Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations

Dong Zhang[1] , **Liangqing Wu**[1] , **Changlong Sun**[2] ,
Shoushan Li[1] , **Qiaoming Zhu**[1*] and **Guodong Zhou**[1]
[1]School of Computer Science and Technology, Soochow University, China
[2]Alibaba Group, China
[1]{dzhang17, lqwu}@stu.suda.edu.cn, {lishoushan, qmzhu, gdzhou}@suda.edu.cn,
[2]changlong.scl@taobao.com

Abstract

Recently, emotion detection in conversations becomes a hot research topic in the Natural Language Processing community. In this paper, we focus on emotion detection in multi-speaker conversations instead of traditional two-speaker conversations in existing studies. Different from non-conversation text, emotion detection in conversation text has one specific challenge in modeling the context-sensitive dependence. Besides, emotion detection in multi-speaker conversations endorses another specific challenge in modeling the speaker-sensitive dependence. To address above two challenges, we propose a conversational graph-based convolutional neural network. On the one hand, our approach represents each utterance and each speaker as a node. On the other hand, the context-sensitive dependence is represented by an undirected edge between two utterances nodes from the same conversation and the speaker-sensitive dependence is represented by an undirected edge between an utterance node and its speaker node. In this way, the entire conversational corpus can be symbolized as a large heterogeneous graph and the emotion detection task can be recast as a classification problem of the utterance nodes in the graph. The experimental results on a multi-modal and multi-speaker conversation corpus demonstrate the great effectiveness of the proposed approach.

1 Introduction

Emotions play an important role in our daily life and emotion detection in text has become a longstanding goal in Natural Language Processing (NLP). In the literature, emotion detection has mainly focused on non-conversation text, such as sentence-level text [Li *et al.*, 2015] and document-level text [Wang *et al.*, 2016]. More recently, emotion detection in conversations has attracted increasing attention in NLP due to its applications in many emerging tasks such as public opinion mining over chat history [Cambria *et al.*, 2017], social media analysis in Facebook , YouTube, Twitter, etc [Majumder *et al.*, 2019], and intelligent systems like smart homes and chatbots [Young *et al.*, 2018].

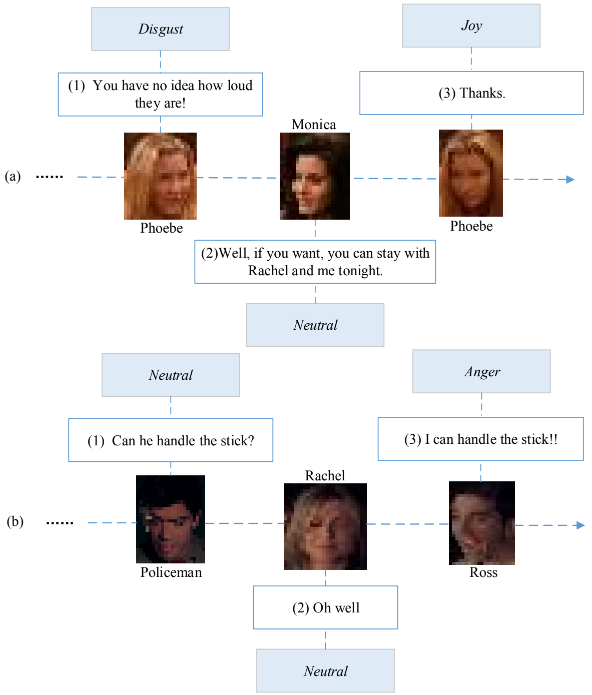
Disgust
(1) You have no idea how loud they are!
Joy
(3) Thanks.
Monica
(a)
Phoebe
Phoebe
(2)Well, if you want, you can stay with Rachel and me tonight.
Neutral
Neutral
Anger
(1) Can he handle the stick?
(3) I can handle the stick!!
Rachel
(b)
Policeman
Ross
(2) Oh well
Neutral

Figure 1: (a) An example for illustrating the context-sensitive dependence; (b) An example for illustrating the speaker-sensitive dependence.

