Structured and Sparse Annotations for Image Emotion Distribution Learning

Haitao Xiong,[1] Hongfu Liu,[2] Bineng Zhong,[3] Yun Fu[4]
[1]School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, China
[2]Michtom School of Computer Science, Brandeis University, Waltham, USA
[3]chool of Computer Science and Technology, Huaqiao University, Xiamen, China
[4]Department of Electrical and Computer Engineering, Northeastern University, Boston, USA
xionghaitao@btbu.edu.cn, hongfuliu@brandeis.edu, bnzhong@hqu.edu.cn, yunfu@ece.neu.edu

Abstract

Label distribution learning methods effectively address the label ambiguity problem and have achieved great success in image emotion analysis. However, these methods ignore structured and sparse information naturally contained in the annotations of emotions. For example, emotions can be grouped and ordered due to their polarities and degrees. Meanwhile, emotions have the character of intensity and are reflected in different levels of sparse annotations. Motivated by these observations, we present a convolutional neural network based framework called Structured and Sparse annotations for image emotion Distribution Learning (SSDL) to tackle two challenges. In order to utilize structured annotations, the Earth Mover's Distance is employed to calculate the minimal cost required to transform one distribution to another for ordered emotions and emotion groups. Combined with Kullback-Leibler divergence, we design the loss to penalize the mis-predictions according to the dissimilarities of same emotions and different emotions simultaneously. Moreover, in order to handle sparse annotations, sparse regularization based on emotional intensity is adopted. Through combined loss and sparse regularization, SSDL could effectively leverage structured and sparse annotations for predicting emotion distribution. Experiment results demonstrate that our proposed SSDL significantly outperforms the state-of-the-art methods.

Introduction

For the reason that an image can contain rich semantics, many people tend to use images to express their opinions and emotions. So, it is important to find out the emotions implied in images, especially for social network sites like Twitter and so on (Zhao et al. 2016; Yang, She, and Sun 2017; Zhao et al. 2018b). In traditional image emotion analysis, an image is in general associated with one or more emotion labels, which belong to single-label and multi-label learning methods (Zhang and Zhou 2014). However, there exists an ambiguity problem for this task, which means the uncertainty of the ground-truth label (Yang, Sun, and Sun 2017; Zhao et al. 2018a). Moreover, images often express mixture of different emotions and different people may have totally different emotional feelings. Through single-label and multi-label learning, the relative importance of different emotions in description of the image is not clear. For this reason, label distribution learning methods are designed to reveal the extent to which each label describes the samples through label distribution prediction and have achieved great success (Geng 2016; Yang, Sun, and Sun 2017; Zhao et al. 2018a). Label distribution learning can provide an alternate view of learning process, where each sample is mapped to a label distribution instead of a single label or multiple labels. Recently, convolutional neural networks (CNNs) based label distribution learning methods have shown superior performance of distribution prediction against traditional label distribution learning methods (Yang, She, and Sun 2017).

Unfortunately, most of label distribution learning based image emotion analysis methods rarely utilize characters of emotions. On the one hand, emotions have the character of polarity and are presented in varying degrees (Mikels et al. 2005). This means that emotions can be grouped and ordered. Ordered emotions and emotion groups are presented in structured annotations. Here we take seven emotions in Emotion6 dataset (Peng et al. 2015) for example. These emotions have ordered sequence according to the degrees of positive or negative polarity. Besides, these emotions can be divided into three groups with different polarities, which are *positive group*, *neutral group* and *negative group*. In image emotion analysis, mis-prediction of one emotion as nearby emotion, such as *joy* and *surprise*, is more adorable than as far away emotion, such as *joy* and *sadness*. Figure 1 shows polarity and group information in seven emotions from Emotion6 dataset. Red-based and gray-based colors are used to represent different degrees of positive and negative polarity, while white color is used to represent neutral polarity. On the other hand, emotions also have the character of intensity (Sheppes et al. 2014) and are reflected in different levels of sparse annotations. Figure 2 shows three images from Flickr_LDL dataset (Yang, Sun, and Sun 2017) and the emotion distribution line chart of these images. Flickr_LDL has eight emotions which are *amusement*, *contentment*, *awe*, *excitement*, *fear*, *sadness*, *disgust* and *anger*. The first four belong to the *positive group* and the last four belong to the *negative group*. In the line chart, emotions are in ordered sequence from emotion *amusement* to emotion *sadness*. The top left image shows strong emotional intensity on single emotion leading to non-zero sparse annotations of emotions. The top right and bottom left image show group strong emotional intensity on emotions from *posi-*

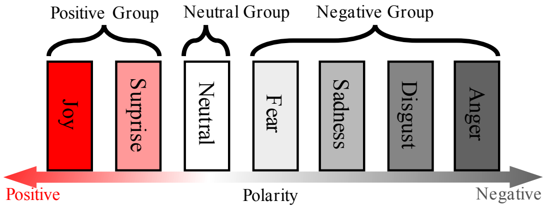

Figure 1: Polarity and group information in seven emotions from Emotion6 dataset.

tive group resulting in non-zero group sparse annotations of emotion groups. Emotional intensity of the bottom left image is weaker than the top right image, because there exist one emotion in the *positive group* is not shown in the bottom left image resulting in spare annotations of emotions in the *positive group*. As shown in Figure 1 and Figure 2, there exist polarity and intensity characters among the emotions, which refers to structured and sparse information in annotations of emotions. However, most label distribution learning methods use Kullback-Leibler (KL) divergence as loss to measure the similarity between the ground-truth and the predicted label distribution (Yang, She, and Sun 2017; Geng 2016). Consequently, these methods can hardly capture the polarity and intensity characters of images emotion, which makes it difficult to learn precise emotional representations for explaining the image emotions. Therefore, how to effectively use image emotional characters in label distribution learning still remains a big challenge.

