Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network

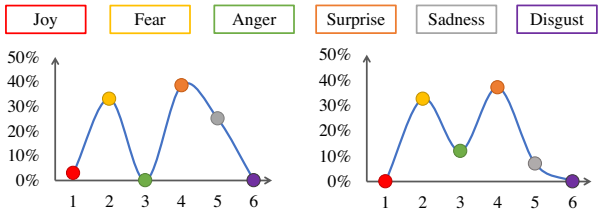
Yuxiang Zhang[1], **Jiamei Fu**[1,2], **Dongyu She**[2], **Ying Zhang**[2], **Senzhang Wang**[3] and **Jufeng Yang**[2*]
[1]College of Computer Science and Technology, Civil Aviation University of China, Tianjin, China
[2]College of Computer and Control Engineering, Nankai University, Tianjin, China
[3]College of Comp. Sci.&Tech., Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract

Emotion analysis of on-line user generated textual content is important for natural language processing and social media analytics tasks. Most of previous emotion analysis approaches focus on identifying users' emotional states from text by classifying emotions into one of the finite categories, *e.g.*, joy, surprise, anger and fear. However, there exists ambiguity characteristic for the emotion analysis, since a single sentence can evoke multiple emotions with different intensities. To address this problem, we introduce emotion distribution learning and propose a multi-task convolutional neural network for text emotion analysis. The end-to-end framework optimizes the distribution prediction and classification tasks simultaneously, which is able to learn robust representations for the distribution dataset with annotations of different voters. While most work adopt the majority voting scheme for the ground truth labeling, we also propose a lexicon-based strategy to generate distributions from a single label, which provides prior information for the emotion classification. Experiments conducted on five public text datasets (*i.e.*, SemEval, Fairy Tales, ISEAR, TEC, CBET) demonstrate that our proposed method performs favorably against the state-of-the-art approaches.

1 Introduction

Text emotion analysis aims to automatically detect and analyze emotions expressed by users towards topics of specific events, services, or other interests [Yadollahi *et al.*, 2017]. Understanding the emotion perspectives plays an important role in human intelligence, decision making, interpersonal communication, which also leads to various potential applications, *e.g.*, customer care service [Jain and Kulkarni, 2014], product recommendation [Russell *et al.*, 2013] and human-machine interaction [Zhou *et al.*, 2018]. In the recent years, the task of emotion classification for text has attracted increasing attention from many researchers, which aims to classify existing emotion in text into one (or more) of a set of pre-defined categories.



(a) Study finds gritty air raises heart disease risk in older women.

(b) Iraq car bombings kill 22 people, wound more than 60.

Figure 1: Examples from the SemEval dataset with their ground truth annotations. We also show the percentage of labels corresponding to each sentence, where each color indicates one of the six emotions.

Most existing approaches towards emotion classification can be regarded as single-label learning (SLL) problem, in which a single dominant emotion is assigned to each sentence. This single-label emotion classification task either depends on emotion lexicons that contain affective words and their corresponding emotion labels [Agrawal and An, 2012] or utilizes existing machine learning classifier to recognize emotions [Perikos and Hatzilygeroudis, 2016]. In practical application, however, a sentence can evoke several emotions at the same time. Recently, multi-label learning (MLL) has been studied extensively, which selects a threshold for the output of a classifier, and assigns multiple emotions with probabilities higher than the threshold to the sentence [Luyckx *et al.*, 2012; Phan *et al.*, 2016].

Both SLL and MLL methods aim to handle the problem that one sentence contains which emotion labels, but they cannot solve the issue of ambiguity related to the emotion [Gao *et al.*, 2017]. Figure 1 shows the examples from the SemEval dataset [Strapparava and Mihalcea, 2007] and the ground truth annotations. For example, surprise accounts for 39% of emotion expressed in the sentence (a), while fear and sadness are presented for 33%, 25%, respectively. The annotation demonstrates that a single sentence contains multiple emotions with different intensities rather than a single representation label, while such ambiguity characteristic is ignored in the SLL and MLL methods.

*Corresponding author: yangjufeng@nankai.edu.cn. This paper was finished in Nankai University.

This paper proposes to address the problem in text emotion analysis field via label distribution learning (LDL), which can represent the intensity associated with each label to describe an instance [Geng, 2016]. In detail, a multi-task convolutional neural network (CNN) is proposed to predict multiple emotions with different degrees in a single sentence. The cross-entropy and Kullback-Leibler (KL) loss are employed as optimization function for the classification and distribution learning, respectively. By combining the two losses, our framework learns both distribution prediction and classification tasks at the same time. During the end-to-end training process, these two tasks can boost each other providing a robust representation for text. Since current datasets are mostly annotated by a single label, we also propose a strategy to transform the ground truth label to the distributions, which provides prior information for emotion analysis. The label with the maximum value of the predicted distribution can also be regarded as the dominant emotion for classification task. Experiment results on distribution dataset (*i.e*., SemEval) and single label datasets (*i.e*., ISEAR, Fairy Tales, TEC, CBET) demonstrate that our proposed method can effectively predict emotion distributions in text and outperforms the state-of-the-art emotion classification approaches. Our contributions are summarized as follows:

- We address the emotion ambiguity in text via label distribution learning. A multi-task convolutional neural network is proposed to learn the tasks of distribution prediction and classification simultaneously.
- We also propose a lexicon-based conversion strategy to generate the emotion distributions from the dominant label for the single-label dataset, which provides context-independent emotion information from the affective words in a sentence.

