Latent Emotion Memory for Multi-Label Emotion Classification

Hao Fei,[1] **Yue Zhang,**[2] **Yafeng Ren,**[3]* **Donghong Ji**[1]*
[1]Key Laboratory of Aerospace Information Security and Trusted Computing,
Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2]School of Engineering, Westlake University, Hangzhou, China
[3]Guangdong Collaborative Innovation Center for Language Research & Services,
Guangdong University of Foreign Studies, Guangzhou, China
{hao.fei, renyafeng, dhji}@whu.edu.cn
yue.zhang@wias.org.cn

Abstract

Identifying multiple emotions in a sentence is an important research topic. Existing methods usually model the problem as multi-label classification task. However, previous methods have two issues, limiting the performance of the task. First, these models do not consider prior emotion distribution in a sentence. Second, they fail to effectively capture the context information closely related to the corresponding emotion. In this paper, we propose a Latent Emotion Memory network (LEM) for multi-label emotion classification. The proposed model can learn the latent emotion distribution without external knowledge, and can effectively leverage it into the classification network. Experimental results on two benchmark datasets show that the proposed model outperforms strong baselines, achieving the state-of-the-art performance.

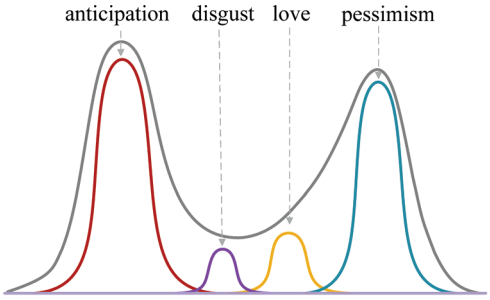

Figure 1: Multiple emotions with different intensities.

Introduction

Emotion classification is an important task in natural language processing (NLP). Automatically inferencing the emotions is the initial step for downstream applications such as emotional chatbots (Zhou et al. 2018), stock market prediction (Nguyen, Shirai, and Velcin 2015), policy studies (Bermingham and Smeaton 2011), etc. However, it is common that there exist more than one emotion in a piece of text. Intuitively, people tend to express multiple emotions in one piece of text. Taking the following sentences as example:

(S1) *How's the new Batman Telltale Series? Looks good but I'm growing weary of this gaming style.*

(S2) *It really is amazing in the worst ways. It was very hard to stifle my laughter after I overheard this comment.*

In sentence S1, multiple emotions are conveyed, including *anticipation, disgust, love* and *pessimism*. Sentence S2 contains two emotions: *joy* and *sadness*. How to identify the multiple co-existing emotions in a sentence remains a challenging task.

There has been work considering multi-label emotion classification (He and Xia 2018; Almeida et al. 2018; Yu et al. 2018). However, there still are two limitations. 1) They assume each emotion occurs with equal prior probability, and fail to consider prior emotion distribution in a sentence. Intuitively, different emotion in a sentence has different intensity. Figure 1 illustrates the emotion distribution of sentence S1, where there are four different emotions with different intensities. In particular, the emotions *anticipation* and *pessimism* receive higher intensities than *disgust* and *love*. Emotion labels with higher intensity should deserve higher probabilities at the final prediction for the model. 2) Previous work does not effectively capture the context information closely related to the corresponding emotion, which is crucial for the prediction. In sentence S2, the clues indicating the *sadness* emotion, '*worst*', are scattered broadly, and surrounded by the words '*laughter*' and '*amazing*' that support the *joy* emotion. Correct predictions can be made only when these features can be sufficiently mined and properly selected. If we can sufficiently capture effective features for each emotion, the final prediction will be relatively easy. This requires a strong ability of the model on features extraction.

To address these issues, we propose a Latent Emotion Memory network (LEM) for multi-label emotion classification. LEM consists of two main components: a latent emotion module and a memory module, which are shown in Figure 2. First, the latent emotion module learns emotions distribution by reconstructing the input via variational au-

*Corresponding author

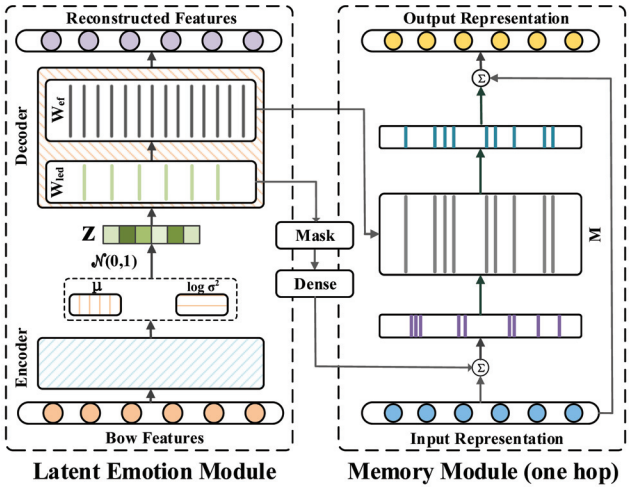

Figure 2: The basic unit of LEM, including the latent emotion module and the memory module with one hop. $\boldsymbol{W}_{led}$ is the latent emotion distribution embedding. $\boldsymbol{W}_{ef}$ is the emotion feature embedding, also used as memory representation for memory module.

toencoder. Second, the memory module captures emotion-related features for the corresponding emotion. Finally, the feature representation from the memory module concatenated with emotion distribution representation from the latent emotion module is fed into a bi-directional Gated Recurrent Unit (BiGRU) to make prediction.