In this paper, we develop a deep model to handle the above challenges. For utilization of structured annotations in emotional polarity, the Earth Mover's Distance (EMD) is employed to calculate the minimal cost required to transform one distribution to another for ordered emotions and emotion groups. Then we design the combined loss based on EDM and KL divergence to penalize the mis-predictions according to the dissimilarities of same emotions and different emotions simultaneously. Moreover, we design the sparse regularization based on emotional intensity to handle different levels of sparse annotations. Through combined loss and sparse regularization, structured and sparse annotations could be effectively leveraged for learning. Our contributions are summarized as follows. First, we design a novel CNN-based framework called Structured and Sparse annotations for image emotion Distribution Learning (SSDL) to analyze image emotions, which integrates information of structured and sparse annotations into a CNN based model for effectively learning. Second, structured annotations are effectively utilized via combining the EMD and KL divergence in loss. Furthermore, sparse regularization based on emotional intensity is designed to handle the sparse annotations of emotions. Experimental results demonstrate the superior results of SSDL compared with several state-of-the-art methods in image emotion distribution learning.

Related Work

Image Emotion Classification

Through the development of computer vision, image emotion classification has been an interesting and meaningful research topic recently. The solutions mainly concentrate on two kinds of approaches: dimensional models and categorical models. Dimensional models use a few basic spaces for emotion description, and the categorical models classify emotion into one of the typical emotion categories, which are obvious for common people understanding and thus have been mainly used by most previous work. According to that whether the image can be associated with one or more emotion labels, categorical models can be divided into single-label and multi-label classification (Cambria et al. 2017; Zhang and Zhou 2014). Feature extraction is an important factor that affects the analysis performance and low-level features of image are most used (Zhao et al. 2014b). These features can be generated automatically, but lack in describing emotions. So hand-crafted features based on art and psychology theory are proposed and show better performance than low-level features in several situation, such as emotions in art images (Zhao et al. 2014a). Due to the different benefits of these features, fused multi-modal features are generated for image emotion analysis (Zhao et al. 2017).

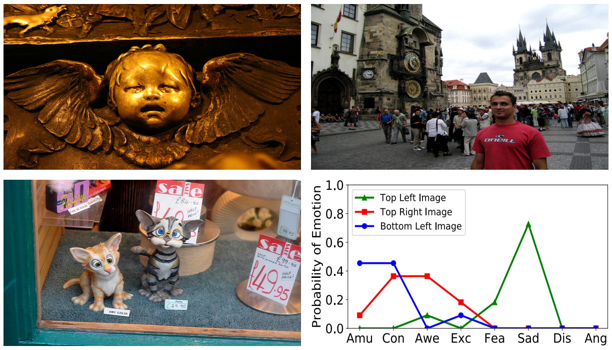

Figure 2: Three images from the Flickr_LDL dataset are annotated in eight ordered emotions. The line chart on bottom right shows emotion distributions of these images.

On the other side, deep models like CNNs have shown strong ability to generate high-level features automatically (Chen, Zhang, and Allebach 2015; Zhao et al. 2018c). Meanwhile, these features show good performance in many computer vision tasks (Wang et al. 2016). So, CNN-based methods have been frequently employed in emotion classification and show great success. In order to use large scale yet noisy training data to, solve the emotion prediction, You *et al.* designed a robust CNN-based model called Progressive CNN (PCNN) for visual sentiment analysis and got obvious improvement (You et al. 2015). Architecture-Frame-Transformer Emotion Classification Network (FT-EC-net) was proposed in (Tripathi et al. 2017) to solve three highly correlated emotion analysis tasks: emotion recognition, emotion attribution and emotion-oriented summarization. However, these work did not consider structured annotations of ordered emotions.

Label Distribution Learning

Label distribution learning was proposed by Ref. (Geng 2016) to reveal the extent to which each label describes the samples through label distribution prediction, and has been a hot topic in machine learning. Generally speaking, current label distribution learning methods can be roughly generalized into three strategies which are problem transformation, algorithm adaptation and designing specialized algorithms (Geng 2016). Representative methods build on problem transformation strategy include, PT-SVM and PT-Bayes, which come from traditional classification methods SVM and Naive Bayes (Geng 2016). Through algorithm adaptation strategy, classification algorithms, such as kNN and backpropagation (BP) neural network, can be naturally extended to deal with label distributions (Geng 2016). Geng *et al.* adopted designing specialized algorithms strategy to propose SA-IIS, SA-BFGS and SA-CPNN for label distribution learning (Geng, Yin, and Zhou 2013; Geng 2016).

Recently, with the development of CNNs, CNN-based methods are proposed for label distribution learning. Deep label distribution learning (DLDL) method utilized the label ambiguity in both feature learning and classifier learning (Gao et al. 2017). Experiments showed that even when the training set is small, DLDL could still achieve better performance for age estimation. Specially for image emotion analysis, convolutional neural network regressions (CNNR) based on CNN with Euclidean loss for each emotion was presented in work of Ref. (Peng et al. 2015). Through normalization, the regression results were transformed to probabilities of all emotions. Yang *et al.* proposed a deep CNN approach through joint optimization of image emotion classification and distribution learning , and the model showed better performance (Yang, She, and Sun 2017).

