	0.5877	0.7168	0.5868
	ACPNN	**0.2442**	**1.7389**	**4.0849**	**0.4797**	**0.7727**	**0.6212**	**0.2344**	1.7675	4.1753	**0.5134**	**0.7628**	**0.6176**
Emotion6	CPNN	0.3495	1.7203	3.9554	0.7399	0.6544	0.5369	0.3561	1.6734	3.8118	0.6627	0.6684	0.5498
	BCPNN	0.3542	1.6921	3.8667	0.7235	0.6591	0.5372	0.2983	1.6639	3.7135	0.5429	0.7454	0.6074
	ACPNN	**0.3452**	**1.6700**	**3.7845**	**0.6363**	**0.6833**	**0.5575**	**0.2799**	**1.6540**	**3.6909**	**0.5057**	**0.7658**	**0.6214**
Flickr_LDL	CPNN	0.4192	2.1369	5.3584	0.9399	0.5894	0.4472	0.3874	2.1315	5.3346	0.8315	0.6508	0.4847
	BCPNN	0.3097	2.1337	5.2701	0.6550	0.7676	0.5842	0.2520	**2.1089**	**5.1561**	0.4732	0.8380	0.6550
	ACPNN	**0.2991**	**2.1196**	**5.2022**	**0.6156**	**0.7857**	**0.5997**	**0.2462**	2.1158	5.1791	**0.4686**	**0.8397**	**0.6617**
Twitter_LDL	CPNN	0.5177	2.4134	6.3725	1.2089	0.5250	0.3604	0.3873	2.4044	6.2942	0.8543	0.7332	0.5050
	BCPNN	0.3555	2.4034	6.2696	0.7815	0.7794	0.5476	0.2828	**2.3866**	**6.1661**	**0.5550**	**0.8499**	0.6422
	ACPNN	**0.3500**	**2.4016**	**6.2529**	**0.7540**	**0.7874**	**0.5628**	**0.2781**	2.4035	6.2453	0.5835	0.8432	**0.6481**

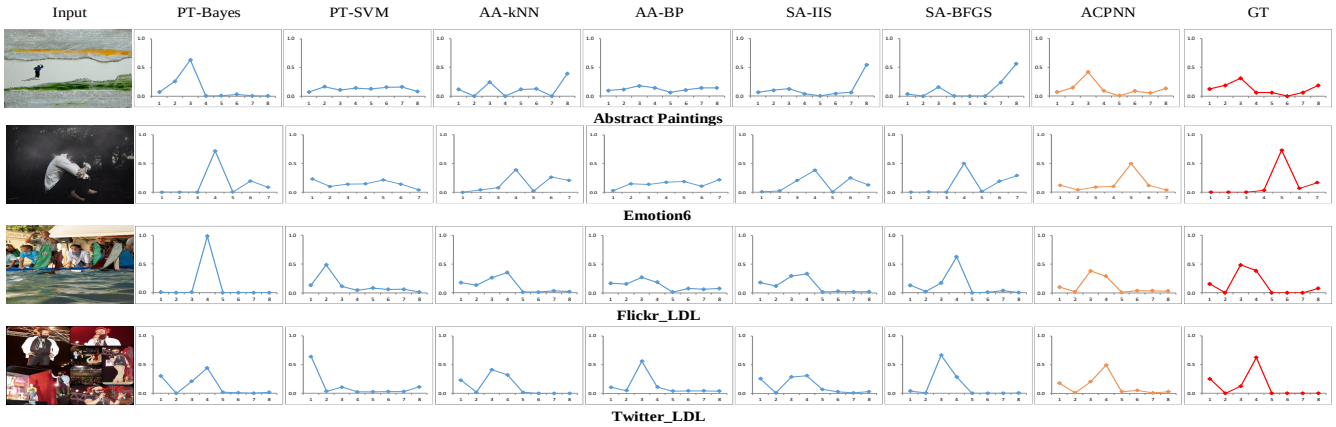


Figure 4: Predicted affective label distributions using our proposed ACPNN and the state-of-the-art approaches. Images and the corresponding ground truth distributions are shown in the first and last columns, respectively. "GT" indicates ground truth.

Experiments

Implementation Details

We execute our experiments on four datasets which have multiple affective labels, including Abstract Paintings (Machajdik and Hanbury 2010), Emotion6 (Peng et al. 2015), and the proposed Flickr_LDL and Twitter_LDL. While Emotion6 labels images in a new definition system containing seven kinds of emotions {*Anger, Disgust, Fear, Joy, Sadness, Surprise and Neutral*}, the other datasets use the typical eight emotions space. In similar fashion with previous works, we randomly select 80% of images as training set and the others for testing.

