AAAS
Research
Volume 2021, Article ID 3067943, 9 pages
https://doi.org/10.34133/2021/3067943

Research
A SCIENCE PARTNER JOURNAL

Research Article

Hierarchical Network with Label Embedding for Contextual Emotion Recognition

Jiawen Deng and Fuji Ren

Faculty of Engineering, Tokushima University, Tokushima 770-8506, Japan

Correspondence should be addressed to Jiawen Deng; c501847002@tokushima-u.ac.jp and Fuji Ren; ren@is.tokushima-u.ac.jp

Received 29 August 2020; Accepted 16 November 2020; Published 6 January 2021

Copyright © 2021 Jiawen Deng and Fuji Ren. Exclusive Licensee Science and Technology Review Publishing House. Distributed under a Creative Commons Attribution License (CC BY 4.0).

Emotion recognition has been used widely in various applications such as mental health monitoring and emotional management. Usually, emotion recognition is regarded as a text classification task. Emotion recognition is a more complex problem, and the relations of emotions expressed in a text are nonnegligible. In this paper, a hierarchical model with label embedding is proposed for contextual emotion recognition. Especially, a hierarchical model is utilized to learn the emotional representation of a given sentence based on its contextual information. To give emotion correlation-based recognition, a label embedding matrix is trained by joint learning, which contributes to the final prediction. Comparison experiments are conducted on Chinese emotional corpus RenCECps, and the experimental results indicate that our approach has a satisfying performance in textual emotion recognition task.

1. Introduction

As an essential element in human nature, emotions have been wildly studied in psychology. Emotion recognition involves the identification of detailed emotional states, which mainly refer to a wide range of mental states, such as happiness, anger, and fear [1]. Textual emotion recognition (TER) is a kind of fine-grained sentiment analysis. It aims to classify a textual expression into one or several emotion classes depending on the underlying emotion theories employed [2].TER should be the most common application in the field of natural language processing, such as mental health monitoring [3], emotional management [4, 5], sinister tone analysis in social networks [6], and human-computer interaction systems [7]. In recent decades, TER tasks have gained considerable interest in the research community.

Recent researches about TER mainly conducted on sentence level, which are aimed at recognizing subtle emotions based on word and concept-based features extracted from the given sentence. However, emotional expression is complicated, and the same sentence could present different emotions in different contexts. In the absence of contextual information, even humans cannot give confident emotional judgments. Therefore, it is necessary to utilize contextual information for sentence-level emotion recognition.

Given a sentence, its context generally refers to the sentences that appear around it. For example, given a sentence from a blog, its context refers to those sentences that appeared around the current sentence. Given an utterance from a dialogue, its context generally refers to the preceding occurred utterances. Such contextual information has been explored in some preceding works, such as HANs [8], TreeLSTM [9], and CLSTM [10]. Under different circumstances, the contextual sentences of current sentence have different contributions to final prediction, and attention mechanism-based networks are widely utilized to address this problem. Inspired by HANs (Hierarchical Attention Networks), we explore effective encoders for sentence-level encoding and contextual-level encoding to generate more accurate emotion representation expressed in the given sentence.

Emotional expression is very complicated. Some emotions often cooccurred with each other, such as the emotion pair of "Joy" and "Love," while some are opposed and rarely appear together, such as "Joy" and "Anxiety." Emotion correlation has always been a significant factor in emotion recognition tasks. To accurately recognize emotions,

it is necessary to fully consider the correlations between each emotion.

This paper explores a hierarchical model to learn contextual representations, which encodes the emotional information of a given sentence based on its context. Besides, to realize emotion correlation learning, we trained a label embedding matrix by joint learning, which is beneficial to emotion correlation-based emotion prediction. Therefore, our contributions are summarized below:

(1) This paper proposes a hierarchical model to learn contextual representations for sentence-level emotion recognition. We take pretrained language model BERT as the sentence-level encoder and take attention-based bidirectional LSTM as the context-level encoder, which are aimed at learning the emotional information of the given sentence based on its context

(2) To give emotion correlation-based prediction, the label embedding matrix is learning by joint learning. Emotion correlation is obtained by calculating the similarity features between sentence representation and each label embedding, which contributes to the final prediction

(3) To guarantee the effectiveness of both emotion prediction and label embedding, the proposed network is trained by an assemble training objective. The experimental results indicate that our approach has a satisfying performance in TER task

The rest of this paper is organized as follows: Section 2 presents some related works of textual emotion recognition and contextual modeling. Section 3 describes the methodology of proposed hierarchical network with label embedding. Experimental results and discussion are shown in Section 4 and Section 5. Finally, Section 6 concludes this paper.