All the components are trained jointly in a supervised end-to-end learning: the latent variable representation from the latent emotion module guides the prediction of the memory module, and the emotion memory module in return encourages the latent emotion module to better learn the emotion distribution through back-propagation. Our model can learn latent emotion distribution information without external knowledge, effectively leveraging it into the classification network.

We conduct experiments on the SemEval 2018 task 1C English dataset and the Ren-CECps Chinese dataset. Experimental results show that our model outperforms strong baselines, achieving the state-of-the-art performance.

Related Work

Multi-label Emotion Classification Emotion detection has been extensively researched in recent years (Ren et al. 2017; Tang et al. 2019). Existing work mainly includes lexicon-based methods (Wang and Pal 2015), graphical model-based methods (Li et al. 2015) and linear classifier-based methods (Quan et al. 2015). More recently, various neural networks models have been proposed for this task, achieving highly competitive results on several benchmark datasets (Ren et al. 2016; Felbo et al. 2017; Baziotis et al. 2018; He and Xia 2018). For example, Wang et al. (2016) employed the TDNN framework by constructing a convolutional neural network (CNN) for multiclass classification. Yu et al. (2018) proposed a transfer learning architecture to improve the performance of multi-label emotion classification. However, these methods do not consider the prior emotion distribution information in a sentence.

Our work is related to the work proposed by Zhou et al. (2016). They proposed an emotion distribution learning (EDL) method, which first learned the relations between emotions based on the theory of Plutchik's wheel of emotions (Plutchik 1980), and then conducted multi-label emotion classification by incorporating these label relations into the cost function (Zhou et al. 2016a). Nevertheless, our method differs from theirs in three aspects: 1) our model effectively learns the emotion distribution, which is free from the restraint of any theory. 2) the emotion intensities distribution is automatically captured during the reconstruction of inputs in VAE model. 3) multi-hop memory module ensures that each emotion makes full use of context information for the corresponding emotion.

Variational Models Our proposed method is also related to work on variational models in NLP applications. Bowman et al. (2015) introduced a RNN-based VAE model for generating diverse and coherent sentences. Miao et al. (2016) proposed a neural variational framework incorporating multilayer perceptrons (MLP), CNN and RNN for generative models of text. Bahuleyan et al. (2017) proposed an attention-based variational seq2seq model for alleviating the attention bypassing effect. Different from the above methods, we first employ the VAE model to make reconstruction for original input, and then make use of the intermediate latent representation as a prior emotion distribution information for facilitating downstream prediction.

Method

The proposed model consists of two main components: a latent emotion module and a memory module. The mechanism of the basic unit of LEM is shown in Figure 2.

Latent Emotion Module

Since we cannot measure the emotion distribution explicitly, we model it as a set of latent variables. We employ variational autoencoder (VAE) to learn the latent multinomial distribution representation $\boldsymbol{Z}$ during the reconstruction of the input.

Encoding The input of latent emotion module is emotion-BoW (eBoW) features of the sentence. Before being fed into VAE, the BoW features are preprocessed so that stopwords or meaningless words are excluded from the vocabulary. The reasons are two fold: 1) our target is to capture the emotion distribution, and the latent representation should be emotion-rich rather than general semantic meaning. 2) reducing the size of BoW vocabulary is beneficial to the training of VAE.

The encoder $f_e(\cdot)$ consists of multiple non-linear hidden layers, transforming the input $\boldsymbol{X}_{eBoW} \in \mathbb{R}^L$ (L is the max length of feature sequence) into prior parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$:

$$\begin{aligned} \boldsymbol{\mu} &= f_{e,\mu}(\boldsymbol{X}_{eBoW}), \\ log\boldsymbol{\sigma} &= f_{e,\sigma}(\boldsymbol{X}_{eBoW}). \end{aligned} \tag{1}$$

We define (Bowman et al. 2015) an emotion latent variable $\boldsymbol{Z}^{'} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \epsilon$, where ϵ is Gaussian noise variable sampled from $\mathcal{N}(0, 1)$, $\boldsymbol{Z}^{'} \in \mathbb{R}^K$ (K denotes the number of emotion labels). Then, the variable is normalized:

$$\boldsymbol{Z} = softmax(\boldsymbol{Z}^{'}). \tag{2}$$

Correspondingly, the latent emotion distribution $p(e_k|, k = 1, \cdots, K)$ is reflected in $\boldsymbol{Z}$.

Decoding Variational inference is used to approximate a posterior distribution over $\boldsymbol{Z}$. We first use a linear hidden layer $f_{led}(\cdot)$ to transform $\boldsymbol{Z}$ into embeddings:

$$\boldsymbol{R}^{led} = f_{led}(\boldsymbol{Z}; \boldsymbol{W}_{led}), \tag{3}$$

where $\boldsymbol{W}_{led} \in \mathbb{R}^{K \times E_{led}}$ (E_{led} is the corresponding embedding dimension) is the learned embedding of the latent emotion distribution. This embedding will later be used to guide the feature learning of memory module, and to control the overall prediction of emotions.