Problem Definition

Given an input image, we are interested in predicting emotion distribution considering structured and sparse annotations of image emotions. Assume that we have C emotions $e = \{e_1, e_2, ..., e_C\}$, and N images for training $x = \{x_1, x_2, ..., x_N\}$. According to different emotional polarities, e can be divided into R emotion groups $g = \{g_1, g_2, ..., g_R\}$. For consideration of structured annotations of ordered emotions and emotion groups, the comparison operator between different emotions and emotion groups should be defined firstly. Suppose e_i and e_j are different emotions, we define $e_i \prec e_j$ which means the positive polarity degree of e_i is less than e_j, and also means the negative polarity degree of e_i is more than e_j. Then we can define $g_m \prec g_n$ which means the positive polarity degree of each emotion in g_m is less than each emotion in g_n. Through comparison operator, image emotions can be reordered. If there is no special explanation, e are ordered emotions and g are ordered emotion groups. On the other hand, in order to express sparse annotations of different levels of emotional intensity, we define i_s as single strong emotional intensity, i_g as group strong emotional intensity and i_w as weakly strong emotional group intensity. So i_s, i_g and i_w are collections of images from x_n and $i_s \cap i_g \cap i_w = \varnothing$.

The probability distribution $d_{x_n} = \{d_{x_n}^{e_1}, d_{x_n}^{e_2}, ..., d_{x_n}^{e_C}\}$ of emotions for the image x_n is regarded as the emotion distribution, where $d_{x_n}^{e_c}$ is the probability of emotion e_c for image x_n and represents the extent to which emotion e_c describes x_n. $d_{x_n}^{e_c}$ is under the constraints $d_{x_n}^{e_c} \geqslant 0$ and $\sum_{c=1}^{C} d_{x_n}^{e_c} = 1$ which mean that the emotion probability is non-negative and e can describe the emotions of the image fully. Emotion group distribution $d_{x_n}^{g}$ can be got by adding probabilities of all emotions that belong to the same emotion group.

Let us denote by $f(x_n; w)$ the activation values of the last fully connected layer for image x_n. And w is the weight. Specially, $f_{e_c}(x_n; w)$ is the activation value for emotion e_c. In order to reveal structured annotations representing by emotion groups, $f_{g_r}(x_n; w)$ is defined as the sum of activation values for all emotions in group g_r. The network is trained by minimizing the following objective function:

$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} L(d_{x_n}, f(x_n; w)) + R(w), \qquad (1)$$

where $L(\cdot, \cdot)$ is the suitable loss function for image emotion distribution learning, and $R(\cdot)$ is the proper regularization.

In image emotion distribution learning, the purpose of our proposed method is using structured sparse annotations to capture polarity and intensity characters for predicting emotion distribution. In order to consider characters of image emotions in learning, we propose a deep CNN-based framework that can extract and integrate structured and sparse annotations from polarity and intensity characters for learning. For challenge of reflecting image emotional characters in modeling of label distribution learning properly, structured and sparse annotations are used in the loss function and regularization for image emotion distribution learning.

Image Emotion Distribution Learning

As illustrated in Figure 3, we design a deep CNN-based framework called Structured and Sparse annotations for image emotion Distribution Learning (SSDL), to utilize polarity and intensity characters of emotions for learning. In the SSDL framework, image emotions are reordered according to the emotional character of polarity, then images are fed into a pre-trained CNN model with some modifications. The number of outputs of last fully-connected layer, which is classification layer, is assigned to the number of emotions. Through information of emotional polarity, SSDL extracts structured annotations from ordered emotions and emotion groups, which are used for EMD-based loss computing. Then EMD-based loss is combined with KL loss for penalizing the mis-predictions according to the dissimilarities of same emotions and different emotions simultaneously. In addition, through information of emotional intensity based on ground-truth distributions, SSDL uses sparse annotations of different levels of emotional intensity for sparse regularization. Finally, combined loss and sparse regularization are both used for distribution learning optimization.

Combined Loss for Structured Annotations

As a distance measure, KL divergence is widely used as training loss called KL loss. KL loss can measure the similarity between the ground-truth and emotion distribution prediction, which is used widely in label distribution learning. But it ignores emotions' group and polarity ranking

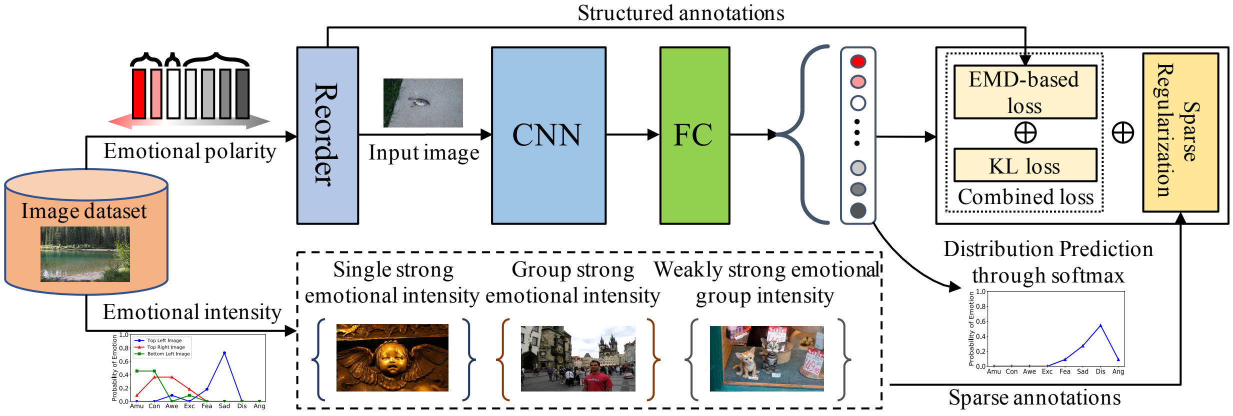

Figure 3: The deep framework of our proposed SSDL. SSDL extracts structured annotations from emotional polarity for EMD-based loss computing. Then EMD-based loss is combined with KL loss for penalizing the mis-predictions according to the dissimilarities of same emotions and different emotions simultaneously. Meanwhile, SSDL also uses sparse annotations of different levels of emotional intensity for sparse regularization.