2. Related Work

2.1. Textual Emotion Recognition. Deep learning-based techniques have achieved significant improvement on TER tasks [11]. Word embedding techniques are aimed at learning latent, low-dimensional representations from the language structure and alleviating the problems of sparse features and high-dimensional representation in traditional bag-of-words models. Some well-established embedding models are wildly used in many NLP tasks, such as Word2vec [12, 13] and GloVe [14]. They are trained on a large scale of unlabeled data and aimed to capture fine-grained syntactic and semantic regularities. Recently, the emergence of pretrained language model opened the pretraining era in the NLP field. The pretrained language models provide useful general knowledge and can be fine-tuned to almost all downstream tasks. CoVe and ELMo [15] generate dynamic and context-sensitive word embedding, by which the same words with different contexts are given different word vectors. They greatly alleviate the occurrence of ambiguity. BERT utilizes a large amount of unlabeled data during the training, which helps the model to learn useful linguistic knowledge. Bert performs well in encoding contextual grammatical knowledge and has achieved satisfactory results in many NLP task.

Emotion label correlation is always a critical problem in TER tasks. Deep canonical correlation analysis (DCCA) performs well in feature-aware label embedding and label-correlation aware prediction [16, 17]. Some studies try to explore the label correlations by transforming multilabel classification tasks into ranking or generation tasks. In [18], they transform multilabel emotion classification tasks into a relevant emotion ranking task, which is aimed at generating a ranked list of relevant emotions according to emotional intensity. In [19], multilabel classification is regarded as a sequence generation task, and Sequence Generation Model (SGM) is proposed to learn label correlations. Some approaches attempt to estimate label correlations by the modification of loss function implicitly. The label-correlation sensitive loss function is first proposed in [20] with the BP-MLL algorithm. Joint binary cross-entropy (JBCE) loss [21] is proposed to train the joint binary neural network (JBNN) to capture label relations. To reduce the computational complexity, partial label dependence can also contribute to this task, which is demonstrated in [22]. A semisupervised multilabel method is proposed in [23], while label correlations are incorporated by modifying the loss function. Multilabel classification and label correlations learning can also be realized in a joint learning framework [24, 25].

2.2. Contextual Modeling. Emotion recognition of each sentence in hierarchical texts, such as document or dialogue, highly depends on contextual cues, and context modeling is indispensable. To realize contextual emotion recognition, not only need to model the current sentence but also the contextual sentences, which helps to know the overall emotional tendency of the document or dialogue.

Context modeling architectures can mainly be summarized into two kinds: flatten context modeling and hierarchical context modeling. By flatten context modeling, context sentences and current sentence are concatenated, and all tokens are flattened into a word sequence. This sequence is fed into neural networks for contextual information extraction and final prediction [26, 27]. However, emotions flow naturally in the contextual sentences, and the sequential nature is nonnegligible. Flatten context processing makes the sequence of words too long and ignores the time step. It destroyed the hierarchical structural information of contextual sentences.

The hierarchical structure is a natural characteristic of text: words form sentences and sentences from contexts. This structure knowledge can be incorporated into model architecture for better representation. Inspired by this fact, some works try to stack deep learning architectures to provide specialized understanding at each level [28]. By hierarchical context modeling, each sentence is embedded into a vector representation by sentence-level encoder, and context information is further extracted by hierarchy context encoder [29]. Hierarchical Attention Networks (HANs) is proposed in [8], which mirrors the documents' hierarchical structure.

HANs mainly consists of two layers applied at word and sentence level, respectively. In each layer, there is a GRU-based sequence encoder along with an attention layer, which is aimed at paying more attention to important words and sentences that benefit to the final prediction.

3. Methodology

3.1. Problem Definition. Assume that we have N training sample X along with their contextual information C. Each sample $x \in X$ is often a sentence from the hierarchical text, such as dialogue or blog, and its contextual information $c = \{c_1, c_2, \cdots c_n\} \in C$ often means the preceding n sentences appeared before x. Each sample x is annotated with K emotional labels: $\{e_1, \cdots e_k, \cdots e_K\}$, which is denoted as a one hot vector $y = \{y_1, \cdots y_k, \cdots y_K\} \in \mathbb{R}^{1\times K}$, in which $y_k = 1$ is x contains emotion e_k otherwise $y_k = 0$.

For each sample $x \in X$, a multilabel emotion recognition model F is trained to transform x into predicted distributions $p = \{p_1, \cdots p_k, \cdots p_K\}$ based on its contextual information c and then give a final prediction of all possible emotion labels. The function F is denoted as

$$F(x, c) = \{p_1, \cdots p_k, \cdots p_K\}. \tag{1}$$

3.2. Hierarchical Network with Label Embedding. To model a sentence x along with its contextual information $c = \{c_1, c_2, \cdots c_n\}$, the simplest way is to utilizing flatten context modeling, by which x and contextual sentence c are concatenated as $x' = \{c_1, c_2, \cdots c_n, x\}$, and all tokens in x' are flattened into a word sequence. However, emotions flow naturally in each sentence. Such flatten processing makes the sequence of words too long and ignores the time step, which destroyed the hierarchical structural information. The sequential nature of context is nonnegligible, and such hierarchical information could better contribute to the emotion prediction.