A good decoder can learn a bunch of high level representation of rich emotion features during the process of reconstructing data. With this purpose, we use a non-linear hidden layer $f_{ef}(\cdot)$ to further decode $\boldsymbol{R}^{led}$:

$$\boldsymbol{R}^{ef} = f_{ef}(\boldsymbol{R}^{led}; \boldsymbol{W}_{ef}), \tag{4}$$

where $\boldsymbol{W}_{ef} \in \mathbb{R}^{L \times E_{ef}}$($E_{ef}$ denotes the corresponding embedding dimension) is the global embedding of emotion features, which is later utilized for the memory module as external memory. Finally, the decoding is formulated as:

$$\widehat{\boldsymbol{X}}_{eBoW} = softmax(f_{rec}(\boldsymbol{R}^{ef})), \tag{5}$$

where $f_{rec}(\cdot)$ is the terminal hidden layer, $\widehat{\boldsymbol{X}}_{eBoW}$ denotes the reconstructed features.

Training Following the work of Le et al. (2018), parameters in the latent emotion module are learned by maximizing the variational lower bound on the marginal log likelihood of features:

$$\begin{aligned} logp_\theta(\boldsymbol{X}) \geq \mathbb{E}_{\boldsymbol{Z} \sim q_\phi(\boldsymbol{Z}|\boldsymbol{X})}[logp_\theta(\boldsymbol{X}|\boldsymbol{Z})] - \\ KL(q_\phi(\boldsymbol{Z}|\boldsymbol{X})||p(\boldsymbol{Z})), \end{aligned} \tag{6}$$

where ϕ and θ are the parameters of the encoder and decoder respectively, and the KL-divergence term ensures that the distributions $q_\phi(\boldsymbol{Z}|\boldsymbol{X})$ is near to the prior probability $p(\boldsymbol{Z})$, $p_\theta(\boldsymbol{X}|\boldsymbol{Z})$ describes the decoding process.

Since the training objective of the decoder is to reconstruct the input, it has direct access to the source features. Thus, when the decoder is trained, we assume that $q(\boldsymbol{Z}|\boldsymbol{X}) = q(\boldsymbol{Z}) = p(\boldsymbol{Z})$, which means that the KL loss is zero. It makes the latent variables $\boldsymbol{Z}$ fail to capture information. To combat this, we employ KL cost annealing and word dropout for the encoder (Goyal et al. 2017).

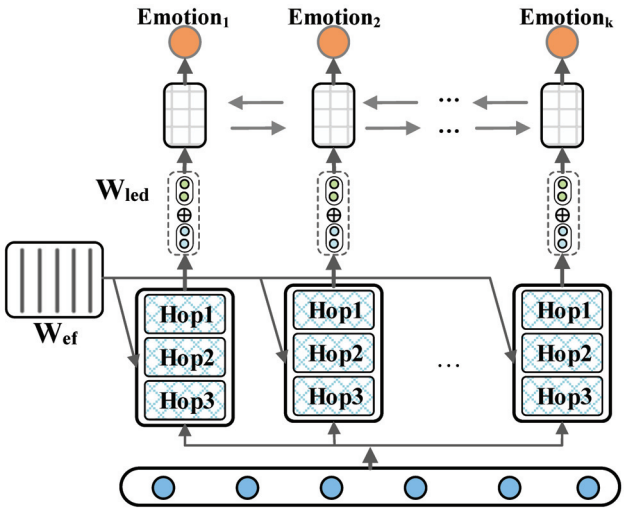

Figure 3: The overall framework of the LEM model. The memory module are private for each k-th emotion. The latent emotion module is shared globally.

Emotion Memory Module

Based on our observation, the evidences and clues for the corresponding emotion are scattered, or even mingled with the features that indicate other emotions in a sentence. This brings challenges for capturing and extracting effective features. We thus leverage multiple-hop memory module (Sukhbaatar et al. 2015) to mine rich context information for the corresponding emotion.

In a memory module, a mask operation is performed on $\boldsymbol{W}_{led}$ for selecting out the k-th emotion embedding $\boldsymbol{W}_{led,k}$. Specifically, the mask:

$$\boldsymbol{Mask} = \underbrace{[\vec{\mathbf{1}}; \cdots; \vec{\mathbf{0}}]}_{K} \tag{7}$$

is a matrix ($\boldsymbol{Mask} \in \mathbb{R}^{K \times E_{led}}$) consisting of K vectors, where $\vec{\mathbf{1}} \in \mathbb{R}^{E_{led}}$ is all-one vector, $\vec{\mathbf{0}} \in \mathbb{R}^{E_{led}}$ is all-zero vector. In the k-th memory module, only the k-th vector is given as $\vec{\mathbf{1}}$, while the others are $\vec{\mathbf{0}}$. The purpose is to guide the k-th memory module to focus on the corresponding emotion on feature learning. Afterwards, a dense layer transforms $\boldsymbol{W}_{led,k}$ into $\boldsymbol{R}^{le}_k$.