2 Related Work

In this section, we focus on reviewing the related approaches to sentiment analysis, which includes sentiment classification and emotion classification. The former is defined to classify the text into positive or negative (or sometimes neutral) opinion. This paper focuses on the latter that aims to classify fine-grained categories of existing emotion in text, *e.g*., anger, fear, joy, sadness *etc*. A general survey is provided in [Yadollahi *et al.*, 2017].

Existing methods for emotion classification are generally for SLL task. There are several approaches proposed based on lexicons, which depend on the emotional words and their corresponding labels to identify emotions in text, *e.g*., WordNet-Affect [Strapparava and Valitutti, 2004], NRC [Mohammad and Turney, 2013] and EmoSenticNet [Poria *et al.*, 2014]. Agrawal *et al*. [Agrawal and An, 2012] classify emotional and non-emotional sentences using the constructed emotion lexicon, while Wang *et al*. [Wang and Pal, 2015] propose a model with several constraints based on an emotion lexicon for emotion classification. Other approaches treat the emotion classification as a supervised learning task, in which a learning model is trained on the features of the labeled data to identify the emotion state for the sentences or documents. For example, Choudhury *et al*. [Choudhury *et al.*, 2012] employ a maximum entropy classification framework to detect human affective states in social media. However, there are scenarios where multiple emotions exist in text are desired. Aforementioned methods assume that each sentence only has one emotion, which is sub-optimal method for learning multiple emotions in text.

MLL methods are employed to assign multiple labels to an instance simultaneously [Zhang and Zhou, 2014], which can be grouped into two categories: problem transformation methods and algorithm adaption methods. The first category aims to transform multi-labeled data into single-label problem space and then employ single-label classifier, such as [Luyckx *et al.*, 2012; Phan *et al.*, 2016]. The second category adapt the existing single-label classification algorithm to classify multi-labeled data. Multi-label learning algorithms, such as ML-KNN and Rank-SVM [Zhang and Zhou, 2014], can be employed to detect multiple emotions in text. However, these methods assign multiple labels to a single instance while cannot detect different intensities associated with each label. Geng [Geng, 2016] propose LDL to represent the degree to which each label describes the instance, which have been employed in many fields, *e.g*., image sentiment [Yang *et al.*, 2017b], age estimation [He *et al.*, 2017], *etc*. Emotion distribution learning is proposed to specially identify multiple emotions with their intensities in sentence [Zhou *et al.*, 2016]. However, they need complex designed textual feature, which costs a lot of human efforts.

Recently, numerous works employ deep framework for sentiment classification and achieve remarkable results. Kim [Kim, 2014] propose a series of CNN variants on top of pretrained word vectors for sentence-level sentiment classification tasks. Qian *et al*. [Qian *et al.*, 2017] present linguistically regularized Long Short-Term Memory models (LSTMs) that employ sentiment lexicons to model the linguistic rules in the neural network models. Several methods on multi-task learning are also proposed to improve sentence-level sentiment classification. For example, Chen *et al*. [Chen *et al.*, 2015] propose lifelong learning to retain the knowledge from past learning to help future learning in the Bayesian framework. Li *et al*. [Li and Lam, 2017] utilize separate LSTM network to handle different task with extended memory interactions and jointly learn three tasks. Yu *et al*. [Yu and Jiang, 2016] adopt separate neural network to learn parameters and combine the feature embedding to improve cross-domain sentiment classification and sentiment classifier. Different from the previous work on multi-task learning, our proposed method learn common representation in a deep learning framework and promote optimization of two loss function simultaneously in an end-to-end manner.