Different from non-conversation cases, nearby utterances in a conversation are closely semantic-related and thus the emotion categories of nearby utterances are closely related. It is important and also challenging to effectively model the context-sensitive dependence in nearby utterances in a conversation. For example, as shown in Figure 1(a), the emotion of the first utterance (said by *Phoebe*) is *disgust*. After the second utterance (said by *Monica*) for comforting Phoebe, the third utterance (said by *Phoebe*) becomes *joyful*. From this example, we can see that the emotion categories could be changed from a *negative* one to a *positive* one when meeting a comforting utterance. A few recent studies, such as [Poria *et al.*, 2017] and [Hazarika *et al.*, 2018b], have realized this

[*]Corresponding author

challenge and proposed some approaches to handling this.

However, all existing studies in emotion detection focus on two-speaker conversations where only two speakers are involved. In this study, we aim to tackle emotion detection in conversations where three or more speakers are involved. In brief, we refer to the conversation in this scenario as multi-speaker conversation. Compared with two-speaker conversations, emotion detection in multi-speaker conversations needs to well model the interactive influence of multiple speakers for a better performance. This is mainly because each speaker has a specific personality and characteristic of uttering which has a significant impact on emotion expressing. Therefore, besides the context-sensitive dependence, it is also important and challenging to effectively model another dependence, namely speaker-sensitive dependence, in utterances for emotion detection in multi-speaker conversations. For example, as shown in Figure 1(b), given a *neutral* utterance (said by *Policeman*), *Rachel* gives a *neutral* reply "*Oh well*" while *Ross* gives an *angry* reply "*I can handle the stick*". The difference of the replying emotions is mainly due to the difference of the two speakers. Specifically, according to the role setting, it is not difficult to notice that *Rachel*'s personality is afraid of trouble while *Ross* is a bit irritable and impatient.

In this paper, we aim to overcome above two challenges in properly modeling both the context-sensitive and speaker-sensitive dependence in multi-speaker conversations. Specifically, we propose a graph-based convolutional neural network towards conversations, namely ConGCN, to model both context-sensitive and speaker-sensitive dependence for emotion detection. On the one hand, each utterance of the whole conversation corpus is represented as a node in a graph, with an edge between the two utterances in the same conversation to symbolize the contextual dependence. On the other hand, each speaker of the whole corpus is represented as a node, and we bridge the specific-speaker dependence between each utterance and its speaker with an undirected edge. On this basis, the entire conversational corpus can be symbolized as a large heterogeneous graph and the emotion detection task can be recast as a classification problem of the utterance nodes in the graph. Experimentation on a multi-modal and multi-speaker conversation corpus shows that our approach could effectively capture the contextual dependence of the utterances in a conversation and the specific-speaker dependence of its corresponding utterance simultaneously. Furthermore, it also shows that our approach is superior in both unimodality and multi-modality.

2 Related Work

As an interdisciplinary research field, emotion detection has been drawing more and more attention in natural language processing and multi-modal communication [Picard, 2010] with focus on exploring various types of features for different-level emotion classification, such as document-level [Alm *et al.*, 2005], sentence-level [Li *et al.*, 2015], and short text-level [Felbo *et al.*, 2017]. Compared with above studies in non-conversational text, the studies in conversations are much less and limited to two-speaker conversations. In the following, we give an overview of emotion detection in two-speaker conversations, together with related studies on graph-based neural networks.

2.1 Emotion Detection in Two-speaker Conversations

Various studies have attributed emotional dynamics as an interactive phenomenon, rather than being within-person and one-directional [Richards, 2010], and modeled them by observing transition properties. Recently, several studies employ memory networks to model the self-speaker and inter-speaker dynamics in two-speaker conversations [Hazarika *et al.*, 2018b; Hazarika *et al.*, 2018a]. Although Majumder *et al.* [2019] claim that their approach can be easily extended to multiple speakers in a conversation, the characteristics of multiple speakers in the whole conversation corpus cannot be effectively highlighted.

Unlike above studies, we distinguish multiple speakers of the whole corpus and model context relationship between utterances of the same conversation with a graph-based neural network. To our best knowledge, this is the first to consider both context-sensitive and speaker-sensitive dependence in multi-speaker conversations.