Given an input text representation $\boldsymbol{X} = \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_L\}$ ($\boldsymbol{e}_i \in \mathbb{R}^{E_e}$ is the word vector, E_e is embedding dimension), we then compute the relatedness between the text embedding and the corresponding latent emotion representation $\boldsymbol{R}^{le}_k$ via attention:

$$\alpha_{input} = softmax(\boldsymbol{W}_{input}[\boldsymbol{R}^{le}_k; \boldsymbol{X}]), \tag{8}$$

where $\boldsymbol{W}_{input}$ is parameter matrix, $[\boldsymbol{a}; \boldsymbol{b}]$ denotes the concatenating operation. Now we obtain the input context representation:

$$\boldsymbol{R}^{input} = \sum \alpha_{input} \cdot \boldsymbol{X}. \tag{9}$$

On the other hand, we obtain the memory representation $\boldsymbol{M}$ from the global emotion feature embedding $\boldsymbol{W}_{ef}$ via a

linear transformation:

$$\boldsymbol{M} = f_\beta(\boldsymbol{W}_{ef}). \quad (10)$$

Next, we compute the match between the input context representation and the memory, to obtain an output context representation:

$$\alpha_{output} = softmax(\sigma(\boldsymbol{W}_{output}[\boldsymbol{R}^{input}; \boldsymbol{M}])), \quad (11)$$

$$\boldsymbol{R}^{output} = \sum \alpha_{output} \cdot \boldsymbol{R}^{input}, \quad (12)$$

where $\sigma(\cdot)$ denotes the activation function.

Finally, we combine the output context representation with the input text embedding into a new representation $\boldsymbol{R}_k^m$ for the k-th emotion.

$$\boldsymbol{R}_k^m = \boldsymbol{R}^{output} + \boldsymbol{X}. \quad (13)$$

The output representation of current memory hop carries features representation that is relevant to the corresponding emotion.

Latent Emotion Memory Network

The overall framework of LEM is illustrated in Figure 3. The full model maintains K private memory modules for K emotion categories. Different emotions learn their own features, but share one common latent emotion module. The latent emotion embedding $\boldsymbol{W}_{led}$ and emotion feature embedding $\boldsymbol{W}_{ef}$ are shared globally.

We utilize multiple hop memories to make comprehensive exploration of features, repeating computation steps based on previous memory states (Tang, Qin, and Liu 2016).The final representation $\boldsymbol{R}_H^m$ can be obtained via multiple memory hops:

$$\boldsymbol{R}_{h,k}^m = Mem_{h,k}(\boldsymbol{R}_{h-1,k}^m), \quad (14)$$

where k denotes the k-th emotion, $h = (1, \cdots, H)$ is the current hop. The number of memory hops H is decided empirically based on the used dataset.

Finally, we use a bi-directional GRU to learn emotion coherence. As briefly described above, the continuous emotion intensity representation $\boldsymbol{W}_{led}$ can be deemed as an important indication for guiding the label prediction. Thus, we concatenate the latent emotion distribution embedding $\boldsymbol{W}_{led}$ and the features representation $\boldsymbol{R}_{H,k}^m$ learned from each memory module, which is represented as:

$$\boldsymbol{R}_{total,k} = [\boldsymbol{R}_{H,k}^m; \boldsymbol{W}_{led,k}], \quad (15)$$

and feed the results into BiGRU. At the k-th time step, softmax function at the bottom of BiGRU outputs the final prediction for the k-th emotion:

$$p_k = softmax(BiGRU(\boldsymbol{R}_{total,k})). \quad (16)$$

Experiments

Training

We employ a joint cross entropy for the binary softmax classifier for each emotion. The emotion-inference module updates the parameters by minimizing the following loss function:

$$L_{clz} = -\frac{1}{K}\sum_j^K (\hat{p}_j log p_j + (1-\hat{p}_j) log(1-p_j)), \quad (17)$$

where $\hat{p}_j$ is the true label of the j-th emotion.

Note that the latent emotion module is closely joint to the memory module. This ensures that the supervised learning in memory module can also guide the distribution learning of latent emotion module via back-propagation. For example, the emotions *anticipation* and *pessimism* in sentence S1 are labeled as positive, which in return leads to higher weights assigned for these two emotion at corresponding positions in $\boldsymbol{Z}$ during training.

The components elaborated above can be jointly trained. However, directly training the whole framework with cold-start can be difficult and will cause high variance. Thus we first pre-train the latent emotion module until it is close to the convergence via Eq.6. Afterwards, we jointly train all the components via Eq.6 and Eq.17. Once the classification loss is close to convergence, we again train the latent emotion module alone, until it converges. We then co-train the overall LEM. We keep such training iterations until the overall performance reaches its plateau.