3 Methodology

In this section, we will explain the details of the proposed approach of the multi-task CNN model, which can learn the text emotion classification and distribution tasks simultaneously. Figure 2 shows our proposed framework, starting from raw text and ending with its emotion distribution prediction. Given a sentence annotated with label distribution, the CNN model first represents the sentence embedding vector matrix,

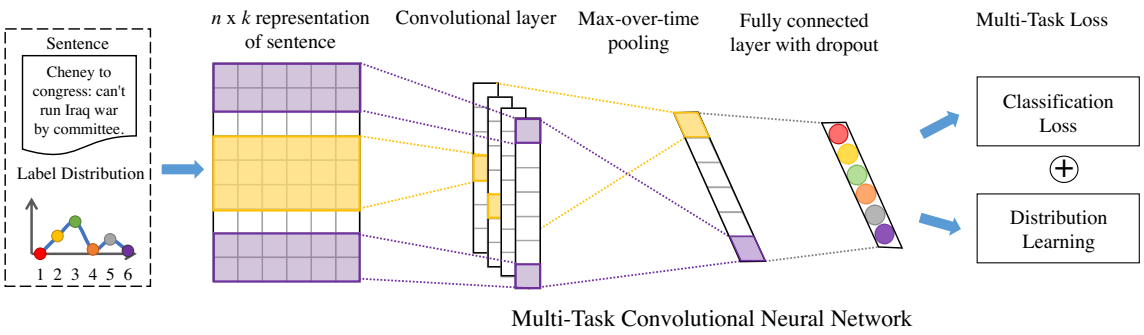


Figure 2: Illustration of the proposed method. Given a sentence annotated with label distribution, our framework represents the sentence with vector matrix using pre-trained word2vec embeddings, then employs different widths of filters and max-pooling operation on it. The multi-task loss is employed as the optimization function, where cross-entropy loss and KL loss are applied for classification and distribution learning, respectively.

then employs multiple widths of filters and max-pooling operation on it. The model combines cross-entropy loss for classification with KL loss for distribution learning as the optimization function. After the forward pass in the model, the features generated from the penultimate layers will be finally fed into the multi-task loss layer. During training, the framework will take back its specific parameters through backpropagation in the backward pass. By employing the stochastic gradient descent (SGD) algorithm, our framework can be optimized in an end-to-end framework.

3.1 Problem Definition

We first formally define the studied problem as follows.

Definition 1 *Emotion Distribution Learning (EDL):* Let $\{(s_i, y_i)\}_{i=1}^{N}$ be a set of N sentences with the corresponding emotion labels of C classes, where $y_i \in \{1, 2, \cdots, C\}$. The goal of emotion distribution learning is to find a function to map each sentence $s_i \in S$ into an emotion distribution $\mathbf{d}_i = \{d_{s_i}^j\}_{j=1}^{C}$. We use d_s^j to represent the intensity of j-th emotion for sentence s. The emotion intensity is normalized to $d_s^j \in [0, 1]$ and $\sum_j d_s^j = 1$ to constitute the emotion distribution. Note that d_s^j is the proportion that emotion j accounts for in the full emotion class y of the sentence s, but not the probability that j correctly labels s. In other words, emotion distribution allows multiple emotions in one sentence, while probabilistic label implies that only one emotion label is correct for each sentence.

3.2 Multi-Task CNN Model

The proposed multi-task CNN model mainly includes several layers, *e.g.*, input layer, convolutional layer, max-pooling layer, and loss layer. Given a training set $S = \{(s_i, \mathbf{d}_i)\}_{i=1}^{N}$, where s_i is the i-th sentence and $\mathbf{d}_i = \{d_{s_i}^1, d_{s_i}^2, ..., d_{s_i}^C\}$ is the associated label distribution. For the distribution dataset, we define the maximum intensity value of the label distributions of the sentence s as its dominant label y. For the single-label dataset, we define the ground-truth label as y.

Input Layer. An input sentence of length M contains a word sequence $s =< w_1, w_2...w_M >$. The m-th word w_m is represented by a real-valued vector $\mathbf{x}_m$ known as word embedding, where $\mathbf{x}_m \in \mathbb{R}^k$ is the k-dimensional word vector corresponding to w_m in the sentence. In this paper, we use fixed-length word2vec word embeddings [Mikolov *et al.*, 2013]. Finally, we concatenate all the word vectors to form the input of the model:

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M]. \quad (1)$$

If the short text is not long enough to up to M, we will pad 0 in the end. Then we get a $k \times M$ embedding matrix for s_i as the input sentence representation.

Convolutional Layer. Let $\mathbf{x}_{p:p+q}$ refer to the concatenation of words $\mathbf{x}_p, \mathbf{x}_{p+1}, ..., \mathbf{x}_{p+q}$. The convolutional layer involves a set of filters $\mathbf{w} \in \mathbb{R}^{h \times k}$, which is applied to a window of h to produce a new feature. For example, a feature v_p is generated from a window of words $\mathbf{x}_{p:p+h-1}$ by

$$v_p = f(\mathbf{w} \cdot \mathbf{x}_{p:p+h-1} + b), \quad (2)$$

where $f(\cdot)$ is a non-linear function such as the sigmoid or ReLu and $b \in \mathbb{R}$ is a bias term. This filter is applied to each possible window of words in the sentence $\mathbf{x}_{1:h}, \mathbf{x}_{2:h+1}, ..., \mathbf{x}_{M-h+1:M}$ to produce a feature map $\mathbf{v}$ as

$$\mathbf{v} = [v_1, v_2, ..., v_{M-h+1}]. \quad (3)$$

Max-pooling Layer. A standard max-over-time pooling operation following [Kim, 2014] is applied over the feature map v to obtain the maximum value by

$$\hat{v} = max(\mathbf{v}), \quad (4)$$

where $\hat{v}$ is treated as the feature corresponding to this particular filter. The pooling scheme aims to capture the most important feature associated with the highest value for each feature map, which can deal with the sentences with variable lengths.