Motivated by Hierarchical Attention Networks (HANs), we focus on hierarchical context modeling. Each sentence in $x' = \{c_1, c_2, \cdots c_n, x,\}$ is first encoded into sentence-level representation $h^s = \{h^s_{c1}, h^s_{c2}, \cdots h^s_{cn}, h^s_x\}$ by a sentence-level encoder En_s, and then, contextual information is further encoded by a hierarchy context encoder. The framework of proposed hierarchical network with label embedding is shown in Figure 1.

3.2.1. Sentence-Level Modeling. At the sentence level, for each sentence $s = \{w_1, w_2, \cdots\}$ in $x' = \{c_1, c_2, \cdots c_n, x\}$, the function En_s encodes s into sentence-level representations h_s, which is denoted as:

$$h^s = En_s(s). \tag{2}$$

Inspired by the pretrained language model and transfer learning techniques, pretrained BERT model [30] is taken as sentence-level encoder En_s in this paper. BERT stands for Bidirectional Encoder Representations from Transformers, and it is designed to pretrain deep bidirectional representations from unlabeled textual data by jointly conditioning on both left and right context in all layers. It remedies the limitation of insufficient training corpora and contributes to syntactic and semantic sentence representation.

In this way, for the sentences in $x' = \{c_1, c_2, \cdots c_n, x\}$, sentence-level representation $h^s = \{h^s_{c1}, h^s_{c2}, \cdots h^s_{cn}, h^s_x\}$ is generated by pretrained BERT model.

3.2.2. Contextual-Level Modeling. At the contextual level, the function En_c encodes the sentence-level representation $h^s = \{h^s_{c1}, h^s_{c2}, \cdots h^s_{cn}, h^s_x\}$ into a context-level representation h^c, which is denoted as:

$$h^c = En_c(h^s). \tag{3}$$

In the proposed model, the function En_c mainly is consisting of two-layer networks: BiGRU (Bidirectional Gated Recurrent Neural Networks) and attention network.

BiGRU is aimed at dealing with the sequential information of contexts. Take sentence-level representation $h^s = \{h^s_{c1}, h^s_{c2}, \cdots h^s_{cn}, h^s_x\}$ as input, the output of the hidden state of BiGRU in each step is $h_i = [\overrightarrow{h}_i : h_i]$, in which $\overrightarrow{h}_i$ and h_i are the output of hidden states from forward and backward directions, respectively.

The attention network is aimed at making the network pay more attention to essential contexts. The attention mechanism considers the contributions of previous occurred contextual sentences $c_i \in c$ to the prediction of current sentence x. More attention weight will be assigned to related contexts. Attention weight a_i and weighted emotional feature vector h^c are defined as follows:

$$\begin{aligned} h^c &= \sum_i a_i h_i, \\ a_i &= \frac{\exp\,(e_i)}{\sum_{k=1}^{n} \exp(e_k)}, \\ e_i &= w_2^T \left[\sigma\left(w_1^T \bullet h_i + b_1\right)\right] + b_2, \end{aligned} \tag{4}$$

in which σ indicates the sigmoid activation function; w_1, b_1, w_2, b_2 indicate the model parameters.

In a typical contextual network, h^c is fed into the classifier for final prediction. The classifier typically consists of a linear transformation. It is followed by a sigmoid operation to normalize the outputs so that each element in will be in the scale of [0,1]. A multilabel neural network is typically trained by minimizing the Binary Cross Entropy (BCE) between the true labels distribution Y and predicted distribution P as the following:

$$\mathrm{BCE}(P, Y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \bullet \log\,(p_{ik}) + (1 - y_{ik}) \bullet \log\,(1 - p_{ik}), \tag{5}$$

in which p_{ik} is the predicted probability of emotion e_k in the ith sample, and y_{ik} is the true label, $y_{ik} \in \{0, 1\}$.

Above-mentioned typically network is intuitive and straightforward and wildly utilized in multilabel classification problems. However, emotion recognition is a more complex