which could lead to misclassification. Simply suppose there exist three ordered emotions $e = \{e_1, e_2, e_3\}$. The ground-truth distribution $d = \{0.6, 0.2, 0.2\}$, and two predicted distributions are $\hat{d}_1 = \{0.5, 0.2, 0.3\}$ and $\hat{d}_2 = \{0.5, 0.3, 0.2\}$. The KL loss values between d, $\hat{d}_1$ and d, $\hat{d}_2$ are same. However, in the case of image emotion analysis, the mis-prediction of e_1 as e_2 and e_3 should be different, and the mis-prediction as e_2 is more adorable because of the more similarity between e_1 and e_2. To handle the above challenge, EMD is employed which is defined as the minimum cost to transport the mass of one distribution to the other (Levina and Bickel 2001) and has shown better performance than softmax cross-entropy loss in ordinal classification problems.

Firstly, for the reason of considering structured annotations of ordered emotions, EMD can be used to penalize mis-predictions of distribution according to distances of different emotions. As shown in problem definition, image emotions are ordered as $e_1 \prec e_2 \prec ... \prec e_C$, and we define the distance between different emotions e_i and e_j as $|i - j|$. Given the ground-truth distribution d and predicted distribution $\hat{d}$, the expression of EMD for emotions is defined as:

$$\mathrm{EMD}_e(d, \hat{d}) = (\frac{1}{C}\sum_{k=1}^{C} \left|\mathrm{CDF}_d(k) - \mathrm{CDF}_{\hat{d}}(k)\right|^q)^{\frac{1}{q}}, \quad (2)$$

where $\mathrm{CDF}_d(k)$ means the cumulative distribution function $\sum_{c=1}^{k} d^{e_c}$. Secondly, we also want to penalize mis-predictions of distribution between different groups. Ordered emotions e can be divided into R ordered emotion groups g as $g_1 \prec g_2 \prec ... \prec g_R$. The distance between different emotion groups g_i and g_j is defined as $|i - j|$. The expression of EMD for emotion groups is defined as:

$$\mathrm{EMD}_g(d^g, \hat{d}^g) = (\frac{1}{R}\sum_{k=1}^{R} \left|\mathrm{CGDF}_{d^g}(k) - \mathrm{CGDF}_{\hat{d}^g}(k)\right|^q)^{\frac{1}{q}}, \quad (3)$$

where $\mathrm{CGDF}_{d^g}(k)$ means the cumulative group distribution function $\sum_{r=1}^{k}\sum_{e_c \in g_r} d^{e_c}$. Finally, we use the EMD loss of emotions and emotion groups to generate the EMD-based loss for image emotion distribution learning:

$$L_{EMD} = \frac{1}{C}\mathrm{EMD}_e(d, \hat{d}) + \frac{1}{R}\mathrm{EMD}_g(d_g, \hat{d}_g), \quad (4)$$

where C is the number of emotions and R is the number of emotion groups.

For penalizing the mis-predictions according to the dissimilarities of same emotions and different emotions simultaneously, loss function L in SSDL combines EMD-based loss and KL loss through a weighted combination:

$$L = \mu L_{EMD} + (1 - \mu) L_{KL}, \quad (5)$$

where L_{KL} is KL loss and weight μ is used to adjust the importance of the two components in combined loss.

Regularization based on Emotional Intensity

Traditional ℓ_1 and ℓ_2 regularizations are mainly used for preventing over-fitting of training data. And ℓ_1 regularization called Lasso (Candès and Wakin 2008) has the ability of producing sparse outputs at single-level. In order to get sparse outputs at group-level, group Lasso are proposed and show better performance in weight sparse learning (Baldassarre et al. 2016). Since group Lasso loses the guarantee of sparsity at single-level, it may still be sub-optimal. To address this problem, sparse group Lasso considers sparsity at both single-level and group-level (Scardapane et al. 2017). As we known, images show different levels of emotional intensity reflected in sparse annotations. In order to reveal different levels of emotional intensity, we present sparse regularization using Lasso, group Lasso and sparse group Lasso.

Figure 4 shows visual comparison between Lasso, group Lasso, and sparse group Lasso penalizations. Four emotions on left come from *positive group* and the other four emotions on right come from *negative group*. We show a possible combination of emotions that are zeroed (white squares)

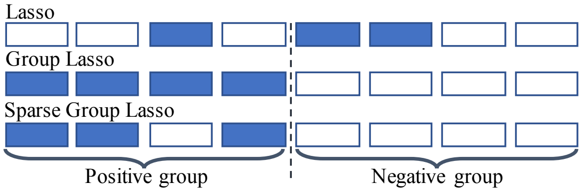

Figure 4: Comparison between Lasso, group Lasso, and sparse group Lasso for prediction. Blue square means the predicted probability of the emotion is non-zero.

and non-zeroed (blue squares) out by the corresponding penalization. In Lasso example, three emotions are predicted in sparsity by removing emotions with optimizing single-level emotion considerations. Through group lasso, whole emotions in *positive group* are predicted and none of emotions in *negative group* are predicted. Sparse group Lasso example combines the usages of Lasso and group Lasso, and gets sparsity at single-level and group-level simultaneously.