As SentiBank (Borth et al. 2013) has shown its superiority to low-level features, we use it to extract mid-level features in our experiments. Meanwhile, deep features extracted with VGGNet (Simonyan and Zisserman 2015) are also applied. For each image, we use the last fully connected layer output as the sentiment representation and reduce it to 280 dimensions using principle component analysis (PCA). Table 2 demonstrates the effects of the two kinds of features.

We compare our ACPNN and BCPNN with seven state-of-the-art approaches, including CPNN, PT-Bayes, PT-SVM, AA-kNN, AA-BP, SA-IIS and SA-BFGS (Geng, Yin, and Zhou 2013). For fair comparison, the numbers of hidden layer units of CPNN, BCPNN and ACPNN are set to the same value 100. Since evaluation measures for single-label classification are not applicable, we employ Chebyshev distance (Cheb), Clark distance (Clark), Canberra metric (Canber), Kullback-Leibler divergence (KLdiv), cosine coefficient (Cosine), and intersection similarity (Intersec) to evaluate the methods in this work, following the same routine of existing LDL algorithms (Cha 2007; Geng, Yin, and Zhou 2013; Geng 2016). That is, we evaluate the performance of LDL by computing the similarity or distance between the predicted label distribution and the real label distribution. In these measures, the first four are distance measures and the last two are similarity measures. As shown in Table 2, 3, the "↓" after the distance measures indicates "the smaller the better", and the "↑" after the similarity measures indicates "the larger the better".

Table 3: Experimental Results on four datasets are shown as mean (rank). Since each of the measures may reflect a certain aspect of an algorithm, "Average Rank" is used to indicate the overall performance. In these experiments, VGGNet is employed to extract deep features. For each measure, the best performance is highlighted by boldface.

Dataset	Criterion	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS	ACPNN
Abstract	Cheb ↓	0.3595(6)	0.2982(5)	0.2451(2)	0.2969(4)	0.2961(3)	0.4722(7)	**0.2344(1)**
	Clark ↓	1.9067(6)	1.7884(3)	**1.7567(1)**	1.8140(4)	1.8673(5)	2.3589(7)	1.7675(2)
	Canber↓	4.7501(6)	4.2884(3)	**4.1063(1)**	4.3479(5)	4.5087(4)	6.2439(7)	4.1753(2)
	KLdiv ↓	3.2678(7)	0.7076(4)	0.5154(2)	0.7815(5)	0.6443(3)	2.2815(6)	**0.5134(1)**
	Cosine ↑	0.6531(4)	0.6429(5)	0.7533(2)	0.6364(6)	0.6916(3)	0.5727(7)	**0.7628(1)**
	Intersec ↑	0.5176(6)	0.5393(5)	0.6090(2)	0.5403(4)	0.5771(3)	0.4170(7)	**0.6176(1)**
	Average Rank	5.83	4.17	1.67	4.67	3.50	6.83	**1.33**
Emotion6	Cheb ↓	0.3581(5)	0.3835(6)	0.2971(2)	0.3384(3)	0.3392(4)	0.3998(7)	**0.2799(1)**
	Clark ↓	1.9368(6)	1.7779(5)	**1.6260(1)**	1.6794(3)	1.7593(4)	1.9600(7)	1.6540(2)
	Canber↓	4.5906(6)	4.1740(5)	**3.5440(1)**	3.8244(3)	4.0494(4)	4.6715(7)	3.6909(2)
	KLdiv ↓	2.7178(7)	0.9286(4)	1.8683(6)	0.6832(2)	0.6937(3)	1.2111(5)	**0.5057(1)**
	Cosine ↑	0.6826(3)	0.5445(7)	0.7088(2)	0.6733(5)	0.6793(4)	0.6118(7)	**0.7658(1)**
	Intersec ↑	0.5473(5)	0.4554(7)	0.5965(2)	0.5498(4)	0.5644(3)	0.5133(6)	**0.6214(1)**
	Average Rank	5.33	5.67	2.33	3.33	3.67	6.33	**1.33**
Flickr_LDL	Cheb ↓	0.4336(6)	0.4647(7)	0.2552(2)	0.3227(5)	0.3016(3)	0.3133(4)	**0.2462(1)**
	Clark ↓	2.1309(3)	2.2333(6)	**1.7143(1)**	2.1360(4)	2.1960(5)	2.2647(7)	2.1158(2)
	Canber↓	5.3584(4)	5.7418(6)	3.6099(1)	5.1559(2)	5.5495(5)	5.7963(7)	5.1791(3)
	KLdiv ↓	1.0086(5)	1.1680(6)	3.7965(7)	0.7052(3)	0.6302(2)	0.7593(4)	**0.4686(1)**
	Cosine ↑	0.5619(6)	0.4443(7)	0.8097(2)	0.7498(5)	0.7722(3)	0.7547(4)	**0.8397(1)**
	Intersec ↑	0.4225(6)	0.3981(7)	0.6567(2)	0.5823(5)	0.6079(3)	0.6035(4)	**0.6617(1)**
	Average Rank	5.00	6.50	2.50	4.00	3.50	5.00	**1.50**
Twitter_LDL	Cheb ↓	0.5433(6)	0.6167(7)	0.2896(2)	0.3189(4)	0.2975(3)	0.3578(5)	**0.2781(1)**
	Clark ↓	2.4028(2)	2.5392(7)	**1.4929(1)**	2.4072(4)	2.4190(6)	2.5063(5)	2.4035(3)
	Canber↓	6.1905(2)	6.9866(7)	**2.7822(1)**	6.2677(4)	6.3051(5)	6.6617(6)	6.2453(3)
	KLdiv ↓	1.3221(4)	1.6761(6)	5.2344(7)	0.7007(3)	0.6404(2)	1.1326(5)	**0.5835(1)**
	Cosine ↑	0.4967(6)	0.3067(7)	0.8125(3)	0.8116(4)	0.8205(2)	0.7341(5)	**0.8432(1)**
	Intersec ↑	0.3423(6)	0.2585(7)	**0.6562(1)**	0.5968(4)	0.6347(3)	0.5893(5)	0.6481(2)
	Average Rank	4.33	6.83	2.50	3.83	3.50	5.17	**1.83**