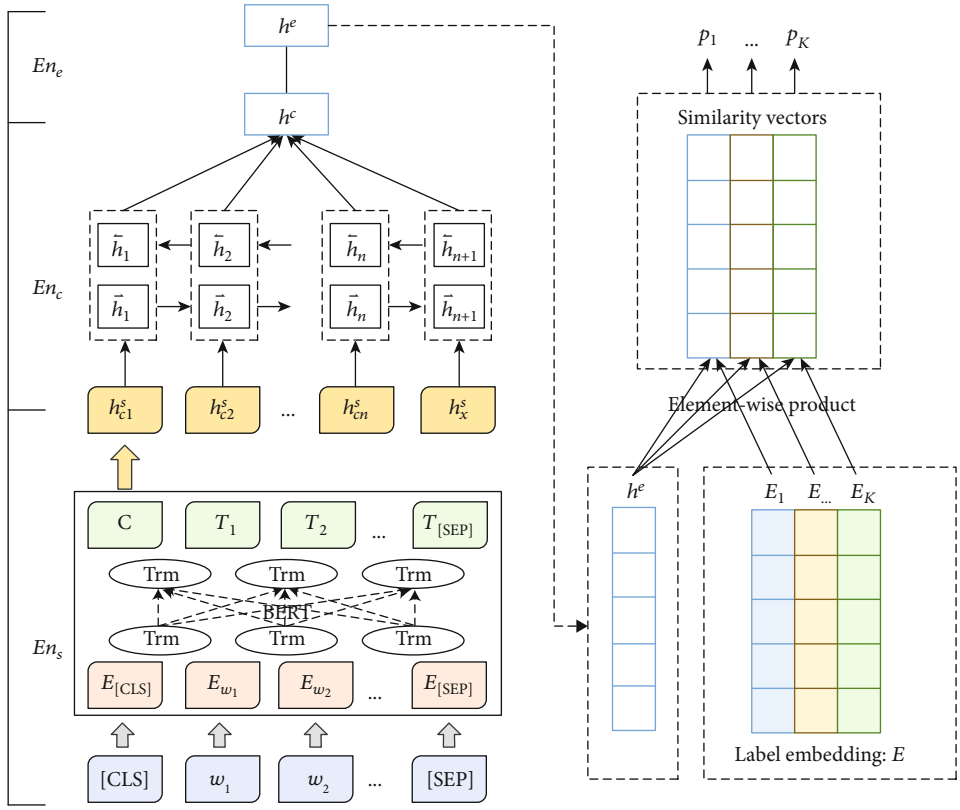


Figure 1: The framework of proposed hierarchical network with label embedding.

problem. This typical network with BCE loss function can be less effective and poor generalization due to its ignorance of label correlations. To capture label correlations, a joint learning label embedding network is proposed, which is detailed in Section 3.2.3.

3.2.3. Label Embedding Network. The label embedding is supposed to represent the semantics and relations between emotion labels. The embedding is denoted by

$$E = \{E_1, \cdots E_k \cdots E_K\} \in \mathbb{R}^{K \times d}, \tag{6}$$

in which K is the number of emotion labels, and d is the dimension of label embedding. Each row in E represents an emotion label.

To make label embedding contribute to the emotion recognition network, the most intuitive way is to compare the emotion representation of contextual input with the label embedding by emotional interaction.

Let the function En_e as emotion projector maps contextual-level representation h^c into emotion representation h^e.

$$h^e = En_e(h^c) = w_e^T \bullet h^c + b_e. \tag{7}$$

Thus, the prediction of all possible emotion labels could be given based on the feature interaction between emotional feature h^e and label embedding matrix E. We firstly implement element-wise product operation on h^e and label embedding E_k of each emotion e_k:

$$h^{e,k} = h^e \odot E_k, k \in [1, K]. \tag{8}$$

$h^{e,k}$ denotes the label-aware feature representation, which incorporates the information of input and a particular emotion label e_k. In this way, the probability of containing emotion e_k is defined as

$$\begin{aligned} p_k &= \sigma\left(w_c^T \bullet h^{e,k} + b_c\right), \\ \sigma(x) &= \frac{1}{1 + e^{-x}}, \end{aligned} \tag{9}$$

in which σ indicates the sigmoid activation function that normalizes the prediction of each emotion p_k in the scale of [0,1]. w_c, b_c indicate the model parameters. The final prediction is given: $P = \{p_1, \cdots p_k, \cdots p_K\}$.

3.3. Training Objectives. For multilabel emotion recognition task, the training objective is often based on binary cross-entropy (BCE). However, BCE loss function takes each emotion as an independent individual and does not consider their relationships. Emotion correlation plays an essential role in this task, which makes emotion recognition be a more complex problem than traditional text classification. To guide the model to learn the emotion correlation during the training process, we propose an assembled training objective to consider all aspects.

3.3.1. Training Objective on the Output Layer. To minimize the loss between the true label distribution and the output distribution, label-correlation aware multilabel loss function is applied at the output layer, which is determined as follows:

$$\text{loss}_{\text{ML}} = \sum_{i=1}^{N} \frac{1}{|Y_i||\bar{Y}_i|} \sum_{(k,l)\in Y_i \times \bar{Y}_i} \exp\left(-\left(p_k^i - p_l^i\right)\right), \tag{10}$$

where Y_i denotes the set of positive emotions for ith sample x_i, and $\bar{Y}_i$ denotes the negative emotions. p_k^i and p_l^i are the output possibility of positive emotion e_k and negative emotion e_l, respectively. Therefore, training with the above loss function is equivalent to maximizing the difference of $(p_k^i - p_l^i)$, which leads the system to output larger values for positive emotions and smaller values for others.