The basic idea of sparse regularization based on emotional intensity is considering different emotion sparsities for sparse annotations of different levels of emotional intensity. Unlike ℓ_1 at single-level, emotion sparsity at group-level forces all probabilities of emotions from the same group to be either all non-zeros, or all zeros. Specifically, we consider three different levels of emotional intensity:

- **Single strong emotional intensity** (i_s): as the emotions of top left image in Figure 2, there exists only single prominent emotion that an image shows, which can be understood as that the probability of this emotion exceeds the threshold and few emotions can describe the image;
- **Group strong emotional intensity** (i_g): group strong emotional intensity means that emotions of the image are focused on all emotions from one emotion group, like the intensity of top right image in Figure 2.
- **Weakly strong emotional group intensity** (i_w): not like the strict requirement of group strong emotional intensity, weakly strong emotional group intensity shows partial emotions from one emotion group, as the bottom right image in Figure 2 shows.

For *single strong emotional intensity*, Lasso R_{ℓ_1} penalizes the absolute magnitude of the predicted probabilities of all emotions, and is calculated as:

$$R_{\ell_1} = \|f(x_n, w)\|_1 = \sum_{c=1}^{C} |f_{e_c}(x_n, w)|. \tag{6}$$

For *group strong emotional intensity*, group Lasso aiming at keeping the group sparse structure is suitable (Hu et al. 2017). The group sparsity of the predicted probabilities of all emotions with group structure g can be measured through $\ell_{2,1}$ norm. Group Lasso $R_{\ell_{2,1}}$ is calculated as:

$$R_{\ell_{2,1}} = \sum_{r=1}^{R} \|f_{g_r}(x_n, w)\|_2 = \sum_{r=1}^{R} \sqrt{\sum_{e_c \in g_r} (f_{e_c}(x_n, w))^2}. \tag{7}$$

For *weakly strong emotional group intensity*, only using Lasso or group Lasso cannot handle this emotional intensity. Sparse group Lasso keeps the group sparsity structure. At the same time, it permits single sparsity. Sparse group Lasso R_{sgl} is calculated by adding R_{ℓ_1} to $R_{\ell_{2,1}}$.

Sparse regularization R_{sr} uses Lasso, group Lasso and sparse group Lasso simultaneously for different levels of emotional intensity. The equation of R_{sr} is given below:

$$R_{sr} = \sum_{x_n \in i_s} R_{\ell_1} + \sum_{x_n \in i_g} R_{\ell_{2,1}} + \sum_{x_n \in i_w} R_{sgl}. \tag{8}$$

Moreover, ℓ_2 regularization $\|w\|_2$ is used to prevent overfitting. The overall regularization R in SSDL is given below:

$$R = \xi_1 R_{sr} + \xi_2 R_{\ell_2}(w), \tag{9}$$

where ξ_1 and ξ_2 are the regularization parameters to balance the importance of the two components in objective function.

Experiments and Results

Implementation Details

To make this comparison, three image emotion distribution datasets are selected, including Emotion6 (Levi and Hassner 2015), Flickr_LDL and Twitter_LDL (Yang, Sun, and Sun 2017). Emotion6 is widely used as a benchmark dataset for emotion classification, which contains 1,980 images collected from Flickr. Flickr_LDL and Twitter_LDL are two datasets collected mainly for emotion distribution learning. Flickr_LDL contains 11,150 images and Twitter_LDL contains 10,045 images. In Emotion6, used emotions can be divided into three groups: *positive*, *negative* and *neutral group*. In Flickr_LDL and Twitter_LDL, used emotions can be divided into two groups: *positive* and *negative group*. All datasets are randomly split into 75% training, 20% testing and 5% validation sets. The validation set is used for choosing the best parameters of our methods. The votes from the annotators are integrated to generate the ground-truth of image emotion distribution d through dividing votes for each emotions by total votes. If the probability of a single emotion exceeds 0.6, these images have single strong emotional intensity. For the rest images, if the emotions shown in the image come from the same emotion group and all emotions in the group are shown, this image have group strong emotional intensity; if the emotions shown in the image come from the same emotion group but not all emotions in the group are shown, this images have weakly strong emotional group intensity.

In the experiments, SSDL is built on VGG-19 architecture (Simonyan and Zisserman 2014). We change the number of last fully-connected layer outputs to the number of emotions. The original loss layer is replaced by the combined loss. The learning rates of the convolution layers, the first two fully-connected layers and the classification layer are initialized as 0.001, 0.001 and 0.01. We fine-tune all layers by back propagation through the whole net using mini-batches of 32 and the total number of epochs is 20 for learning.

In order to check the superiority of SSDL, we conduct experiments to compare it against several baseline methods. All these methods can be divided into four types:

Table 1: Performance comparison between SSDL and the state-of-the-art methods for emotion distribution learning on Emotion6, Flickr_LDL, Twitter_LDL datasets measured by Chebyshev, Clark, Canberra, KL divergence, Cosine, Intersection, average rank of all measures and accuracy.