Loss Layer. The formula below calculates the sum of the optimization losses for the network. This layer produces a probability distribution over the emotion labels. Our loss function

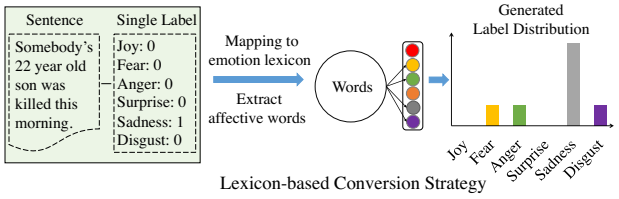


Figure 3: Illustration of proposed lexicon-based conversion strategy. Given a single-label of a sentence, we utilize emotion lexicon to obtain the affective words and the corresponding emotion labels, which are then normalized to the emotion distributions.

composed of cross-entropy loss and KL loss with different weights as follows:

$$E = (1-\lambda)E_{cls}(s,y) + \lambda E_{edl}(s,\mathbf{d}), \tag{5}$$

where E_{cls} represents the cross-entropy loss for classification, and E_{edl} represents the KL loss for distribution prediction. The λ is the weight to control the importance of the two parts of losses. Generally, cross-entropy loss is always applied to maximize the probability of the correct class y, which can be defined by

$$E_{cls}(s,y) = -\frac{1}{N}\left[\sum_i \sum_j \mathbf{1}(y_i = j) \ln \frac{e^{a_j^{(i)}}}{\sum_t e^{a_t^{(i)}}}\right], \tag{6}$$

where the indicator function $\mathbf{1}(\delta) = 1$ if δ is true, otherwise 0. Note that $\left\{a_j^{(i)} | j = 1, 2, ..., C\right\}$ is the activation values of units in the last fully connected layer for s_i. The loss of cross-entropy sums the negative log-likelihood for all the training sentences and penalizes the classification error for each class equally. Therefore, the estimated probability of each emotion class is not considered in the cross-entropy loss, while such variance is useful to deal with the ambiguity among labels in predicting the entire label distribution.

For the distribution learning, we employ the KL loss used in [Gao *et al.*, 2017] to measure the distance between the predicted and true distributions. The KL loss for emotion distribution learning is defined as follows:

$$E_{edl}(s,\mathbf{d}) = -\frac{1}{N}\sum_i \sum_j d_{s_i}^j \ln \frac{e^{a_j^{(i)}}}{\sum_t e^{a_t^{(i)}}}, \tag{7}$$

where $d_{s_i}^j$ indicates the sum of loss of each label for sentence s_i. The optimization of $E_{edl}(s,\mathbf{d})$ assembles all training sentences $\{s_i\}_{i=1}^N$ counting the similarity between two distributions aforementioned.

To solve the optimization problem stated in (5), we employ SGD algorithm to minimize the compositional loss function E. In the final output, the emotion distribution of the given sentence can be predicted, where the label with the maximum probability is treated as the dominant emotion label.

3.3 Lexicon-based Conversion Strategy

Since most datasets only provide single label for a sentence, we propose a lexicon-based conversion strategy that utilizes linguistic resources (*i.e.*, emotion lexicons) to generate distributions from the ground-truth single emotion label. Generally, there can be multiple affective words in a single sentence, and one word can also be corresponding to multiple emotion labels in the lexicon. Such features have been shown highly effective in determining emotion of the sentence [Teng *et al.*, 2016]. Thus we utilize existing lexicons to generate a weak label distribution from the single label as demonstrated in Figure 3. We first map the words from each sentence to the specific emotion according to the emotion lexicon. Then we assign probabilities to the corresponding classes if there exists other affective words other than the ones with the ground-truth emotions. Otherwise, we use the traditional one-hot distribution for the sentences. Such probabilities are then normalized to the emotion distributions considering the ambiguity of each sentence. Note that we adopt the combination of the lexicons NRC [Mohammad and Turney, 2013], Emosenticnet [Poria *et al.*, 2014] and synonyms of emotion label to form our emotion lexicon in experiment.

