CONSK-GCN: CONVERSATIONAL SEMANTIC- AND KNOWLEDGE-ORIENTED GRAPH CONVOLUTIONAL NETWORK FOR MULTIMODAL EMOTION RECOGNITION

Yahui Fu[1,2], *Shogo Okada*[2], *Longbiao Wang*[1], *Lili Guo*[1], *Yaodong Song*[1], *Jiaxing Liu*[1] *and Jianwu Dang*[1,2]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{fuyahui,longbiao_wang,liliguo,songyaodong,jiaxingliu}@tju.edu.cn,{okada-s,jdang}@jaist.ac.jp

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428438

ABSTRACT

Emotion recognition in conversations (ERC) has received significant attention in recent years due to its widespread applications in diverse areas, such as social media, health care, and artificial intelligence interactions. However, different from nonconversational text, it is particularly challenging to model the effective context-aware dependence for the task of ERC. To address this problem, we propose a new Conversational Semantic- and Knowledge-oriented Graph Convolutional Network (ConSK-GCN) approach that leverages both semantic dependence and commonsense knowledge. First, we construct the contextual inter-interaction and intradependence of the interlocutors via a conversational graph-based convolutional network based on multimodal representations. Second, we incorporate commonsense knowledge to guide ConSK-GCN to model the semantic-sensitive and knowledge-sensitive contextual dependence. The results of extensive experiments show that the proposed method outperforms the current state of the art on the IEMOCAP dataset.

Index Terms— Graph convolutional network, commonsense knowledge, conversational multimodal emotion recognition

1. INTRODUCTION

Emotion recognition, which is the subtask of affective computing, has remained the subject of active research for decades. In the literature, emotion recognition has mainly focused on nonconversational text, audio, or visual information extracted from a single utterance while ignoring contextual semantics. Deep learning methods such as the deep neural network (DNN)[1], convolutional neural network (CNN)[2], and recurrent neural network (RNN)[3] are the most commonly used architectures for emotion recognition and usually achieve competitive results.

Shogo Okada and Longbiao Wang are corresponding authors. This work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers 19H01120, 19H01719 and JST AIP Trilateral AI Research, Grant Number JPMJCR20G6, Japan.

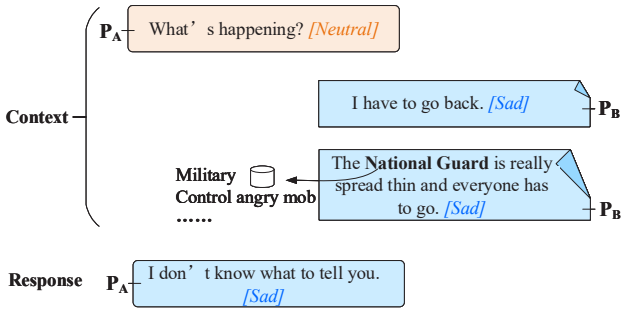

Fig. 1. An example conversation with annotated labels from the IEMOCAP dataset [4]. In this conversation, P_A 's emotion changes are influenced by the contextual information of P_B. By incorporating an external knowledge base, the concept "National Guard" in the third utterance is enriched by associated terms such as "Military" and "Control angry mob". Thus, the implicit emotion in the third utterance can be inferred more easily via its enriched meaning.

More recently, emotion recognition in conversations (ERC) has attracted increasing attention because it is a necessary step for a number of applications, including opinion mining over chat history, social media threads (such as YouTube, Facebook, Twitter), human-computer interaction, and so on. Different from nonconversation cases, nearby utterances in a conversation are closely related to semantics and emotion. Furthermore, we assume that the emotion of the target utterance is usually strongly influenced by the nearby context (Fig. 1). Thus, it is important but challenging to effectively model the context-sensitive dependence among the conversations.

RNN-based methods such as bc-LSTM[5] apply bidirectional long short-term memory (BLSTM) to propagate contextual information to the utterances and process the constituent utterances of a dialogue in sequence. However, this approach faces the issue of context propagation and may not perform well on long-term contextual information [6]. To mitigate this issue, some variants like DialogueRNN [7] integrate with an attention mechanism that can dynamically fo-

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

cus on the most relevant contexts. However, this attention mechanism does not consider the relative position of the target and context utterances, which is important for modeling how past utterances influence future utterances and vice versa. DialogueGCN [8] and ConGCN [9] employ a graph convolutional neural network (GCN) to model the contextual dependence and all achieve a new state of the art, proving the effectiveness of the GCN on context structure. However, both DialogueGCN and ConGCN only consider the semantic information between utterances. Thus, for implicit emotional texts that do not contain obvious emotional terms, and the words are relatively objective and neutral, it is difficult to correctly distinguish the emotions if only the semantics of the utterances are considered. Both semantic context and commonsense knowledge are essential for the machine to analyze emotion in conversations. Figure 1 shows an example demonstrating the importance of context and knowledge in the detection of the accurate emotion of implicit emotional texts. In the literature, only a limited number of studies have explored the incorporation of context and commonsense knowledge via GCN for the ERC task.