Datasets	Measures	PT		AA		SA			CNN-based			
		PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS	SA-CPNN	CNNR	DLDL	JCDL	SSDL
Emotion6	Chebyshev ↓	0.35(9)	0.39(11)	0.29(5)	0.3(6)	0.32(8)	0.38(10)	0.3(6)	0.26(4)	0.25(3)	**0.24(1)**	**0.24(1)**
	Clark ↓	0.73(10)	0.69(9)	0.62(3)	0.64(8)	0.63(6)	0.74(11)	0.63(6)	**0.61(1)**	0.62(3)	**0.61(1)**	0.62(3)
	Canberra ↓	0.66(10)	0.62(9)	0.51(2)	0.54(6)	0.55(8)	0.67(11)	0.54(6)	**0.49(1)**	0.52(5)	0.51(2)	0.51(2)
	KL ↓	2.32(11)	1.07(9)	0.85(8)	0.63(6)	0.61(5)	1.16(10)	0.56(4)	0.67(7)	0.43(3)	0.42(2)	**0.41(1)**
	Cosine ↑	0.69(6)	0.48(11)	0.75(4)	0.68(8)	0.69(6)	0.63(10)	0.66(9)	0.74(5)	0.79(3)	0.8(2)	**0.81(1)**
	Intersection ↑	0.56(9)	0.42(11)	0.62(4)	0.59(8)	0.61(5)	0.52(10)	0.6(6)	0.6(6)	0.65(2)	0.65(2)	**0.67(1)**
	Average Rank	9.17(9)	10(10)	4.33(5)	7(8)	6.33(7)	10.33(11)	6.17(6)	4(4)	3.17(3)	1.67(2)	**1.5(1)**
	Accuracy	0.39(9)	0.37(10)	0.44(5)	0.4(8)	0.41(7)	0.35(11)	0.42(6)	0.45(4)	0.46(3)	0.52(2)	**0.53(1)**
Flickr_LDL	Chebyshev ↓	0.44(10)	0.55(11)	0.28(5)	0.36(8)	0.31(7)	0.37(9)	0.3(6)	0.25(3)	0.25(3)	0.24(2)	**0.23(1)**
	Clark ↓	0.89(11)	0.87(10)	**0.57(1)**	0.82(5)	0.82(5)	0.86(9)	0.82(5)	0.84(8)	0.78(3)	0.77(2)	0.78(3)
	Canberra ↓	0.85(11)	0.83(10)	**0.41(1)**	0.75(7)	0.75(7)	0.82(9)	0.74(6)	0.73(5)	0.7(3)	0.7(3)	0.69(2)
	KL ↓	1.88(10)	1.69(9)	3.28(11)	0.82(7)	0.66(4)	1.06(8)	0.71(6)	0.7(5)	0.54(3)	0.53(2)	**0.46(1)**
	Cosine ↑	0.63(10)	0.32(11)	0.79(4)	0.72(6)	0.78(5)	0.7(8)	0.7(8)	0.72(6)	0.81(3)	0.82(2)	**0.85(1)**
	Intersection ↑	0.49(10)	0.29(11)	0.64(3)	0.53(9)	0.6(6)	0.56(8)	0.6(6)	0.62(5)	0.64(3)	0.65(2)	**0.68(1)**
	Average Rank	10.33(10)	10.33(10)	4.17(4)	7(8)	5.67(6)	8.5(9)	6.17(7)	5.33(5)	3(3)	2.17(2)	**1.5(1)**
	Accuracy	0.47(10)	0.37(11)	0.61(3)	0.52(8)	0.58(6)	0.5(9)	0.58(6)	0.61(3)	0.61(3)	0.64(2)	**0.7(1)**
Twitter_LDL	Chebyshev ↓	0.53(10)	0.63(11)	0.28(4)	0.37(8)	0.28(4)	0.37(8)	0.36(7)	0.28(4)	0.26(3)	**0.25(1)**	**0.25(1)**
	Clark ↓	0.85(6)	0.91(11)	**0.58(1)**	0.89(9)	0.86(8)	0.89(9)	0.85(6)	0.84(3)	0.84(3)	0.83(2)	0.84(3)
	Canberra ↓	0.77(5)	0.88(11)	**0.41(1)**	0.84(9)	0.79(8)	0.84(9)	0.78(7)	0.76(2)	0.77(5)	0.76(2)	0.76(2)
	KL ↓	1.31(9)	1.65(10)	3.89(11)	1.19(7)	0.64(4)	1.19(7)	0.85(6)	0.67(5)	0.54(3)	0.53(2)	**0.51(1)**
	Cosine ↑	0.53(10)	0.25(11)	0.82(4)	0.71(8)	0.82(4)	0.71(8)	0.75(7)	0.82(4)	0.83(3)	0.85(2)	**0.86(1)**
	Intersection ↑	0.4(10)	0.21(11)	0.66(3)	0.59(6)	0.63(5)	0.57(8)	0.56(9)	0.58(7)	0.65(4)	0.68(2)	**0.69(1)**
	Average Rank	8.33(10)	10.83(11)	4(4)	7.83(8)	5.5(6)	8.17(9)	7(7)	4.17(5)	3.5(3)	1.83(2)	**1.5(1)**
	Accuracy	0.45(10)	0.4(11)	0.73(4)	0.72(6)	0.7(7)	0.57(9)	0.7(7)	0.74(3)	0.73(4)	0.76(2)	**0.77(1)**

- Methods through problem transformation (PT): PT-Bayes and PT-SVM are based on traditional classification methods SVM and Naive Bayes, and change the training samples into weighted single-label samples for distribution learning (Geng 2016);
- Methods through algorithm adaptation (AA): traditional classification methods kNN and BP neural network are extended to deal with distribution learning and called AA-kNN and AA-BP (Geng 2016).
- Specialized algorithm methods (SA): SA-IIS uses a strategy similar to improved iterative scaling (IIS) that assuming the probability of each emotion to be the maximum entropy model (Geng, Yin, and Zhou 2013); SA-BFGS is based on IIS using the idea of an effective quasi-Newton method Broyden-Fletcher-Goldfarb-Shanno for improving (Geng 2016); SA-CPNN is conditional probability neural network(Geng, Yin, and Zhou 2013).
- CNN-based methods (CNN-based): CNNR uses Euclidean loss for learning (Peng et al. 2015); DLDL uses KL divergence as loss function (Gao et al. 2017); a multi-task CNN-based framework through Joint image emotion Classification and Distribution Learning (JCDL) (Yang, She, and Sun 2017).