Datasets

We conduct experiments on two benchmark datasets, including the English dataset SemEval2018 (Mohammad et al. 2018) and the Chinese dataset Ren-CECps (Quan and Ren 2010). SemEval2018 consists of 11 emotion labels: *anticipation*, *anger*, *fear*, *joy*, *disgust*, *love*, *optimism*, *sad*, *surprise*, *trust* and *pessimism*. Ren-CECps contains 8 emotion labels: *anger*, *expectation*, *anxiety*, *joy*, *love*, *hate*, *sorrow* and *surprise*. Statistics of datasets is shown in Table 1.

In our experiments, the latent emotion module takes emotional BoW as input. We first filter out stopword tokens[1], keeping the words existed in a sentiment dictionary as the lexicons. For English lexicon, we employ GI (Stone, Dunphy, and Smith 1966), LIWC (Pennebaker, Francis, and Booth 2001), MPQA (Wilson, Wiebe, and Hoffmann 2005), Opinion Lexicon (Hu and Liu 2004) and SentiWordNet (Baccianella, Esuli, and Sebastiani 2010). For Chinese lexicon, we use HowNet. Note that we also keep the emotion-rich symbols (e.g. emoji, emoticons). For the memory module, we keep the original token sequence as the input.

Experimental Settings

For English, we use the publicly available GloVe[2] 300-dimensional embeddings trained on 6 billion words from Wikipedia and web text. For Chinese, we train 300-dimensional word embeddings on Chinese Wikipedia on 3.1 billion words using word2vec[3]. We set the same max length for $\boldsymbol{X}_{eBoW}$ and $\boldsymbol{X}$.

In the learning process, we set 300 epochs for pre-training latent emotion module, and 1000 total training epochs with early-stop strategy (Caruana, Lawrence, and Giles 2001). To mitigate overfitting, we apply dropout with a rate of 0.01. We use Adam (Kingma and Ba 2014) for the optimization with

[1]For Chinese, we perform word segmentation first, by using the gensim package: https://radimrehurek.com/gensim

[2]http://nlp.stanford.edu/projects/glove/

[3]https://code.google.com/archive/p/word2vec/

Dataset	Sent.	Words	Avg.len.	Emo.	3 co.e.l.(%)	2 co.e.l.(%)	1 co.e.l.(%)	Train	Dev	Test
Ren-CECps	35,096	228,455	24.56	8	1,824(5.2)	11,416(32.5)	18,812(53.6)	24,567	3,510	7,019
SemEval2018	10,983	32,557	16.04	11	3,419(31.1)	4,442(40.4)	1,563(14.2)	6,838	886	3,259

Table 1: Statistics of datasets. **Avg.len.** is the average length of sentences. **Emo.** denotes the numbers of emotion categories. **#N co.e.l.** denotes the N numbers of co-existing emotion labels in one sentence.

System		SemEval2018					REN-CECps				
		HL	**RL**	**miF1**	**maF1**	**AP**	**HL**	**RL**	**miF1**	**maF1**	**AP**
BR	TextCNN	0.198	0.292	0.548	0.465	0.439	0.204	0.292	0.322	0.301	0.626
	BiLSTM	0.245	0.344	0.498	0.437	0.400	0.212	0.370	0.290	0.277	0.582
	RCNN	0.181	0.311	0.512	0.408	0.385	0.201	0.325	0.3437	0.301	0.605
	attLSTM	0.244	0.248	0.557	0.432	0.449	0.191	0.318	0.350	0.296	0.641
	FastText	0.197	0.235	0.522	0.438	0.428	0.206	0.264	0.312	0.281	0.630
JB	TextCNN	0.161	0.263	0.612	0.496	0.502	0.160	0.230	0.413	0.384	0.689
	BiLSTM	0.183	0.208	0.608	0.485	0.492	0.174	0.274	0.374	0.311	0.672
	RCNN	0.151	0.237	0.649	0.505	0.510	0.168	0.237	0.388	0.344	0.686
	JBNN	0.190	0.192	0.632	0.528	0.526	0.165	0.192	0.418	0.380	0.693
ECC		0.210	0.240	0.458	0.376	0.395	0.210	0.348	0.291	0.256	0.597
MLLOC		0.245	0.342	0.484	0.414	0.413	0.185	0.474	0.278	0.234	0.413
ML-KNN		0.196	0.270	0.410	0.387	0.391	0.245	0.290	0.310	0.285	0.591
TMC		0.191	0.219	0.561	0.465	0.482	0.228	0.252	0.391	0.342	0.630
EDL		0.182	0.177	0.581	0.504	0.501	0.187	0.227	0.458	0.365	0.662
SGM		0.165	0.184	0.616	0.492	0.524	0.187	0.234	0.473	0.392	0.673
RERc		0.176	0.170	0.651	0.539	0.530	0.201	0.210	**0.511**	0.416	0.683
DATN		-	-	-	0.551	-	-	-	-	0.441	0.732
LEM (ours)		**0.142**	**0.157**	**0.675**	**0.567**	**0.568**	**0.151**	**0.183**	0.501	**0.448**	**0.751**
LEM-$\boldsymbol{W}_{led}$		0.185	0.197	0.620	0.517	0.527	0.191	0.204	0.408	0.413	0.711
LEM-$\boldsymbol{R}^{le}$		0.167	0.173	0.640	0.534	0.542	0.168	0.198	0.436	0.427	0.735

Table 2: Experimental results on two datasets. LEM-$\boldsymbol{W}_{led}$ represents the model without latent emotion distribution representation $\boldsymbol{W}_{led}$ in Eq.15. LEM-$\boldsymbol{R}^{le}$ denotes the memory module without $\boldsymbol{R}^{le}$ in Eq.8.

the initial rate of 0.001. We employ five widely used metrics for measuring multi-label classification performance, including **Hamming Loss** (**HL**), **Ranking Loss** (**RL**), **Micro F1** (**miF1**), **Macro F1** (**maF1**) and **Average Precision** (**AP**) (Zhang and Zhou 2014; Zhou, Yang, and He 2018).