3.3.2. Training Objective on Label Embedding. Given a contextual input x, its positive label is Y_i and its negative label is $\bar{Y}_i$, and $Y = Y_i \cup \bar{Y}_i$. Emotion representation h^e is learned as in Equation (7). In the proposed network, nonlinear label embedding is utilized in the network to guiding the final prediction P by the similarity feature with h^e. In this way, we assume that h^e can in turn be used in the training of label embedding by being closer to the embedding of positive emotions while farther to other negative emotions.

To measure the distance of emotion representation h^e and label embedding, cosine embedding loss is utilized.

$$\text{loss}_{\text{CosEmbed}} = \sum_{i=1}^{N} \frac{1}{K} \bullet \sum_{k=1}^{K} \text{CosLoss}(h_i^e, E_k), \tag{11}$$

$$\text{CosLoss}(h_i^e, E_k) = \begin{cases} 1 - \cos\left(h_i^e, E_k\right), & y_i \in Y_{\text{pos}}, \\ \max\left(0, \cos\left(h_i^e, E_k\right)\right), & y_i \in Y_{\text{neg}}. \end{cases} \tag{12}$$

To guarantee label embedding can encode semantic features among labels, we introduce an additional network to recognize each emotion from corresponding label embedding. For each emotion e_i, its label embedding is E_i. The prediction $\hat{e}_i$ based on E_i is given as

$$p_{e_k} = \text{softmax}(W_{ek} \bullet E_k + b_{ek}), \tag{13}$$

$$\text{loss}_{\text{LabelEmbed}} = \frac{1}{K} \bullet \sum_{k=1}^{K} -e_k \bullet \log\left(p_{e_k}\right). \tag{14}$$

In summary, the assemble training objective of the proposed method is as follows:

$$\text{Loss}(x, y) = \text{loss}_{\text{ML}} + \text{loss}_{\text{CosEmbed}} + \text{loss}_{\text{LabelEmbed}}. \tag{15}$$

4. Experiments

4.1. Dataset. The experiments are conducted on Chinese emotion corpus RenCECps (http://a1-www.is.tokushima-u.ac.jp/member/ren/Ren-CECps1.0/DocumentforRen-CECps1.0.html) to evaluate the proposed architecture. RenCECps is an annotated emotional corpus with Chinese blog texts. The corpus is annotated in document, paragraph, and sentence level [31]. Each level is annotated with eight emotional categories ("Joy," "Hate," "Love," "Sorrow," "Anxiety," "Surprise," "Anger," and "Expect").

Our experiments are conducted at sentence level, and the preceding two sentences of the current sentence are taken as the context information. After preprocessing, there are 24310 contextual sentences in training data and 6746 in testing data.

For each emotion e_i, its cumulative number CN_i is calculated.

$$CN_i = \sum_{\text{n}=1}^{N} \left(y_{n,i} = 1\right), \tag{16}$$

in which $y_{n,i}$ is the annotation of emotion e_i in the nth sample, and the statistical results are shown in Figure 2. Label cardinality (LCard) is a standard measure of multilabeledness and means the average number of emotions concluded per sentence of the corpus and [32]. In RenCECps, LCard is 1.4468.

4.2. Evaluation Metrics. The global performance is evaluated by micro- and macro-F1 score. F1 score is the harmonic mean of precision and recall. Micro-F1 score gives each sample the same importance, while macro-F1 score takes all classes as equally important.

Some popular evaluation measures typically utilized in multilabel classification are utilized to measure the efficiency of proposed methods. Hamming Loss (HL) is the fraction of labels that are incorrectly predicted. Coverage evaluates how far it is needed to go down the ranked emotion list to cover all the relevant emotions in the instance. One Error (OE) evaluates the fraction of sentences whose top-ranked emotion is not in the relevant emotion set. Ranking Loss (RL) evaluates the average fraction of label pairs that are reversely ordered for instance.

4.3. Experimental Details. For a given sentence, its preceding two sentences are taken as contextual sentences. There are total 8 emotion labels annotated for each sentence, and the dimension of label embedding is set to 256. The dimension of hidden state of GRU cell is set to 768/2, and 768 is the dimension of sentence-level embedding.

During the model training, the learning rate is set to 2e-5, and the batch size is set to 128. Adam optimization method is applied to train the model by minimizing the proposed training objective.

4.4. Baselines. In this section, we report the experimental results of our proposed method and baseline models. Additionally, we analyze the influence of training objectives on output layer and label embedding.

We compare our proposed model with six baseline methods as follows.