Results and Analysis

Since our proposed ACPNN and BCPNN are built based on the typical CPNN algorithm, all these methods are applied to four datasets and evaluated by the six measures. Table 2 shows the superiority of our methods, where both mid-level and deep features are extracted and the latter performs sightly better in our experiments. The reason mid-level features generate comparable results with deep features is that SentiBank (Borth et al. 2013) is a well-designed affective representation, while VGGNet is built for object recognition. The performance will be further improved if a special network for affective prediction, e.g. progressive CNN (You et al. 2015b), is applied to extract deep features. Note although the datasets vary on scale and emotional categories, ACPNN has shown its superiority to BCPNN and both of them outperform CPNN.

Furthermore, we compare ACPNN with six working methods which employ three strategies, i.e. problem transformation, algorithm adaptation, and specialized algorithm design. The results are shown in Table 3. For fair comparison, deep features extracted from VGGNet are applied to all approaches, and the best performances are highlighted by boldface. The ranks are given in the parentheses right after the measure values, and the average ranks are given in the last row of each subtable. As can be seen from Table 3, for each particular dataset, ACPNN performs the best, while the rankings of other six LDL algorithms are often different on different measures. AA-kNN performs better than others because AA-kNN keeps the label distribution and thus keeps the overall labeling structure for each instance, while others break down the original label distributions by weighted resampling. Figure 4 shows the predicted affective label distributions on images from various datasets, in which ACPNN generates the most similar distributions as the ground truth.

Conclusion

In this work, we introduce a challenging problem of predicting multi-emotions evoked in images. Due to the absence of large-scale datasets with multiple affective labels, we build two new datasets named Flickr_LDL and Twitter_LDL, where LDL indicates we employ label distribution learning to address the problem. Based on the state-of-the-art conditional probability neural network (CPNN), we encode its input label into a binary vector (BCPNN), and then develop our ACPNN by augmenting distributions with label noises. Mid-level features and deep features are employed on Abstract Paintings, Emotion6, Flickr_LDL and Twitter_LDL in our experiments, demonstrating ACPNN is superior to BCPNN, and both of them outperform CPNN and other contrastive methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61301238, 61201424), China Scholarship Council (No.201506205024) and the Natural Science Foundation of Tianjin, China (No. 14ZCDZGX00831).