Formally, given a sentence s_i and the emotion lexicon, we first calculate the intensity score of the j-th emotion according to the mapped emotion D. For $j = y_i$, we define the intensity of the dominant label as follows:

$$\mathbf{d}_{s_i}^{j=y_i} = \begin{cases} \varepsilon, & \text{if } D/y_i \neq \varnothing \\ 1, & \text{otherwise} \end{cases}, \tag{8}$$

where y_i is the ground-truth label position of sentence s_i. For $j \neq y_i$, we define the intensity of j-th emotion label by:

$$\mathbf{d}_{s_i}^{j\neq y_i} = \begin{cases} (1-\varepsilon)\frac{|y_j|}{D/y_i}, & \text{if } D/y_i \neq \varnothing \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where $|y_j|$ is the number of emotional words of the j-th emotion in the sentence. The emotion intensity scores of each label are real numbers, with the sign indicating the strength of each emotion in a sentence. The proposed lexicon-based strategy can provide a word-level emotion to further determine the emotion intensities of the whole sentence. Finally, we obtain the weak emotion distributions for each sentence.

4 Experiments

We conduct extensive experiments on both text distribution and single label datasets to evaluate our proposed framework.

4.1 Experimental Setup

Implementation Details. Employing the pretrained word vectors achieved from unsupervised model as initialized vectors is considered the most popular and effective method to improve performance in a supervised language model. Thus, we use a publicly available word embeddings [Mikolov *et al.*, 2013], which has been trained on 100 billion words from Google News using the continuous bag-of-words model. This dictionary provides a 300-dimensional vector for each word[1]. For the CNN framework, we use filter windows of 3, 4, 5 with 100 feature maps each, dropout rate of 0.5, and mini-batch size of 50 following the same routing in [Kim, 2014]. Our framework is implemented using Torch7.

[1]https://code.google.com/p/word2vec/

Method	Distribution Prediction						Classification			
	Euclidean(↓)	Sϕrensen(↓)	Squared$\mathcal{X}^2$(↓)	K-L(↓)	Cosine(↑)	Intersec(↑)	Pre.(%)	Rec.(%)	F1(%)	Acc.(%)
PT-Bayes	0.7724(12)	0.7036(12)	1.1776(12)	2.5013(12)	0.3798(12)	0.2964(12)	0.1128(8)	0.1660(10)	0.1267(8)	0.2200(9)
PT-SVM	0.5627(10)	0.5600(11)	0.8427(11)	1.0318(9)	0.5685(10)	0.4340(11)	0.1376(7)	0.1780(6)	0.1366(7)	0.2040(10)
AA-KNN	0.5483(9)	0.5457(9)	0.8006(9)	1.3988(11)	0.5897(9)	0.4543(9)	0.2667(3)	0.1901(5)	0.1833(5)	0.2440(8)
AA-BP	0.5216(6)	0.5294(6)	0.7401(5)	0.8310(5)	0.6367(6)	0.4706(6)	0.0686(12)	0.1526(12)	0.0936(11)	0.2610(5)
SA-LDSVR	0.5306(8)	0.5374(8)	0.7520(8)	0.8358(6)	0.6212(8)	0.4626(8)	0.0775(10)	0.1671(9)	0.0563(12)	0.1440(11)
SA-IIS	0.5175(3)	0.5277(4)	0.7324(3)	0.8047(3)	0.6447(3)	0.4723(4)	0.0690(11)	0.1594(11)	0.0944(10)	0.2800(4)
SA-BFGS	0.5754(11)	0.5555(10)	0.8366(10)	1.1165(9)	0.5605(11)	0.4445(10)	0.2155(5)	0.2108(4)	0.2069(3)	0.2560(5)
SA-CPNN	0.5215(5)	0.5252(3)	0.7449(7)	0.8623(8)	0.6417(4)	0.4748(3)	0.2307(4)	0.1711(7)	0.1641(6)	0.2440(8)
BCPNN	0.5207(4)	0.5281(5)	0.7399(4)	0.8377(7)	0.6383(5)	0.4719(5)	0.1813(6)	0.2240(3)	0.1844(4)	0.3000(3)
ACPNN	0.5227(6)	0.5320(6)	0.7428(5)	0.8277(3)	0.6346(6)	0.4680(6)	0.0759(8)	0.1695(7)	0.1048(8)	0.2600(5)
Ours(KL)	0.4623(2)	**0.4060(1)**	0.5627(2)	0.7425(2)	**0.7310(1)**	0.5534(2)	0.4669(2)	0.3851(2)	0.3827(2)	0.4720(2)
Ours	**0.4438(1)**	0.4196(2)	**0.5519(1)**	**0.7306(1)**	0.7291(2)	**0.5804(1)**	**0.4833(1)**	**0.4223(1)**	**0.4141(1)**	**0.5160(1)**

Table 1: Results of various LDL methods on the SemEval dataset. We evaluate the performance using six distance measures (*i.e.*, Euclidean distance, Sϕrensen, Squared$\mathcal{X}^2$, KL divergence, cosine and intersection) and four classification measures (*i.e.*, precision, recall, F-score, accuracy). Note that ↓ after the distance measures indicate smaller is better, and ↑ after the similarity measures indicate larger is better.