2.2 Graph-based Neural Networks

Graph-based Neural Networks have received growing attentions recently [Cai *et al.*, 2018]. As a pioneer, Kipf and Welling [2017] present a simplified graph neural network model, called graph convolutional networks (GCN), and achieve the state-of-the-art classification results on a number of benchmark graph datasets. GCN is also exported to several NLP tasks such as semantic role labeling [Marcheggiani and Titov, 2017], text classification [Yao *et al.*, 2019] and machine translation [Bastings *et al.*, 2017], where GCN is used to encode the syntactic structure of sentences.

However, to our best knowledge, there are no studies in the literature adopting a graph-based approach to bridge the contextual relationship between two utterances in a conversation and the speaker-dependent relationship between an utterance and its speaker.

3 Emotion Detection in Multi-speaker Conversations with Conversational GCN

Figure 2 shows the architecture of our proposed ConGCN for emotion detection in multi-speaker conversations.

3.1 Multi-modal Feature Extraction

In this study, we focus on a multi-modal scenario for emotion detection in multi-speaker conversations where both the linguistic context (text) and acoustic characteristics (audio) are employed to infer the emotion of each utterance. Both networks for textual and acoustic feature extraction are trained at utterance level with the emotion labels separately.

Textual Features

The textual features are generated by following [Hazarika *et al.*, 2018a] where these textual features achieve the state-of-the-art performance in multi-modal emotion detection. First,

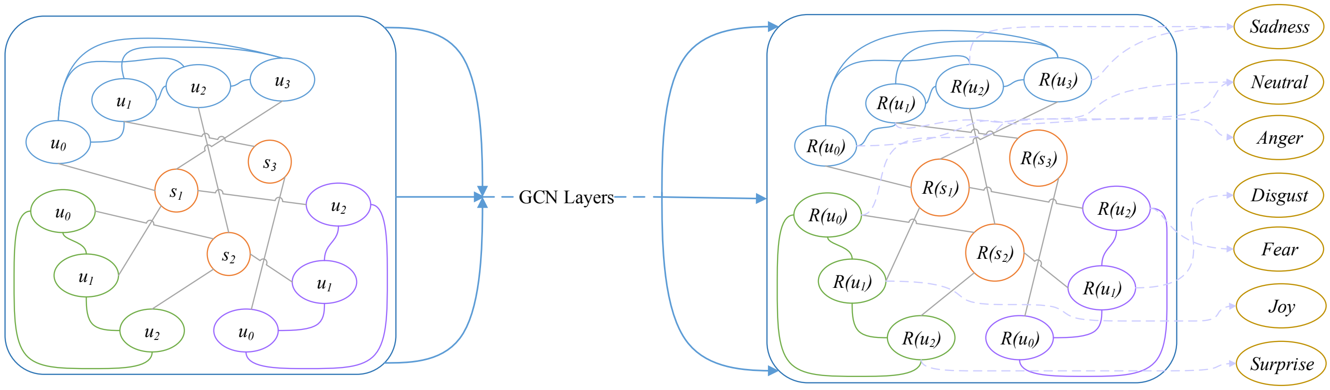

Figure 2: Overall architecture of our proposed approach ConGCN. Nodes with prefix "u" are utterance nodes, others are speakers nodes with prefix "s" . The curve edges denote the utterance-utterance context relationship and the beeline edges denote the speaker-utterance specific speaking relationships. Different colors mean different conversations. $R(x)$ means the representation (embedding) of x. After graph transformation, each node will get a class, i.e., *Sadness*, *Neutral*, *Anger*, *Disgust*, *Fear*, *Joy* or *Surprise*.

the input is obtained with the pre-trained word embeddings extracted from the 300-dimensional GloVe embeddings[Pennington *et al.*, 2014]. Then, a convolutional layer consists of three filters with sizes $f^1; f^2; f^3$ and each filter has f_{out} feature maps. Finally, a max-pooling layer is employed and the pooled features are propagated into two stacked bi-directional LSTM layers. Formally, textual representation of an utterance is denoted as $\boldsymbol{u_t} \in \mathbb{R}^{d_t}$ of dimension d_t.