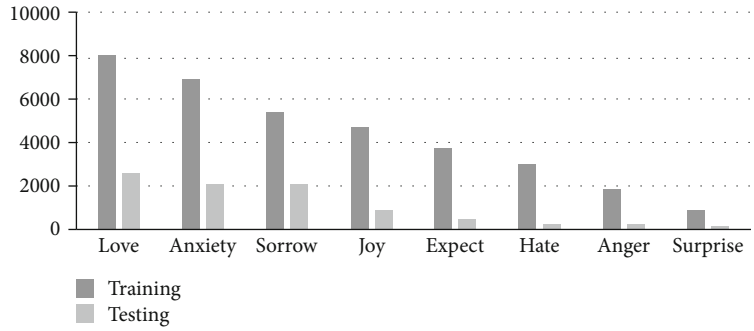


Figure 2: Cumulative number of each emotion in RenCECps.

Table 1: Experimental results in RenCECps.

Metrics	Ours	RERc	HANs	EDL	EmoDetect	ML-KNN	Rank-SVM
Micro-F1 (↑)	*0.5665*	0.5116	0.5573	0.4620	0.4552	0.4720	0.4962
Macro-F1(↑)	*0.4186*	0.4161	0.4003	0.3923	0.3622	0.3632	0.3965
Ranking loss (↓)	*0.1132*	0.2102	0.1136	0.2589	0.2781	0.2928	0.3024
One-error (↓)	*0.3559*	0.4550	0.3623	0.5227	0.5352	0.5543	0.5606
Coverage (↓)	2.1272	*2.1268*	2.1272	2.1699	2.8956	2.4448	2.5962
Hamming loss (↓)	*0.1998*	0.2014	0.2075	0.2102	0.2202	0.2409	0.2585

(1) RERc [18]: a novel framework based on relevant emotion ranking to identify multiple emotions and produce the rankings of relevant emotions from text.

(2) HANs [8]: it has a hierarchical structure that mirrors the hierarchical structure of documents and has two levels of attention mechanisms applied at the word-and sentence-level. In our experiments, sentence-level encoder of HANs is replaced by pretrained BERT model.

(3) EDL [33]: Emotion Distribution Learning, it learns a mapping function from texts to their emotion distributions describing multiple emotions and their respective intensities based on label distribution learning.

(4) EmoDetect [34]: it outputs the emotion distribution based on a dimensionality reduction method using nonnegative matrix factorization, which combines several constraints such as emotions bindings, topic correlations, and emotion lexicons in a constraint optimization framework.

(5) ML-KNN [35]: Multi-Label k-Nearest Neighbor, which adapts the traditional k-nearest neighbor (KNN) algorithm to deal with multilabel data.

(6) Rank-SVM [36]: adapts maximum margin strategy to deal with multilabel data and focuses on distinguishing relevant from irrelevant while neglecting the rankings of relevant ones.

The comparison experiments of baseline HANs are implemented based on the open-source codes shared on GitHub, and the results of other baselines are adopted from the published papers.

5. Experimental Results and Discussions

5.1. Results Analysis. The experimental results of our model compared with the baselines on the RenCECps dataset are shown in Table 1. The best result on each metric is in italics.

As the results shown in Table 1, it indicates that our proposed method significantly outperforms baseline models to a great extent and achieves satisfactory performance on RenCECps dataset. For example, compared to the baseline RERc, our model achieves an improvement of 10.73% micro-F1 score. On multilabel evaluation measures, our model achieves a reduction of 46.15% ranking loss and 21.78% one error. Compared to other baselines, our model achieved satisfactory results as well, which demonstrated the effectiveness of the proposed method.

5.2. The Effectiveness of Label Embedding Layer. Our proposed model is an extension of the baseline of HANs. In our experiments, sentence-level encoder of HANs is replaced by pretrained BERT model. Therefore, by comparing the results of these two models, it can be revealed whether the addition of label embedding layer is effective on the subtask of emotion correlation learning.

As we can see from the results shown in Table 1, the proposed model significantly outperforms baseline HANs, which achieves the improvement of micro-F1 score from 0.5573 to 0.5665 and macro-F1 score from 0.4003 to 0.4186. On multilabel evaluation measures, our model achieves a reduction of ranking loss from 0.1136 to 0.1131, one error from 0.3623 to 0.3559, and hamming loss from 0.2075 to 0.1998.