To further the above problems, we propose a new multimodal Semantic- and Knowledge-oriented Graph Convolutional Network (ConSK-GCN) to effectively structure the semantic-sensitive and knowledge-sensitive contextual dependence in each conversation. On the one hand, we construct the contextual inter-interaction and intradependence of the interlocutors via a conversational semantic-oriented GCN (ConS-GCN). In this context graph, each utterance can be seen as a single node, and the relational edges between a pair of nodes/utterances represent the dependence between the speakers of these utterances. On the other hand, we incorporate an external knowledge base that is fundamental to understand conversations to enrich the semantic meaning of the tokens in the utterance via a conversational knowledge-oriented GCN (ConK-GCN). Furthermore, we introduce an affective lexicon into knowledge graph construction to enrich the emotional polarity of each concept. Furthermore, we leverage the semantic edge weights and affect enriched knowledge edge weights to construct a new adjacency matrix of our ConSK-GCN for better performance in the ERC task.

2. RELATED WORK

In the literature, graph convolution networks [10] such as text classification [11], emotion recognition in conversations [8] [9] have been widely used in recent years, and have achieved competitive performance, where GCN is used to encode the syntactic structure of sentences. Unlike the above studies, our approach encodes both knowledge-sensitive and semantic-sensitive contextual dependence via GCN.

The knowledge base has attracted increasing attention in several research areas such as open-domain dialogue systems [12], and emotion detection in conversations [13]. Knowledge bases provide a rich source of background concepts related by commonsense links, which can enhance the semantics of a piece of text by providing context-specific concepts. [13] makes use of knowledge base by concatenating the concept embedding and word embedding as the input to the Transformer architecture. However, using external knowledge as the initial input of the model has limited utility in helping the model to build effective contextual dependence. Different from these studies, we incorporate the knowledge base and semantic dependence via new ConSK-GCN to capture both semantic-aware and knowledge-aware contextual emotion features.

3. DATA PREPROCESSING

To better mine the information of the raw data and capture efficient contextual traits, we preprocess the textual features, acoustic features, and knowledge base respectively.

3.1. Multimodal features extraction

In this study, we focus on multimodal emotion recognition in conversations with acoustic and textual characteristics, which are complementary to emotion information and result in a decent performance. Furthermore, we train separate networks to extract linguistic and acoustic features at the utterance level with emotion labels.

3.1.1. Textual features

We employ the most used CNN [2] to extract textual embeddings of the transcripts. First, we use the publicly available pretrained word2vec [14] to initialize the word vectors. Then, we use one convolutional layer followed by one max-pooling and two fully connected layers to obtain deep feature representations for each utterance. We use convolutional filters of size 3, 4, and 5 with 100 feature maps in each. The window size of max-pooling is set to 2 followed by the ReLU activation [15]. These are then concatenated and fed into two fully connected layers with 500 and 100 hidden nodes separately followed by the ReLU activation. Formally, the textual representation of an utterance is denoted as $\boldsymbol{\mu_t}$.

3.1.2. Acoustic features

In this paper, we use the same audio preprocessing method introduced in [16]. The time of each segment is 265-*ms* and the input spectrogram has the following *time* $\times$ *frequency*: 32 $\times$ 129. A CNN is utilized to extract deep acoustic features from the segment-level spectrograms. Then, the segment-level features are propagated into the BLSTM with 200 dimensions to extract sequential information within each utterance. Finally, the features are fed into a single fully connected layer with 512 dimensions at the utterance level for emotion classification. Formally, the acoustic representation of an utterance is denoted as $\boldsymbol{\mu_a}$.