Acoustic Features

The acoustic features are generated by following [Gu *et al.*, 2018] where these acoustic features achieve the state-of-the-art performance in multi-modal emotion detection. First, each utterance-video is formatted as a 16-bit PCM WAV file. Then, the synchronized frames of static, delta, and double delta feature maps are extracted from the 3D-array MFSCs map of audio file to form the low-level acoustic feature vector $A_j = [s_j, \Delta_j, \Delta\Delta_j]$. Finally, two stacked bi-directional LSTMs are employed to compute the bidirectional acoustic contextual states ($[\overrightarrow{a_j}; \overleftarrow{a_j}]$). Formally, acoustic representation of an utterance is denoted as $\boldsymbol{u_a} \in \mathbb{R}^{d_a}$ of dimension d_a.

3.2 Conversational GCN

Graph neural networks have proved to be effective on tasks that have rich relational structure and should well preserve global structure information in a graph with graph embeddings. In this work, we employ GCN to perform emotion detection in multi-speaker conversations, which considers both the context-sensitive and speaker-sensitive dependence.

Graph Construction

A large graph is constructed for the entire conversation corpus. Specifically, in this graph, all utterances and all speakers are considered as nodes. The edge between two utterance nodes in a conversation is built by contextual information while the edge between an utterance node and its speaker node is built using speaker-dependent relationship. On this basis, the emotion detection task on utterances becomes a node classification problem. Formally, let $G = (V, E)$ represent a conversational graph, where V ($|V| = n$) denotes a set of utterance and speaker nodes. $E \subset V \times V$ is a set of relationships containing context-sensitive and speaker-sensitive dependence. Every node is assumed to be connected to itself, i.e., $(v; v) \in E$ for any v.

Node Representations

(1) For the utterance node, we generate its final representation u by concatenating both textual and acoustic features at multi-modal setting: $u = [\boldsymbol{u_t}; \boldsymbol{u_a}] \in \mathbb{R}^{d_t+d_a}$ of dimension $d_t + d_a$. (2) For the speaker node, we randomly initialize the embedding of different speakers $s \in \mathbb{R}^{d_t+d_a}$ of dimension $d_t + d_a$.

Edge Weighting

Two kinds of edges exist in our conversational graph. The first kind of edges represents the context-sensitive dependence when the two utterances are from the same conversation. In our approach, following [Skianis *et al.*, 2018], we adopt the angular similarity to represent the weight of an edge between two utterances, i.e.,

$$A_{ij} = 1 - \frac{arccos(sim(u_i, u_j))}{\pi} \tag{1}$$

where u_i and u_j denote the vector representations of i-th and j-th utterances in the same conversation.

The second kind of edges represents the speaker-sensitive dependence between one utterance and its speaker. In our approach, we employ the inverse speaking frequency of the speaker, i.e.,

$$A_{ij} = \frac{c}{\boldsymbol{Freq}} \tag{2}$$

where either component of (i, j) is an utterance or its speaker. $\boldsymbol{Freq}$ denotes the utterance number of a speaker in the whole corpus. c is a speaking coefficient to avoid over-fitting, which is a pre-specified hyper-parameter.

Graph Learning

Let $X \in \mathbb{R}^{n \times m}$ be a matrix containing all n nodes with their features, where m is the dimension of the feature vectors, and each row $x_v \in \mathbb{R}^m$ is the feature vector for v. We introduce an adjacency matrix A of G and its degree matrix D, where $D_{ii} = \sum_j A_{ij}$. The diagonal elements of A are set to be 1 because of self-loops.

For a one-layer GCN, the new k-dimensional node feature matrix $X^{(1)} \in \mathbb{R}^{n \times k}$ is computed as:

$$X^{(1)} = \rho(\widetilde{A}XW_0) \tag{3}$$

where $\widetilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix and $W_0 \in \mathbb{R}^{m \times k}$ is a weight matrix. ρ is an activation function, e.g. a ReLU $\rho(x) = max(0, x)$.