For PT, AA and SA methods, features extracted from the last fully connection layer based on VGGNet deep model (Simonyan and Zisserman 2014) are used for distribution learning. Moreover, PCA approach is also used to get the first 280 principal features for more comparable and effective learning. All our experiments are carried out on a NVIDIA GTX Titan Xp GPU with 12GB memory.

In order to compare different label distribution learning methods, measures should be able to calculate the average distance or similarity between the ground-truth and predicted label distributions. As suggested in (Geng 2016), six distribution learning measures are applied in our experiments including Chebyshev distance(↓), Clark distance(↓), Canberra metric(↓), and KL divergence(↓). The similarity measures include Cosine coefficient(↑) and Intersection similarity(↑). Down arrow ↓ means lower is better, and up arrow ↑ means higher is better. Moreover, since KL divergence is not well defined when zero occurs, we use a small value $\epsilon = 10^{-10}$ to replace zero value. In addition, the maximum values of Clark distance and Canberra metric are determined by the number of emotions. For standardized comparison, we divided Clark distance by the square root of number of emotions and divided Canberra metric by the number of emotions. Moreover, We can utilize the max emotion in ground-truth and predicted distributions as the single emotion label for classification. In this way, accuracy can be calculated by classification result of the max emotion.

Results and Analysis

On Image Emotion Distribution Learning. Table 1 presents the performance of SSDL and baseline emotion distribution learning methods measured by Chebyshev, Clark, Canberra, KL divergence, Cosine, Intersection, average rank of these measures and accuracy. The performance is shown as mean values of each measure and the ranks of each measure which are described in the rackets. Besides, results in

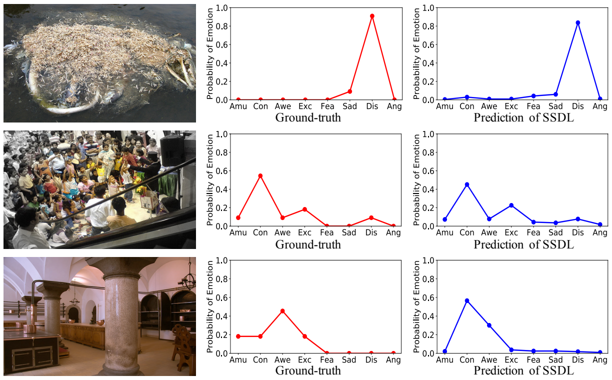

Figure 5: Predicted emotion distributions of SSDL on image examples are shown in last column. The ground-truth distributions is shown in the middle column.

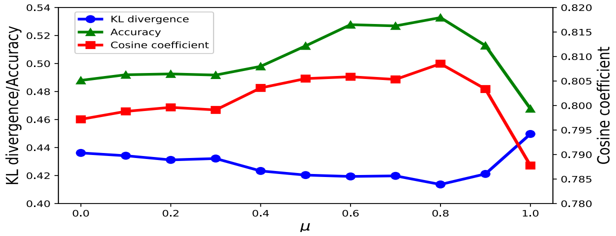

Figure 6: Effect of μ for combined loss on Emotion6 dataset. Note that $\mu = 1$ means only using EMD-based loss, and $\mu = 0$ means only using KL loss.

bold indicate the best values of each measure.

From the results, we can observe that: (1) SSDL shows superiority in most of measures on all datasets and ranks first on all average rank measures, which demonstrates the effectiveness of SSDL in image emotion distribution learning by leveraging structured and sparse annotations; (2) SSDL achieves the best max emotion classification performance, even though the main goal of our proposed method SSDL is predicting the emotion distribution. (3) for Clark and Canberra measures, SSDL shows worse performance than AA-kNN because that AA-kNN predicts emotion distribution depending on k nearest samples and is suitable for predicting emotions with low probabilities which are beneficial to the improvement of Clark and Canberra; (4) four CNN-based methods rank top 5 performance except that the performance of CNNR in few measures, which reveals the strong capacity of CNNs in image emotion distribution learning.

Several images from Flickr_LDL dataset are shown in Figure 5, followed by the ground-truth and predicted emotion distribution by SSDL. From the results, we can see that SSDL predicts distribution similar to the ground-truth distribution and captures the sparse annotations of emotions. Specially, predicted distribution of the bottom image in Figure 5 demonstrates that mis-predictions of SSDL for the image mainly move to nearby emotions in the *positive group* which is caused by the usage of EMD-based loss.

Table 2: Effect of sparse regularization R_{sr} for SSDL on Emotion6 dataset.