Both the proposed method and baseline HANs give prediction based on the contextual representation learned from a hierarchical network. HANs directly fed it into output layer for final prediction, which mainly consists of a fully connected layer and an activate function like sigmoid. This implementation is intuitive and straightforward, and it is also a common processing method in most multilabel classification tasks. However, such implementation treats emotion

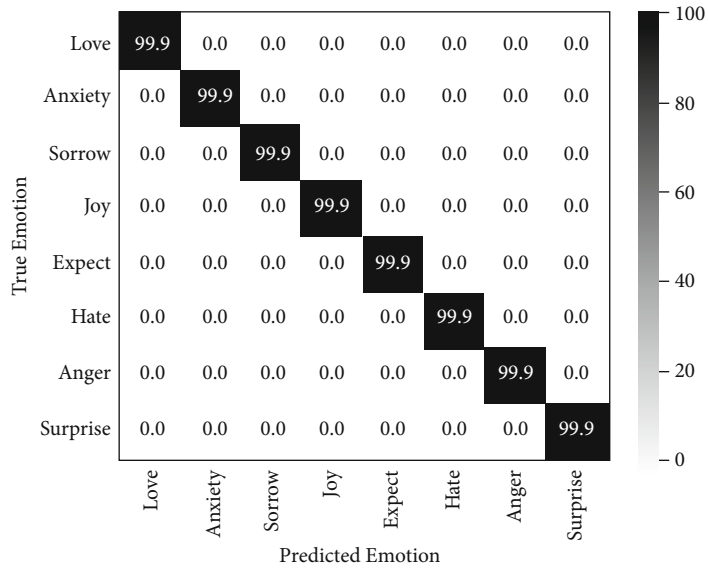


Figure 3: The prediction probability given by label embedding matrix.

recognition task as a general text classification task. It does not consider the correlation between emotion labels, such as the probability of cooccurrence of "Love" and "Happy" is higher than that of "Love" and "Sad." In our proposed model, label embedding space is introduced for emotion correlations learning. The final prediction is based on the interaction of the emotion representation of input text and the label embedding matrix. To guarantee that the label embedding matrix learned the semantic features among labels, training objective on label embedding is utilized to guide the training. The predicted probability given by label embedding matrix, as Equation (13), is visualized in Figure 3. The results in the figure clearly show that the label embedding matrix can accurately predict the corresponding emotion, which means that the emotional information of each label has been actually learned in the label embedding matrix.

5.3. The Effectiveness of Training Objectives. As described in Section 3.3, we proposed an assembled training objective to realize the joint learning of both emotion recognition task and label embedding task. To evaluate the effectiveness of training objectives and label embedding network, we train the proposed model with different training objectives. The results are shown in Table 2. The symbols "M," "C," and "L" denote the loss function of multilabel loss, as in Equation (10), cosine embedding loss, as in Equation (11), and label embedding loss, as in Equation (14), which are utilized for training.

From Table 2, compared with the assembled training objective ("M+C+L"), the proposed model with only multilabel loss ("M") on output layer achieves a reduction of 2.22% micro-F1 and 1.39% macro-F1 and an improvement of 12.37% ranking loss, 6.41% one-error, and 4.52% coverage. It suggests that the proposed ensemble training objective can contribute to the classification improvement.

The experimental results of the proposed model trained on "M+C" and "M" indicate the contribution of cosine embedding loss on the training of the label embedding matrix. Cosine embedding loss guides the training of label embedding by making the emotion representation of input being closer to the embedding of positive emotion labels while farther to other negative emotion labels.

Table 2: Ablation experimental results with different training objectives.

Metrics	M+C+L	M+C	M+L	M
Micro-F1 (↑)	*0.5665*	0.5655	0.5570	0.5539
Macro-F1(↑)	0.4186	*0.4246*	0.4156	0.4128
Ranking loss (↓)	*0.1132*	0.1209	0.1194	0.1272
One-error (↓)	*0.3559*	0.3734	0.3719	0.3787
Coverage (↓)	*2.1272*	2.1778	2.1638	2.2234
Hamming loss (↓)	0.1998	0.1959	0.2040	*0.1957*

"M": multilabel loss; "C": cosine embedding loss; "L": label embedding loss.

The comparison results of the proposed model trained on "M+L" and "M" indicate that the addition of label embedding loss is effective. Label embedding loss guarantees that the trained label embedding matrix is able to encode semantic features among emotion labels.

6. Conclusions

In this paper, we proposed a hierarchical network with label embedding for contextual emotion recognition. Our method involves hierarchically encoding the given sentence based on its contextual information and training a label embedding matrix with an assembled training objective to realize emotion correlation learning. The experimental results show the strong ability of the proposed method to learn emotional features for contextual emotion recognition. In the future, it shall be interesting to incorporate background resources, such as emotion lexicon and knowledge graph, to make the system more satisfactory and robust.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported in part by the Research Clusters program of Tokushima University under grant no. 2003002. This research has been partially supported by NSFC-Shenzhen Joint Foundation (Key Project) (Grant no. U1613217).

References

[1] C. Quan and F. Ren, "An exploration of features for recognizing word emotion," in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 922–930, Beijing, August 2010.

[2] F. Ren and X. Kang, "Employing hierarchical Bayesian networks in simple and complex emotion topic analysis," *Computer Speech & Language*, vol. 27, no. 4, pp. 943–968, 2013.