For a multi-layer GCN, the node feature matrix is updated by the following formula, i.e.,

$$X^{(l+1)} = \rho(\widetilde{A}X^{(l)}W_l) \tag{4}$$

where l denotes the layer number and $X^{(0)} = X$. In our approach, we employ a two-layer GCN to perform emotion detection in conversations.

Model Training

After transformation by two-layer GCN, the final layer node embeddings are fed into a softmax classifier:

$$Z = softmax(\widetilde{A}\,\mathrm{ReLu}\,(\widetilde{A}XW_0)W_1) \tag{5}$$

where $\widetilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the same as in equation (3), and $softmax(x_i) = \frac{1}{\mathcal{Z}}exp(x_i)$ with $\mathcal{Z} = \sum_i exp(x_i)$. The loss function is defined as the cross-entropy error over all labeled nodes:

$$\mathcal{L} = -\sum_{d \in y_D}\sum_{f=1}^{F} Y_{df} \ln Z_{df} \tag{6}$$

where y_D is the set of node indices that have labels and F is the dimension of the output features, which is equal to the number of classes. Y is the label indicator matrix. The weight parameters W_0 and W_1 can be trained via gradient descent. In equation (5), $R^{(1)}(X) = \widetilde{A}XW_0$ contains the first layer utterance and speaker embeddings and $R^{(2)}(X) = \widetilde{A}\mathrm{ReLU}(\widetilde{A}XW_0)W_1$ contains the second layer utterance and speaker embeddings.

4 Experimentation

In this section, we systematically evaluate our graph-based approach towards emotion detection in multi-speaker conversations.

4.1 Experimental Settings

Dataset

We use a multi-modal and multi-speaker conversational dataset, namely Multi-modal EmotionLines Dataset (MELD)[1] [Poria *et al.*, 2019]. This dataset extends, improves, and further develops the EmotionLines dataset [Hsu *et al.*, 2018] for the multi-modal scenario, including not only textual conversations, but also their corresponding visual and audio counterparts. MELD contains more than 13,000 utterances, 1432 conversations and 304 different speakers. Note that, different from the multi-modal emotion detection datasets, such as IEMOCAP [Busso *et al.*, 2008] and SEMAINE [McKeown *et al.*, 2012], MELD is a multi-speaker conversation dataset where three or more speakers are involved in a conversation. The train and test data split distributed on all different emotions are shown in Table 1. For a better comparison, we perform 10-fold cross-validation in all our experiments.

Emotion	Train+Dev	Test
Anger	1,261	345
Disgust	293	68
Fear	308	50
Joy	1,906	402
Neutral	5,180	1,256
Sadness	795	208
Surprise	1,355	281

Table 1: Statistics of the MELD dataset.

Implementation Details

The hyper-parameters in our approach is set as follows: The embedding size of the convolution layers is set to be 64. The speaking coefficient c is 5. The learning rate is 0.01. The dropout rate is 0.5. L2 loss weight is $5e-4$. In addition, the number of GCN layers is 2 unless otherwise stated. In the built graph, the number of nodes is 15140. Following [Kipf and Welling, 2017], we train ConGCN for a maximum of 200 epochs using Adam [Kingma and Ba, 2015] and stop training if the validation loss does not decrease for 20 consecutive epochs.

Evaluation Metrics and Significance Test

The performance is evaluated using the weighted average F1 measure as [Hazarika *et al.*, 2018a], while the paired t-test is performed to test the significance of the difference between two approaches, with a default significant level of 0.05.

4.2 Baselines

For comparison, we implement following state-of-the-art baseline approaches to emotion detection in conversations in order to comprehensively evaluate the performance of our proposed approach.

Memory Fusion Network (**MFN**)[2] [Zadeh *et al.*, 2018], explicitly accounts for uni-modal and multi-modal interactions in a neural architecture and continuously models them through time. This approach considers neither context-sensitive nor speaker-sensitive dependence. This is the state-of-the-art of multi-modal fusion approach.

Bidirectional Contextual LSTM (**BC-LSTM**)[1] [Poria *et al.*, 2017], performs context-dependent fusion of multi-sequence data. This approach only considers the contextual information in a conversation. As a strong baseline of the MELD dataset, this approach holds the competitive performance.