Baselines

We make comparisons between LEM and the following baseline systems.

Binary Relevance (BR): Zhang and Zhou (2014) transform the multi-label problem into several binary problems (Zhang and Zhou 2014). Following their settings, we employ five neural classifier that are widely used for text classification, including TextCNN (Kim 2014), BiLSTM (Schuster and Paliwal 1997), RCNN (Lai et al. 2015), attLSTM (Zhou et al. 2016b) and FastText (Joulin et al. 2016).

Joint Binary (JB): He and Xia (2018) show that their model (named JBNN) is better than BR by sharing the relations between labels (He and Xia 2018). Following their settings, we employ five neural classifiers that are widely used for text classification, including TextCNN (Kim 2014), BiLSTM (Schuster and Paliwal 1997), RCNN (Lai et al. 2015), attLSTM (Zhou et al. 2016b) and FastText (Joulin et al. 2016).

Multi-label Emotion Classification: There has also been work for multi-label emotion classification: ECC (Read et al. 2009), MLLOC (Huang and Zhou 2012), ML-KNN (Zhang and Zhou 2014), TMC (Wang et al. 2016), EDL (Zhou et al. 2016a), SGM (Yang et al. 2018), RERc (Zhou, Yang, and He 2018) and DATN (Yu et al. 2018).

Results

The results are shown in Table 2. Our proposed model achieves the best performance compared with all baseline systems on almost all measurements, with 0.568 Average Precision, 0.142 Hamming Loss, 0.157 Ranking Loss on SemEval2018, and 0.751 Average Precision, 0.151 Hamming Loss and 0.183 Ranking Loss on Ren-CECps. Based on experimental results, we have the following observations. First, the joint binary classification methods are better than separate binary classifiers. This finding is consistent to the work of He and Xia (2018). Second, the models that integrate the relations between emotions achieve better performance. For example, SGM and RERc achieve better performance than TMC and BR models. Third, sufficiently capturing and utilizing the emotion distribution is useful for multi-label emo-

Model	anxiety	joy	love	expectation	hate	sorrow	anger	surprise	Avg.
JBNN	0.507	0.456	0.446	0.380	0.268	0.442	0.304	0.251	0.380
EDL	0.453	0.402	0.399	0.259	0.255	0.366	0.307	0.239	0.335
TMC	0.438	0.400	0.391	0.331	0.211	0.393	0.237	0.188	0.322
SGM	0.471	0.429	0.422	0.359	0.241	0.406	0.277	0.217	0.352
LEM(our)	**0.571**	**0.530**	**0.509**	**0.454**	**0.336**	**0.507**	**0.369**	**0.321**	**0.448**

Table 3: Results of each emotion on Ren-CECps. The performance is measured by Macro F1.

tion classification task. For example, our model achieves better performance than previous methods.

We also empirically explore the performances without the prior emotion distribution information. Specifically, $\boldsymbol{R}^{total}$ in the BiGRU does not contain the latent distribution representation $\boldsymbol{W}_{led}$ in Eq.15, and the memory module do not contain $\boldsymbol{R}^{le}$ in Eq.8, respectively. We can see that in both situations, the ability of LEM drops dramatically, though the performances are still better than most of the baselines. The above analysis shows the usefulness of prior emotion distribution for multi-label emotion classification.

Further, we compare the performances between LEM and strong baselines on each emotion. Results on the Ren-CECps dataset are shown in Table 3. Our model gives better results on all categories, while the emotions *hate*, *anger* and *surprise* are more challenging, with relatively lower performances in all models.

Analysis and Discussion

Emotion Distribution Learning We conduct experiments on different number of co-existing emotion labels to validate the ability on multiple label learning. Table 4 reports the results. First, the models (EDL and LEM) that use the information of emotion distribution achieve better performance. Second, LEM model gives better performance than all baseline systems. Third, the more emotions co-exist, the more improvements our LEM achieves. This indicates the effectiveness of incorporating the emotion distribution information for the task.

We further analyze how capable the latent emotion module is on emotion distribution learning. We intercept the emotion distribution $\boldsymbol{Z}$ and measure the distance between the learned distribution and the true distribution, with three metrics: **Euclidean (Eu)**, **Squared χ^2 (Sq)** and **Kullback-Leibler (KL)** (Zhou et al. 2016a). Based on Ren-CECps[4], we make comparisons with three baselines: ML-KNN, EDL and RERc, which directly learn the emotion intensity distribution for the task. As seen in Table 5, the latent emotion module of our model achieves the best KL score (0.190) against the baselines. This demonstrates the strong ability of the latent emotion module on distribution learning.