[3] M. D. Choudhury, M. G. Scott, and C. E. Horvitz, "Predicting depression via social media," in *7th International AAAI Conference on Weblogs and Social Media*, pp. 128–137, Massachusetts, USA, 2013.

[4] S. A. Golder and M. W. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.

[5] S. Kim, J. Lee, G. Lebanon, and H. Park, "Estimating temporal dynamics of human emotions," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 168–174, Austin, TX, USA, 2015.

[6] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: a survey," *Social Network Analysis and Mining*, vol. 8, no. 1, p. 28, 2018.

[7] C. S. Montero and J. Suhonen, "Emotion analysis meets learning analytics: online learner profiling beyond numerical data," in *Proceedings of the 14th Koli Calling International Conference on Computing Education Research - Koli Calling '14*, pp. 165–169, New York, NY, USA, November 2014.

[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, San Diego, CA, USA, June 2016.

[9] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," 2015, http://arxiv.org/abs/1503.00075.

[10] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck, "Contextual lstm (clstm) models for large scale nlp tasks," 2016, http://arxiv.org/abs/1602.06291.

[11] Y. Zhou and F. Ren, "CERG: Chinese emotional response generator with retrieval method," *Research*, vol. 2020, pp. 1–8, 2020.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Deam, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111–3119, NIPS, 2013.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, http://arxiv.org/abs/1301.3781.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014.

[15] M. E. Peters, M. Neumann, M. Iyyer et al., "Deep contextualized word representations," 2018, http://arxiv.org/abs/1802.05365.

[16] C. K. Yeh, W. C. Wu, W. J. Ko, and Y. C. F. Wang, "Learning deep latent space for multi-label classification," 2017, http://arxiv.org/abs/1707.00418.

[17] K. Wang, M. Yang, and W. Yang, "Deep correlation structure preserved label space embedding for multi-label classification," *Asian Conference on Machine Learning*, vol. 95, pp. 1–16, 2018.

[18] D. Zhou, Y. Yang, and Y. He, "Relevant emotion ranking from text constrained with emotion relationships," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 561–571, New Orleans, LA, USA, June 2018.

[19] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: sequence generation model for multi-label classification," 2018, http://arxiv.org/abs/1806.04822.

[20] M. L. Zhang and Z. H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[21] H. He and R. Xia, "Joint binary neural network for multi-label learning with applications to emotion classification," in *Natural Language Processing and Chinese Computing. NLPCC 2018. Lecture Notes in Computer Science, vol 11108*, pp. 250–259, Springer, Cham, August 2018.

[22] S. Lian, J. Liu, R. Lu, and X. Luo, "Captured multi-label relations via joint deep supervised autoencoder," *Applied Soft Computing*, vol. 74, pp. 709–728, 2019.

[23] D. A. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for semisupervised multiple emotion detection of conversation transcripts," *IEEE Transactions on Affective Computing*, 2018.

[24] Z. F. He, M. Yang, Y. Gao, H. D. Liu, and Y. Yin, "Joint multi-label classification and label correlations with missing labels and feature selection," *Knowledge-Based Systems*, vol. 163, pp. 145–158, 2019.

[25] M. Rei and A. Søgaard, "Jointly learning to label sentences and tokens," 2018, http://arxiv.org/abs/1811.05949.

[26] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "SemEval-2019 Task 3: emocontext contextual emotion detection in text," in *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pp. 39–48, Minneapolis, MN, USA, June 2019.

[27] P. Agrawal and A. Suri, "NELEC at SemEval-2019 task 3: think twice before going deep," 2019, http://arxiv.org/abs/1904.03223.

[28] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: hierarchical deep learning for text classification," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, Mexico, December 2017.

[29] G. I. Winata, A. Madotto, Z. Lin et al., "CAiRE_HKUST at SemEval-2019 Task 3: hierarchical attention for dialogue emotion classification," 2019, http://arxiv.org/abs/1906.04041.

[30] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," 2018, http://arxiv.org/abs/1810.04805.

[31] C. Quan and F. Ren, "A blog emotion corpus for emotional expression analysis in Chinese," *Computer Speech & Language*, vol. 24, no. 4, pp. 726–749, 2010.

[32] J. Lee, H. Kim, N. Kim, and J. H. Lee, "An approach for multi-label classification by directed acyclic graph with label correlation maximization," *Information Sciences*, vol. 351, pp. 101–114, 2016.

[33] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng, "Emotion distribution learning from texts," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 638–647, Austin, TX, USA, November 2016.

[34] Y. Wang and P. Aditya, "Detecting emotions in social media: A constrained optimization approach," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 996–1002, Austin, TX, USA, June 2015.

[35] M. Zhang and Z. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[36] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.