Conversational Memory Network (**CMN**)[3] [Hazarika *et al.*, 2018b], models separate contexts for both self-speaker

[1] https://github.com/SenticNet/meld

[2] https://github.com/A2Zadeh/CMU-multi-modalSDK

[3] https://github.com/SenticNet/conv-emotion

Modality	Text							
Models	*Anger*	*Disgust*	*Fear*	*Joy*	*Neutral*	*Sadness*	*Surprise*	*w*-average
BC-LSTM	38.9	0.0	0.0	45.8	77.0	0.0	47.3	54.3
CMN	29.8	0.0	0.0	48.3	75.9	21.0	45.7	54.5
ICON	30.1	0.0	0.0	48.5	76.2	18.9	46.3	54.6
DialogueRNN	41.5	0.0	5.4	47.6	73.7	23.4	44.9	55.1
ConGCN w/o context	33.3	0.0	0.0	52.3	71.3	**25.6**	**50.6**	54.3
ConGCN w/o speaker	38.2	0.0	0.0	51.6	**77.3**	0.0	47.3	55.3
ConGCN (Ours)	**43.2**	**8.8**	**6.5**	**52.4**	74.9	22.6	49.8	**57.4**
Modality	Audio							
BC-LSTM	21.9	0.0	0.0	0.0	66.1	0.0	16.0	36.4
CMN	29.6	0.0	0.0	11.8	**67.0**	0.0	2.8	38.3
ICON	31.5	0.0	0.0	8.6	66.9	0.0	0.0	37.7
DialogueRNN	32.1	5.1	0.0	11.2	53.0	8.3	15.6	34.0
ConGCN w/o context	33.6	4.9	3.5	12.4	61.7	4.5	23.9	39.2
ConGCN w/o speaker	31.9	**5.3**	2.4	4.1	59.1	13.5	**26.2**	37.3
ConGCN (Ours)	**34.1**	3.0	**4.7**	**15.5**	64.1	**19.3**	25.4	**42.2**

Table 2: Performance of different approaches to emotion detection on the MELD dataset with uni-modality.

Models	*Anger*	*Disgust*	*Fear*	*Joy*	*Neutral*	*Sadness*	*Surprise*	*w*-average
MFN	40.8	0.0	0.0	46.7	76.2	13.7	40.7	54.7
BC-LSTM	44.5	0.0	0.0	49.7	76.4	15.6	48.4	56.8
CMN	44.7	0.0	0.0	44.7	74.3	23.4	47.2	55.5
ICON	44.8	0.0	0.0	50.2	73.6	23.2	50.0	56.3
DialogueRNN	45.6	0.0	0.0	**53.2**	73.2	24.8	**51.9**	57.0
ConGCN w/o context	43.0	2.4	5.0	49.0	74.9	**29.6**	49.0	57.1
ConGCN w/o speaker	45.9	5.3	7.3	51.6	73.7	28.4	50.5	57.4
ConGCN (Ours)	**46.8**	**10.6**	**8.7**	53.1	**76.7**	28.5	50.3	**59.4**

Table 3: Performance of different approaches to emotion detection on the MELD dataset with multi-modality.

and other-speaker to an utterance for emotion detection in two-speaker conversations, but ignores the global contextual information of the utterance. To our best knowledge, this approach is a pioneer to explore multi-modal emotion detection in two-speaker conversations. In our implementation, for fair comparison, in the context of a test utterance, the different speakers' utterances are distinguished with different gated recurrent units (GRUs).

Interactive Conversational Memory Network (**ICON**)[3] [Hazarika *et al.*, 2018a], models separate contexts for both self-speaker and other-speaker to an utterance for emotion detection in two-speaker conversations, and simultaneously incorporates the global contextual information of the test utterance. This approach represents state-of-the-art of uni-modal and multi-modal emotion detection in conversations. In our implementation, for fair comparison, in the context of a test utterance, the different speakers' utterances are distinguished with different GRUs.

Attentive RNN (**DialogueRNN**)[3] [Majumder *et al.*, 2019], employs three GRUs to model the speaker, the context from the preceding utterances, and the emotion of the preceding utterances. This approach considers different speakers of a conversation in a two-speaker scenario. This approach represents the state-of-the-art of uni-modal and multi-modal emotion detection in conversations. In our implementation, for fair comparison, we set the speaker representation with one-hot encoding of all speakers in the whole corpus.