Impact of Memory Hops We also analyze the impact of different hop numbers in the memory module. Table 6 reports the results. We can see that LEM with 2 hops achieves the best F1 score on Ren-CECps and 3 on SemEval2018.

[4]In Ren-CECps, each emotion in a sentence is labeled with not only a emotion tag, but an intensity score, and thus the normalized intensity scores can be viewed as distribution.

Model	$\geq$ 1 co.e.l.	$\geq$ 2 co.e.l.	$\geq$ 3 co.e.l.
TMC	0.507	0.547	0.608
SGM	0.531	0.581	0.659
EDL	0.538	0.610	0.697
LEM	**0.593**	**0.643**	**0.734**

Table 4: Different number of co-existing emotion labels.

Model	Eu	Sq	KL
ML-KNN	0.247	0.241	0.258
EDL	0.236	0.188	0.206
RERc	**0.186**	**0.151**	0.198
LEM$^\sharp$	0.207	0.168	**0.190**

Table 5: Results on emotion distribution learning. LEM$^\sharp$ denotes the latent emotion module of LEM.

Dataset	1 hop	2 hops	3 hops	4 hops	5 hops
Ren-CECps	0.434	**0.448**	0.440	0.436	0.431
SemEval2018	0.556	0.562	**0.567**	0.560	0.550

Table 6: Impact of different number of memory hops.

This can be explained from two aspects. First, SemEval2018 contains 11 emotion labels, and the co-existing labels in the sentences are more complex than that of Ren-CECps. Second, the cues for supporting the emotions in SemEval2018 are more scattered, though the average length of sentences is shorter than that of Ren-CECps. Besides, we can observe that the increase of hop numbers does not give better performance, and may cause overfitting.

Emotion Discovery We print out the top 10 key words learned from each emotion memory on SemEval2018, by first gathering the top 3 highly lighted tokens at every sentences on each emotion memory, and then collecting all these tokens and sorting them by their frequency. The results are shown in Table 7.

We can find that the words discovered by emotion memories are strong clues for directly indicating the corresponding emotion. This proves that the memory module has strong ability on feature learning and extraction. Besides, some words are simultaneously occurred across multiple emotions. These words can be viewed as evidences, demonstrating the ability of our model on learning the correlations between emotions. Actually, some sentence often involves multiple emotions based on a single indicative clue. For example in sentence S1, the word 'good' is a strong cue for indicating the emotion relevance concurrently on two emotions: *anticipation* and *love*.

Emotion	Key words
anger	angry anger terrible fucking fuming awful can't rage bully shocking
anticipation	new blues good can't watch serious nervous wait horror worry
disgust	angry horrible terrible revenge fuck awful offended shit hate
fear	anxiety fear nervous horror nightmare awful panic afraid terror bad
joy	happy love good smile amazing new hilarious great lol live.ly
love	laughter love smiling best Happy beautiful hilarious glee good great
optimism	happy love good smile optimism life new best great fear
pessimism	sad depression sadness lost depressing awful can't anxiety nervous bad
sadness	nightmare can't depressing unhappy bad life anxiety terrible awful bitter
surprise	shocking amazing hilarious Watch live.ly new broadcast serious awe believe
trust	good :) love best fear optimism worry amazing new faith

Table 7: Top 10 key words on each emotion discovered by emotion memory.

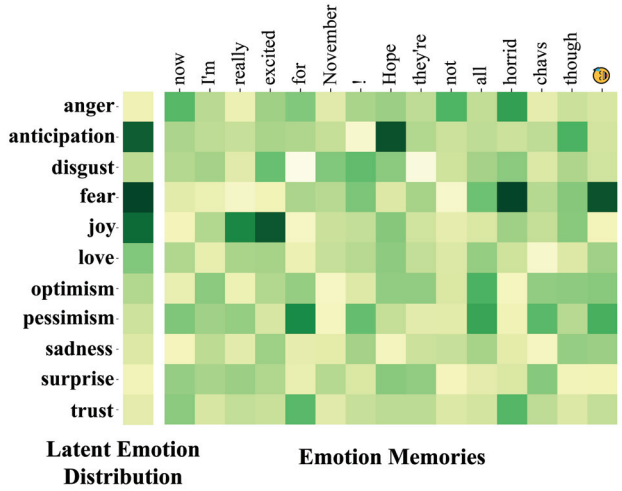

Figure 4: Visualization on an example. The left shows the emotion distribution captured in latent variable $\boldsymbol{Z}$, and the right shows the memories visualization to the corresponding emotion.

Case Study We analyze how the latent emotion module and the emotion memories in LEM together help to make prediction. We select an example from the test set in SemEval2018, which contains three emotions: *anticipation*, *fear* and *joy*. We visualize the latent variable $\boldsymbol{Z}$, and the memories representation from the last hop of the corresponding emotion. Figure 4 shows the results. We can see that the latent emotion distribution correctly captures the emotion intensities distribution. In the example, *anticipation*, *fear* and *joy* receive much higher weights. The corresponding emotion memories provide clues that are closely related to the emotions. For example, the word '*Hope*' indicates the emotion *anticipation*, and the word '*horrid*' and the emoji pattern are strong evidences for the emotion *fear*.