Evaluation Metrics. Six metrics are used for distribution prediction evaluation following [Zhou *et al.*, 2016], *i.e.*, Euclidean, Sϕrensen, Squared$\mathcal{X}^2$, KL divergence, Cosine and Intersection. Four metrics are used to indicate the classification performance, including precision (Pre.), recall (Rec.), F-score (F1), and accuracy (Acc.).

4.2 Experiments on Distribution Dataset

Experiments in this section investigate the effectiveness of our proposed approach for predicting emotion distribution and classification task on distribution dataset.

Dataset. SemEval [Strapparava and Mihalcea, 2007] is a distribution dataset that contains 1250 news headlines annotated with six emotion labels: *i.e.*, anger, disgust, fear, joy, sadness and surprise, which is taken from SemEval 2007 Task 14, *i.e.*, Affective Text. The dataset contains the score of each emotion for each headline on a 100-point scale, which are normalized to percentages that summed up to 1 in this paper. For the SemEval, we adopt the standard 1000 headlines for training and 250 headlines for testing to run our experiments. We randomly choose 90% of train samples for training, the rest 10% for testing.

Baseline. For the distribution dataset, we compare our method against the state-of-the-art LDL methods, including PT-Bayes, PT-SVM, AA-KNN, AA-BP, SA-LDSVR, SA-IIS, SA-BFGS, SA-CPNN [Geng *et al.*, 2013; Geng, 2016], BCPNN and ACPNN [Yang *et al.*, 2017b]. PT-Bayes and PT-SVM refer to problem transformation that transforms LDL problem into several single-label problems and then employ Bayes classifier and SVM to predict probabilities of each class for learning label distribution, respectively. AA-KNN and AA-BP refer to algorithm adaptation that extends existing k-nearest neighbors (KNN) algorithm and backpropagation (BP) neural network to solve label distribution problem. SA-LDSVR, SA-IIS, SA-BFGS and SA-CPNN is specialized algorithm for LDL to optimize parametric model directly, which is different from aforementioned problem transformation and algorithm adaption. BCPNN replace the integer labels in conditional probability neural network(CPNN) to binary representation and ACPNN add noises to the ground truth labels to augment learning label distributions. Ours(KL) represents that only KL loss is employed in our framework.

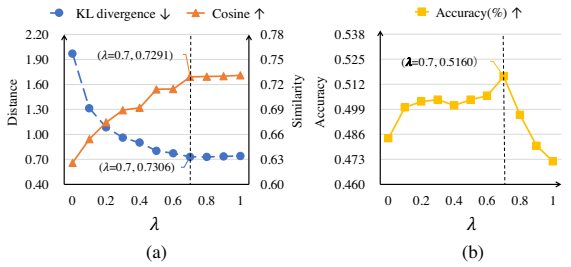


Figure 4: Impact of parameter λ on the distribution prediction (a) and classification (b). Note that $\lambda = 0$ represents that only cross-entropy loss for classification is employed in the framework. Figure (a) indicates the performance of distribution prediction, and (b) demonstrates the performance of classification.

We utilize the output from the penultimate layer of CNN, *i.e.*, a 300-dimensional vector as the input feature of these baseline algorithms.

Effect of Parameter λ. Since we integrate cross-entropy loss with KL loss through a weight(λ), the effect of parameter λ on the model performance is shown in Figure 4. We use KL divergence and cosine coefficient to demonstrate the performance of distribution prediction task and use accuracy for the classification task. As shown in Figure 4 (a), when λ increases from 0 to 1, the performance of distribution prediction first increase dramatically and then gradually steady. Figure 4 (b) shows that the classification accuracy can also be boosted with λ increasing since the ambiguity information can be considered for training more robust model. When λ increase from 0 to 0.6, the performance in distribution prediction and classification rise dramatically. The model performance achieves most favorable when $\lambda = 0.7$, which represents the two tasks reach a balance. When $\lambda = 1$, our method achieves sub-optimal classification performance due to the too much distribution loss can not preserve the dominant label. Thus, we set $\lambda = 0.7$ for our framework considering a balance between the distribution prediction and classification performance in the remaining experiment.