A variation of our approach (**ConGCN w/o context**), removes the context-sensitive dependence modeling.

A variation of our approach (**ConGCN w/o speaker**), removes the speaker-sensitive dependence modeling.

4.3 Experimental Results with Uni-modality

Table 2 illustrates the performance of various approaches with uni-modaltiy performed on MELD. From this table, we can see that:

In the text modality, those approaches including **CMN**, **ICON**, and **DialogueRNN**, which deal with the speaker information, are all superior than **BC-LSTM**, which highlights the importance of using the speaker information. Among all baseline approaches, **DialogueRNN** averagely performs best, achieving 55.1% in terms of *w*-average. In comparison, our approach **ConGCN** performs better than all baseline approaches and the significance test shows that the average improvements over the baseline approaches are all significant (p-$value < 0.05$). Moreover, our approach outperforms two ablated approaches **ConGCN w/o context** and **ConGCN w/o speaker**. This indicates the necessity to consider both the context-sensitive and speaker-sensitive dependence for emotion detection in multi-speaker conversations.

In the acoustic modality, among the baseline approaches, it seems that speaker information is not always helpful since **DialogueRNN** performs no better than **BC-LSTM**. This may be due to the fact that **DialogueRNN** cannot well deal with

Our approach (GCN layers' number)	Textual modality	Acoustic modality	Multi-modality
ConGCN (l=1)	55.8	38.8	56.1
ConGCN (l=2)	**57.4**	**42.2**	**59.4**
ConGCN (l=3)	56.1	39.8	57.6
ConGCN (l=4)	48.0	36.2	52.2

Table 4: Weighted average performance with different features and different numbers of GCN layers.

Speaker Name	Type	*Anger*	*Disgust*	*Fear*	*Joy*	*Neutral*	*Sadness*	*Surprise*
Doug	Number of his/her utterances	1	0	0	27	15	1	0
	Probability distribution	0.180	0.094	0.025	0.285	0.211	0.111	0.094
Emily	Number of his/her utterances	13	0	2	6	27	4	7
	Probability distribution	0.250	0.046	0.055	0.082	0.340	0.072	0.155
Tag	Number of his/her utterances	4	0	0	10	35	4	7
	Probability distribution	0.073	0.061	0.030	0.148	0.515	0.062	0.112

Table 5: The numbers of some speakers' utterances in each emotion category and the predicted probability distribution of these speakers.

the speaker information under the acoustic modality. However, among all approaches, our approach performs best and it averagely outperforms the best-performed baseline approach (**CMN**) by about 4%. This also applies to two ablated approaches, which indicate that it is necessary to model both the context- and speaker-sensitive dependence.

4.4 Experimental Results with Multi-modality

Table 3 illustrates the performance of various approaches with multi-modaltiy performed on MELD. From this table, we can see that the performance of **MFN** is apparently lower than the other approaches. For example, the F1 score of **MFN** is averagely 2.1% lower than **BC-LSTM**. This indicates that either context or speaker information is helpful when multi-modality features are employed. Among all approaches, our approach performs best. Furthermore, compared with the two ablated approaches **ConGCN w/o context** or **ConGCN w/o speaker**, **ConGCN** is apparently superior. This highlights the importance of employing both the context-sensitive and speaker-sensitive dependence.

5 Analysis and Discussion

5.1 Impact of Multi-layer GCN

Table 4 illustrates the results of our approach when using different numbers of GCN layers. It shows that, whether using uni-modality features or multi-modality features, as the number of GCN layers increases, the performance of our proposed **ConGCN** first rises and then decreases. It also shows our approach yields the best results when using a two-layer GCN for emotion detection in multi-speaker conversations. This indicates that multi-layer GCN is indeed better than one-layer to some extent, but more than two layers of GCN in our approach seems not a good choice.