Conclusion

We proposed a Latent Emotion Memory network for multi-label emotion classification. The proposed model could learn the latent emotion distribution without external knowledge, and effectively leverage it into the classification network. Results on two datasets showed that our model outperformed strong baselines, achieving the state-of-the-art performance, demonstrating the usefulness of modeling rich emotion correlations for the task.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61702121, No.61772378), the National Key Research and Development Program of China (No.2017YFC1200500), the Research Foundation of Ministry of Education of China (No.18JZD015), the Major Projects of the National Social Science Foundation of China (No.11&ZD189) and the Science and Technology Project of Guangzhou (No.201704030002).

References

Almeida, A. M.; Cerri, R.; Paraiso, E. C.; Mantovani, R. G.; and Junior, S. B. 2018. Applying multi-label techniques in emotion identification of short texts. *Neurocomputing* 320:35–46.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, 2200–2204.

Bahuleyan, H.; Mou, L.; Vechtomova, O.; and Poupart, P. 2017. Variational attention for sequence-to-sequence models. *arXiv preprint arXiv:1712.08207*.

Baziotis, C.; Athanasiou, N.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; and Potamianos, A. 2018. Ntua-slp at semeval-2018 task 1: predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

Bermingham, A., and Smeaton, A. 2011. On using twitter to monitor political sentiment and predict election results. In *SAAIP*, 2–10.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Caruana, R.; Lawrence, S.; and Giles, C. L. 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *NIPS*, 402–408.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Goyal, A. G. A. P.; Sordoni, A.; Côté, M.-A.; Ke, N. R.; and Bengio, Y. 2017. Z-forcing: Training stochastic recurrent networks. In *NIPS*, 6713–6723.

He, H., and Xia, R. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In *NLPCC*, 250–259.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *KDD*, 168–177.

Huang, S.-J., and Zhou, Z.-H. 2012. Multi-label learning by exploiting label correlations locally. In *AAAI*, 949–955.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*, 2267–2273.

Li, S.; Huang, L.; Wang, R.; and Zhou, G. 2015. Sentence-level emotion classification with label and context dependence. In *ACL*, 1045–1053.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *ICML*, 1727–1736.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. Semeval-2018 task 1: Affect in tweets. In *SemEval*, 1–17.

Nguyen, T. H.; Shirai, K.; and Velcin, J. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications* 42(24):9603–9611.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001–2002.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier. 3–33.

Quan, C., and Ren, F. 2010. Sentence emotion analysis and recognition based on emotion words using ren-cecps. *International Journal of Advanced Intelligence* 2(1):105–117.

Quan, X.; Wang, Q.; Zhang, Y.; Si, L.; and Wenyin, L. 2015. Latent discriminative models for social emotion detection with emotional dependency. *ACM Transactions on Information Systems* 34(1):1–2.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier chains for multi-label classification. In *ECML-PKDD*, 254–269.

Ren, Y.; Zhang, Y.; Zhang, M.; and Ji, D. 2016. Context-sensitive twitter sentiment classification using neutal network. In *AAAI*, 215–221.

Ren, H.; Ren, Y.; Li, X.; Feng, W.; and Liu, M. 2017. Natural logic inference for emotion detection. In *CCL*, 424–436.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Stone, P. J.; Dunphy, D. C.; and Smith, M. S. 1966. The general inquirer: A computer approach to content analysis. *Information Storage and Retrieval* 4(4):375–376.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *NIPS*, 2440–2448.

Tang, D.; Zhang, Z.; He, Y.; Lin, C.; and Zhou, D. 2019. Hidden topic–emotion transition model for multi-level social emotion detection. *Knowledge-Based Systems* 164:426–435.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.

Wang, Y., and Pal, A. 2015. Detecting emotions in social media: A constrained optimization approach. In *IJCAI*, 996–1002.

Wang, Y.; Feng, S.; Wang, D.; Yu, G.; and Zhang, Y. 2016. Multi-label chinese microblog emotion classification via convolutional neural network. In *APWeb*, 567–580.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *EMNLP*, 347–354.

Yang, P.; Sun, X.; Li, W.; Ma, S.; Wu, W.; and Wang, H. 2018. Sgm: sequence generation model for multi-label classification. In *COLING*, 3915–3926.

Yu, J.; Marujo, L.; Jiang, J.; Karuturi, P.; and Brendel, W. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *EMNLP*, 1097–1102.

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.

Zhou, D.; Zhang, X.; Zhou, Y.; Zhao, Q.; and Geng, X. 2016a. Emotion distribution learning from texts. In *EMNLP*, 638–647.

Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016b. Attention-based bidirectional long short-term memory networks for relation classification. In *ACL*, 207–212.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*, 730–738.

Zhou, D.; Yang, Y.; and He, Y. 2018. Relevant emotion ranking from text constrained with emotion relationships. In *NAACL*, 561–571.