Results and Analysis. The evaluation results of the distribution prediction and the classification on the SemEval dataset

Dataset	Method	Pre.(%)	Rec.(%)	F1(%)	Acc.(%)
ISEAR	NMF	46.10	25.80	16.60	-
	HMM	47.90	26.20	33.90	-
	NB	54.55	55.00	54.78	-
	SVM	60.00	57.50	58.32	-
	CNN	64.21	63.97	63.66	63.90
	Ours(LS)	64.65	64.49	64.57	64.49
	Ours(C1)	65.69	65.76	65.73	65.68
	Ours	**67.11**	**66.91**	**66.80**	**66.75**
Fairy Tales	NMF	74.70	73.10	73.30	-
	CNN	76.68	77.28	76.27	76.82
	Ours(LS)	76.50	77.35	76.92	77.35
	Ours(C1)	76.73	77.36	77.04	77.90
	Ours	**78.21**	**79.23**	**78.72**	**79.21**
TEC	NB	48.74	49.88	49.30	-
	CNN	57.79	51.43	53.55	61.84
	Ours(LS)	59.78	52.04	54.71	62.76
	Ours(C1)	61.37	51.08	53.88	62.99
	Ours	**62.10**	**52.57**	**56.94**	**64.24**
CBET	NB	50.17	48.69	49.92	-
	CNN	59.50	59.53	59.52	59.58
	Ours(LS)	60.01	59.74	60.67	60.05
	Ours(C1)	60.64	60.53	60.56	60.53
	Ours	**61.20**	**61.58**	**61.39**	**61.52**

Table 2: Comparison of classification performance with the state-of-the-art methods on single label datasets.

are shown in Table 1. According to comparison results of six measures, our proposed method performs better in predicting emotion distributions than compared methods. For example, the result of method on Euclidean is lower than BCPNN by 0.0769 indicate the distribution prediction performance better. It demonstrates that our end-to-end network learns better sentence representation than traditional algorithms though we utilize the same text feature. Meanwhile, the result of our method is better than Ours(KL) on most metrics, means compositional losses employed in our method promote learning distribution because two losses can boost each other during training process. Table 1 also shows the superiority of our method in classification performance. Our method performs better than other methods on classification because KL loss enables the model learn emotion ambiguity information.

4.3 Experiments on Single Label Datasets

To evaluate the effectiveness of our lexicon-based conversion strategy and our framework, we also execute experiments on following four single label datasets.

Datasets. We implement our experiments on four single label datasets as follows, including ISEAR, Fairy Tales, TEC, CBET. ISEAR [Scherer and Wallbott, 1994] consists of 7666 sentences annotated by human coders. Different people were asked to report on the circumstances and experiences of the seven major emotions they had experienced, *i.e*., joy, fear, anger, sadness, disgust, shame, and guilt. Fairy Tales [Alm and Sproat, 2005] contains 185 children's stories annotated by five emotion classes, *i.e*., angry, fearful, happy, sad and surprised. Each sentence is annotated by a single emotion label. TEC [Mohammad, 2012] includes 21,051 emotional tweets selected by prespecified hashtags. Each tweet is labeled by one of the following six emotions, *i.e*., anger, disgust, fear, joy, sadness, and surprise. CBET [Shahraki, 2015] consists of 76,860 tweets which are labeled with nine emotions, *i.e*., anger, fear, joy, love, sadness, surprise, thankfulness, disgust, and guilt. Each emotion contains 8,540 tweets. we perform 10-fold cross validation on all the above datasets, and report the average results.

Baseline. For the single label datasets, we compare our approach with the following machine-learning baseline methods on different datasets. NMF is based Non-negative Matrix Factorization to classify emotion-labeled data [Kim *et al.*, 2010]. NB uses Navies Bayes on expanded tweets with the consideration of topics [Shahraki, 2015]. HMM uses a high-order hidden Markov model for emotion detection [Ho and Cao, 2012]. We also implement CNN only with cross-entropy loss as a baseline. To evaluate the effectiveness of the proposed conversion strategy, we evaluate our method against other two strategies, *i.e*., label smoothing Ours(LS) [Szegedy *et al.*, 2016], implication constraint Ours(C1) [Yang *et al.*, 2017a]. For the parameter ε, we all set 0.8 for the dominant label in LS method and our experiments, and rest 0.2 for the other labels for comparison. For the C1 method, we generate label distribution by utilizing different similarity between the pairwise emotion categories.

Results and Analysis. The experimental results of the distribution prediction and the classification on the single label datasets are shown in Table 2. As can be seen, CNN performs better than traditional methods (*i.e*., NMF, HMM, NB, SVM) due to deep network architecture can capture high-level features from affective sentences. Meanwhile, our method outperforms CNN with only the cross-entropy loss. For example, it improves accuracy by 2.85% over CNN and precision by 2.9% on ISEAR dataset. The experimental results demonstrate combining cross-entropy loss with KL loss is more effective to promote classification performance. One can also see that when generating distribution from a single label, our lexicon-based strategy outperforms the LS method [Szegedy *et al.*, 2016], and Constraint 1 used in Yang's method [Yang *et al.*, 2017a], demonstrating the effectiveness of our proposed lexicon-based conversion strategy. This demonstrates that our task-specific method learns more information in sentence with the consideration of affective words.