References

[Bengio, Weston, and Grangier 2010] Bengio, S.; Weston, J.; and Grangier, D. 2010. Label embedding trees for large multi-class tasks. In *NIPS*, 163–171. 2

[Borth et al. 2013] Borth, D.; Ji, R.; Chen, T.; Breuel, T.; and Chang, S. F. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 223–232. 2, 3, 5, 6

[Cha 2007] Cha, S. H. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models Methods in Applied Sciences* 1(4):300–307. 5

[Chen et al. 2014a] Chen, T.; Borth, D.; Darrell, T.; and Chang, S. F. 2014a. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*. 2

[Chen et al. 2014b] Chen, T.; Yu, F. X.; Chen, J.; Cui, Y.; Chen, Y. Y.; and Chang, S. F. 2014b. Object-based visual sentiment concept analysis and application. In *ACM MM*, 367–376. 2

[Chen et al. 2015] Chen, Y.-Y.; Chen, T.; Liu, T.; Liao, H.-Y. M.; and Chang, S.-F. 2015. Assistive image comment robota novel mid-level concept-based representation. *IEEE Transactions on Affective Computing* 6(3):298–311. 1

[Geng, Yin, and Zhou 2013] Geng, X.; Yin, C.; and Zhou, Z. H. 2013. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(10):2401–2412. 1, 2, 3, 5

[Geng 2016] Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28(7):1734–1748. 1, 5

[Joshi et al. 2011] Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q. T.; Wang, J. Z.; Li, J.; and Luo, J. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28(5):94–115. 1

[Jou et al. 2015] Jou, B.; Chen, T.; Pappas, N.; Redi, M.; Topkara, M.; and Chang, S.-F. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *ACM MM*, 159–168. 1

[Lang, Bradley, and Cuthbert 1998] Lang, P. J.; Bradley, M. M.; and Cuthbert, B. N. 1998. Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. *Biological Psychiatry* 44(12):1248–1263. 1

[Lang, Bradley, and Cuthbert 1999] Lang, P.; Bradley, M.; and Cuthbert, B. 1999. International affective picture system (iaps): Technical manual and affective ratings. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*. 2

[Lang 1979] Lang, P. J. 1979. A bio-informational theory of emotional imagery. *Psychophysiology* 16(6):495–512. 1

[Machajdik and Hanbury 2010] Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 83–92. 1, 2, 5

[Modha and Fainman 1994] Modha, D. S., and Fainman, Y. 1994. A learning law for density estimation. *IEEE Transactions on Neural Networks* 5(3):519–523. 4

[Nicolaou, Gunes, and Pantic 2011] Nicolaou, M. A.; Gunes, H.; and Pantic, M. 2011. A multi-layer hybrid framework for dimensional emotion classification. In *ACM MM*, 933–936. 2

[Peng et al. 2015] Peng, K. C.; Chen, T.; Sadovnik, A.; and Gallagher, A. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 860–868. 1, 2, 3, 5

[Plutchik 1980] Plutchik, R. 1980. A general psychoevolutionary theory of emotion. *Theories of Emotion* 1:3–33. 2

[Riedmiller and Braun 1993] Riedmiller, M., and Braun, H. 1993. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *ICNN*, 586–591. 4

[Simonyan and Zisserman 2015] Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*. 5

[Solli and Lenz 2009] Solli, M., and Lenz, R. 2009. Color based bags-of-emotions. In *ICCAIP*, 573–580. 2

[Sun et al. 2016] Sun, M.; Yang, J.; Wang, K.; and Shen, H. 2016. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *ICME*, 1–6. 1

[Xu et al. 2008] Xu, M.; Jin, J. S.; Luo, S.; and Duan, L. 2008. Hierarchical movie affective content analysis based on arousal and valence features. In *ACM MM*, 677–680. 2

[You et al. 2015a] You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015a. Joint visual-textual sentiment analysis with deep neural networks. In *ACM MM*, 1071–1074. 1

[You et al. 2015b] You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015b. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 25–30. 2, 6

[You et al. 2016] You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*. 1, 2

[Zhang and Wu 2015] Zhang, M.-L., and Wu, L. 2015. Lift: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(1):107–120. 2

[Zhao et al. 2014] Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T. S.; and Sun, X. 2014. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 47–56. 2

[Zhou and Zhang 2006] Zhou, Z.-H., and Zhang, M.-L. 2006. Multi-instance multi-label learning with application to scene classification. In *NIPS*, 1609–1616. 1