Methods	KL divergence ↓	Cosine ↑
SSDL(without R_{sr})	0.43	0.79
SSDL(with R_{sr})	**0.41**	**0.81**

On Sensitivity of Combined Loss Parameter μ. In SSDL, μ controls the relative importance between the EMD-based loss and KL loss. The bigger value of μ, the more important of EMD-based loss. We use Emotion6 and three measures, which are KL divergence, Cosine coefficient and Accuracy, to demonstrate how μ influences the performance of SSDL on Emotion6. The results are presented in Figure 6. From the curves, we find that: (1) the performance of only using EMD-based loss or KL loss is worse than using both as a combined loss, which illustrates that EMD or KL divergence can only handle the one aspect of the dissimilarities and using both can improve performance of distribution prediction; (2) performance of combined loss is effective and stable when increases μ from 0.5 to 0.8, which means addressing the usage of EMB-based loss. Through these results, we can find out that our proposed combined loss is robust for image emotion distribution learning in SSDL.

On Effect of Sparse Regularization R_{sr}. In order to check the effect of sparse regularization in SSDL, we perform experiments on Emotion6 using SSDL with and without sparse regularization R_{sr}. KL divergence and Cosine coefficient are used for performance comparing. Firstly, we set $\xi_1 = 0$ to get SSDL without R_{sr}. Secondly, we use both components of regularization to get SSDL with R_{sr}. Then we compare these methods and the results are shown in Table 2. From the results, we can detect that the performance of SSDL in emotion distribution learning is improved by using sparse regularization, which reveals the effectiveness of R_{sr} by capturing different levels of sparse annotations.

Conclusion

In this paper, we explored how to effectively use structured and sparse annotations in image emotion distribution learning. A novel CNN-based framework named Structured and Sparse annotations for image emotion Distribution Learning was proposed, in which emotional characters of polarity and intensity are effectively considered. Combined EMD-based and KL loss were used to take structured annotations into account for emotional polarity. Meanwhile, sparse regularization was designed to take sparse annotations into account for emotional intensity. Extensive experiments on distribution datasets revealed that the effectiveness of SSDL in image emotion distribution learning through utilizing rich information extracted from structured and sparse annotations.

Acknowledgements

This research was supported by Beijing Natural Science Foundation (4172014, 4184084), Support Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five-year Plan (CIT&TCD201804031), NSFC (61572205) and the NVIDIA GPU donation grant.

References

Baldassarre, L.; Bhan, N.; Cevher, V.; Kyrillidis, A.; and Satpathi, S. 2016. Group-sparse model selection: Hardness and relaxations. *IEEE Transactions on Information Theory* 62(11):6508–6534.

Cambria, E.; Poria, S.; Gelbukh, A.; and Thelwall, M. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems* 32(6):74–80.

Candès, E. J., and Wakin, M. B. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine* 25(2):21–30.

Chen, M.; Zhang, L.; and Allebach, J. P. 2015. Learning deep features for image emotion classification. In *Image Processing (ICIP), 2015 IEEE International Conference on*, 4491–4495. IEEE.

Gao, B. B.; Xing, C.; Xie, C. W.; Wu, J.; and Geng, X. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26(6):2825–2838.

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.

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

Hu, Y.; Li, C.; Meng, K.; Qin, J.; and Yang, X. 2017. Group sparse optimization via lp, q regularization. *Journal of Machine Learning Research* 18:1–52.

Levi, G., and Hassner, T. 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 503–510. ACM.

Levina, E., and Bickel, P. 2001. The earth mover's distance is the mallows distance: Some insights from statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, 251–256. IEEE.

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

Peng, K. C.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 860–868.

Scardapane, S.; Comminiello, D.; Hussain, A.; and Uncini, A. 2017. Group sparse regularization for deep neural networks. *Neurocomputing* 241:81–89.

Sheppes, G.; Scheibe, S.; Suri, G.; Radu, P.; Blechert, J.; and Gross, J. J. 2014. Emotion regulation choice: a conceptual framework and supporting evidence. *Journal of Experimental Psychology: General* 143(1):163.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Tripathi, S.; Acharya, S.; Sharma, R. D.; Mittal, S.; and Bhattacharya, S. 2017. Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In *AAAI Conference on Artificial Intelligence*, 4746–4752.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. CNN-RNN: A unified framework for multi-label image classification. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2285–2294. IEEE.

Yang, J.; She, D.; and Sun, M. 2017. Joint image emotion classification and distribution learning via deep convolutional neural network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3266–3272. AAAI Press.

Yang, J.; Sun, M.; and Sun, X. 2017. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI Conference on Artificial Intelligence*, 224–230.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI Conference on Artificial Intelligence*, 381–388.

Zhang, M. L., and Zhou, Z. H. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T. S.; and Sun, X. 2014a. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, 47–56. ACM.

Zhao, S.; Yao, H.; Yang, Y.; and Zhang, Y. 2014b. Affective image retrieval via multi-graph learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, 1025–1028. ACM.

Zhao, S.; Yao, H.; Gao, Y.; Ding, G.; and Chua, T. S. 2016. Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing*.

Zhao, S.; Ding, G.; Gao, Y.; and Han, J. 2017. Learning visual emotion distributions via multi-modal features fusion. In *Proceedings of the 2017 ACM on Multimedia Conference*, 369–377. ACM.

Zhao, S.; Ding, G.; Gao, Y.; Zhao, X.; Tang, Y.; Han, J.; Yao, H.; and Huang, Q. 2018a. Discrete probability distribution prediction of image emotions with shared sparse learning. *IEEE Transactions on Affective Computing* (1):1–1.

Zhao, S.; Gao, Y.; Ding, G.; and Chua, T. S. 2018b. Real-time multimedia social event detection in microblog. *IEEE Transactions on Cybernetics* 48(11):3218–3231.

Zhao, S.; Zhao, X.; Ding, G.; and Keutzer, K. 2018c. Emotiongan: Unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *Proceedings of the 26th ACM international conference on Multimedia*, 1319–1327. ACM.