5 Conclusion

In this work, we propose a multi-task convolutional neural network, which can jointly learn representations for emotion distribution and classification, to predict multiple emotions with their intensities. Moreover, for single-label datasets, we propose a lexicon-based strategy to generate distributions from the single label to further improve the classification performance. Extensive experimental results show that our proposed approach performs favorably against the state-of-the-art methods on both tasks.

Acknowledgments

This work was partially supported by grants from the NSFC (No. U1533104, 61602237), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

[Agrawal and An, 2012] Ameeta Agrawal and Aijun An. Unsupervised emotion detection from text using semantic and syntactic relations. In *WI-IAT*, 2012.

[Alm and Sproat, 2005] Cecilia Ovesdotter Alm and Richard Sproat. Emotional sequencing and development in fairy tales. In *ACII*, 2005.

[Chen *et al.*, 2015] Zhiyuan Chen, Nianzu Ma, and Bing Liu. Lifelong learning for sentiment classification. In *ACL*, 2015.

[Choudhury *et al.*, 2012] Munmun De Choudhury, Michael Gamon, and Scott Counts. Happy, nervous or surprised? classification of human affective states in social media. In *ICWSM*, 2012.

[Gao *et al.*, 2017] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

[Geng *et al.*, 2013] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

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

[He *et al.*, 2017] Zhouzhou He, Xi Li, Zhongfei Zhang, Fei Wu, Xin Geng, Yaqing Zhang, Ming-Hsuan Yang, and Yueting Zhuang. Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image Processing*, 26(8):3846–3858, 2017.

[Ho and Cao, 2012] Dung T. Ho and Tru H. Cao. A high-order hidden markov model for emotion detection from textual data. In *PKAW*, 2012.

[Jain and Kulkarni, 2014] Mukesh C Jain and VY Kulkarni. Texemo: Conveying emotion from text-the study. *International Journal of Computer Applications*, 86(4):43–49, 2014.

[Kim *et al.*, 2010] Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *CAAGET Workshop*, 2010.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *ACL*, 2014.

[Li and Lam, 2017] Xin Li and Wai Lam. Deep multi-task learning for aspect term extraction with memory interaction. In *EMNLP*, 2017.

[Luyckx *et al.*, 2012] Kim Luyckx, Frederik Vaassen, Claudia Peersman, and Walter Daelemans. Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. *Biomedical informatics insights*, 5:BII–S8966, 2012.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Mohammad and Turney, 2013] Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. *NRC Technical Report*, 2013.

[Mohammad, 2012] Saif M Mohammad. Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, 2012.

[Perikos and Hatzilygeroudis, 2016] Isidoros Perikos and Ioannis Hatzilygeroudis. Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence*, 51:191–201, 2016.

[Phan *et al.*, 2016] Duc Anh Phan, Hiroyuki Shindo, and Yuji Matsumoto. Multiple emotions detection in conversation transcripts. In *PACLIC*, 2016.

[Poria *et al.*, 2014] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. Emosenticspace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69:108–123, 2014.

[Qian *et al.*, 2017] Qiao Qian, Minlie Huang, and Xiaoyan Zhu. Linguistically regularized lstms for sentiment classification. In *ACL*, 2017.

[Russell *et al.*, 2013] James A Russell, José-Miguel Fernández-Dols, Anthony SR Manstead, and Jane C Wellenkamp. *Everyday conceptions of emotion: An introduction to the psychology, anthropology and linguistics of emotion*, volume 81. Springer Science & Business Media, 2013.

[Scherer and Wallbott, 1994] Klaus R Scherer and Harald G Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310, 1994.

[Shahraki, 2015] Ameneh Gholipour Shahraki. Emotion mining from text. *Master's thesis,University of Alberta*, 2015.

[Strapparava and Mihalcea, 2007] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of International Workshop on Semantic Evaluations*, 2007.

[Strapparava and Valitutti, 2004] Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, 2004.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[Teng *et al.*, 2016] Zhiyang Teng, Duy-Tin Vo, and Yue Zhang. Context-sensitive lexicon features for neural sentiment analysis. In *EMNLP*, 2016.

[Wang and Pal, 2015] Yichen Wang and Aditya Pal. Detecting emotions in social media: A constrained optimization approach. In *IJCAI*, 2015.

[Yadollahi *et al.*, 2017] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaïane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, 50(2):25:1–25:33, 2017.

[Yang *et al.*, 2017a] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017.

[Yang *et al.*, 2017b] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI*, 2017.

[Yu and Jiang, 2016] Jianfei Yu and Jing Jiang. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *EMNLP*, 2016.

[Zhang and Zhou, 2014] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[Zhou *et al.*, 2016] Deyu Zhou, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. Emotion distribution learning from texts. In *EMNLP*, 2016.

[Zhou *et al.*, 2018] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*, 2018.