5.2 Effectiveness of Speaker Prediction

Our approach treats both the utterances and speakers as nodes for prediction. In this section, we compare the predicted label of a speaker with the distribution of his/her utterances' emotion categories to check whether they are consistent. Table 5 illustrates the number of one speaker's utterances in each emotion category together with the predicted probability distribution over all emotion categories. From this table, we can see that: (1) Some speakers tend to say some words with particular emotion categories. For instance, the speaker *Doug* tens to say '*Joy*' and '*Neural*' utterances while *Emily* tends to say '*Neural*' and '*Anger*" utterances. (2) Our approach performs well in predicting the emotion categories of a speaker. For instance, the speaker *Doug* is classified to be *joy*, which is consistent to the number distribution of *Doug*'s utterances.

6 Conclusion

In this paper, we propose a conversational graph-based neural network, namely ConGCN, to model both context-sensitive and speaker-sensitive dependence, for emotion detection in multi-speaker conversations. In our graph-based approach, both utterances and speakers are modeled as nodes while both the context-sensitive dependence and the speaker-sensitive dependence are modeled as edges. Empirical evaluation on a multi-modal and multi-speaker dataset shows that our approach significantly outperforms several state-of-the-art approaches. This indicates the importance of both the context- and speaker-sensitive dependence to emotion detection in multi-speaker conversations and the effectiveness of our graph-based approach in well modeling such dependence.

In our future work, we would like to improve the performance of emotion detection by using unlabeled data since our graph-based neural network approach is easy to add unlabeled data. Moreover, we would like to apply our approach to other applications in conversations, such as sentiment classification and topic classification.

Acknowledgments

The research work is partially supported by the Key Project of NSFC No.61836007 and two NSFC grants No.61672366, No.61673290. This work is also supported by the joint research project of Alibaba and Soochow University.

References

[Alm *et al.*, 2005] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *HLT/EMNLP*, pages 579–586, 2005.

[Bastings *et al.*, 2017] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. In *EMNLP*, pages 1957–1967, 2017.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[Cai *et al.*, 2018] HongYun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, 30(9):1616–1637, 2018.

[Cambria *et al.*, 2017] Erik Cambria, Soujanya Poria, Alexander F. Gelbukh, and Mike Thelwall. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80, 2017.

[Felbo *et al.*, 2017] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*, pages 1615–1625, 2017.

[Gu *et al.*, 2018] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Hybrid attention based multimodal network for spoken language classification. In *COLING*, pages 2379–2390, 2018.

[Hazarika *et al.*, 2018a] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON: interactive conversational memory network for multimodal emotion detection. In *EMNLP*, pages 2594–2604, 2018.

[Hazarika *et al.*, 2018b] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *NAACL-HLT*, pages 2122–2132, 2018.

[Hsu *et al.*, 2018] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. In *LREC*, 2018.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[Li *et al.*, 2015] Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. Sentence-level emotion classification with label and context dependence. In *ACL*, pages 1045–1053, 2015.

[Majumder *et al.*, 2019] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. Dialoguernn: An attentive RNN for emotion detection in conversations. In *AAAI*, 2019.

[Marcheggiani and Titov, 2017] Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *EMNLP*, pages 1506–1515, 2017.

[McKeown *et al.*, 2012] Gary McKeown, Michel François Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affective Computing*, 3(1):5–17, 2012.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Picard, 2010] Rosalind W. Picard. Affective computing: From laughter to ieee. *IEEE Transactions on Affective Computing*, 1(1):11–17, 2010.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *ACL*, pages 873–883, 2017.

[Poria *et al.*, 2019] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, 2019.

[Richards, 2010] Jane M. Richards. The cognitive consequences of concealing feelings. *Current Directions in Psychological Science*, 13(4):131–134, 2010.

[Skianis *et al.*, 2018] Konstantinos Skianis, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. Fusing document, collection and label graph-based representations with word embeddings for text classification. In *NAACL-HLT*, pages 49–58, 2018.

[Wang *et al.*, 2016] Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xue-Jie Zhang. Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL*, 2016.

[Yao *et al.*, 2019] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *AAAI*, 2019.

[Young *et al.*, 2018] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *AAAI*, pages 4970–4977, 2018.

[Zadeh *et al.*, 2018] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, pages 5634–5641, 2018.