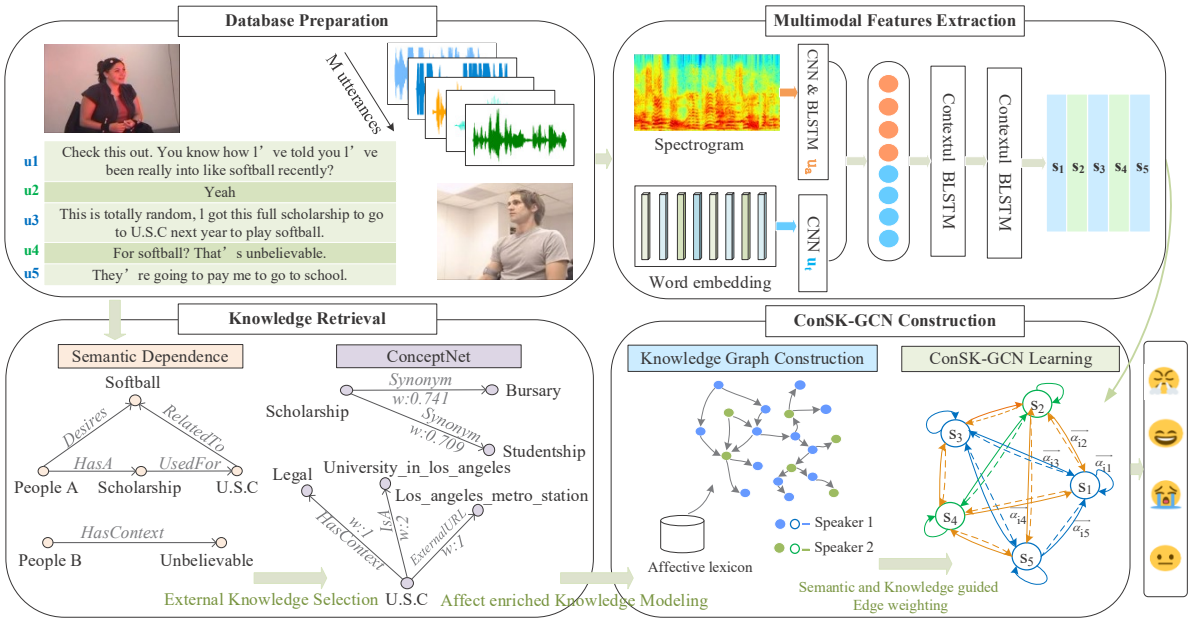

Fig. 2. Overall architecture of our proposed ConSK-GCN approach for multimodal emotion recognition

3.1.3. Multimodal features

After obtaining the textual and acoustic features in an utterance, we concatenate the embeddings of these two modalities $\boldsymbol{\mu} = [\boldsymbol{\mu_t}; \boldsymbol{\mu_a}]$, and then feed the concatenated embeddings into two stacked BLSTM for sequence encoding to obtain the global contextual information. Formally, we denote the context-aware multimodal representations as $\boldsymbol{s}$.

3.2. Knowledge retrieval

In this paper, we utilize external commonsense knowledge base ConceptNet [17] and an emotion lexicon NRC_VAD [18] as the knowledge sources in our approach.

ConceptNet is a large-scale multilingual semantic graph that connects words and phrases of natural language with labeled weighted edges and is designed to assist natural language applications to better understand the meanings behind the words used by people. The nodes in ConceptNet are concepts and the edges represent relation. As shown in Figure2, each <concept1, relation, concept2> triplet is an assertion, and each assertion is associated with a single confidence score. For example, *"scholarship has synonym of bursary with confidence score of 0.741"*. Then we select the corresponding concepts based on the semantic dependence of each conversation.

NRC_VAD lexicon includes a list of more than 20,000 English words with their valence (V), arousal (A), and dominance (D) scores. The real-valued scores for VAD are on a scale of 0-1 for each dimension respectively, corresponding to the degree from low to high.

4. CONVERSATIONAL SEMANTIC- AND KNOWLEDGE-ORIENTED GRAPH CONVOLUTIONAL NETWORK CONSTRUCTION

Figure 2 shows the architecture of our proposed ConSK-GCN approach for multimodal emotion recognition.

4.1. Knowledge graph construction

We build the knowledge graph $G_k = (V_k, E_k, V, A)$ based on the conversational knowledge-aware dependence, where V_k is a concept set and E_k is a link set, and $E_k \subset V_k \times V_k$ is a set of relation that represent the relatedness among the knowledge concepts. In addition, for the concepts in V_k, we retrieve the corresponding valence (V) and arousal (A) scores from NRC_VAD, respectively.

Each node/concept in the knowledge graph is embedded into a single effective semantic space, named ConceptNet Numberbatch [17], that learns from both distributional semantics and ConceptNet. The tokens that are not included in the ConceptNet are initialized by the "fastText" method [19], which is a library for efficient learning of word representations. For the concept not in the NRC_VAD, we set the VAD value to 0.5 as a neutral score.

The edges in the knowledge graph represent the knowledge relatedness between the concepts. First, for each concept $c_{i,m}$ in utterance i, we adopt l_2 norm to compute the emotion intensity emo_m, that is,

$$emo_m = min - max(\|[V(c_{i,m}) - 1/2, A(c_{i,m})/2]\|_2) \quad (1)$$

where $m = 1, ...n$, and n is the number of concepts in each utterance. $||.||_2$ denotes l_2 norm, $V(c_{i,m})$ and $A(c_{i,m})$

represent the corresponding valence and arousal score for each concept in utterance i. Then, we consider the past context window size of p and future context window size of f, and knowledge edge weights $a^k_{i,j}$ are defined as below:

$$k_{i,m} = emo_m c_{i,m} \tag{2}$$

$$a^k_{i,j} = \sum_{m=1}^{n} abs\left(\cos(k^\top_{i,m} W_k [\, k_{i-p,m}, ..., k_{i+f,m}])\right) \tag{3}$$

where $k_{i,m}$ is the affect enriched knowledge of concept m in utterance i, and $j = i-p, ..., i+f$, W_k is a learnable parameters matrix.

4.2. Semantic graph construction

We build the semantic graph $G_s = (V_s, E_s)$ based on the conversational semantic-aware dependence, where V_s denotes a set of utterance nodes, and $E_s \subset V_s \times V_s$ is a set of relations that represent the semantic similarity among the utterances.

The node features in the semantic graph are the multimodal representation $\boldsymbol{s}$. The edges in the semantic graph represent the semantic-sensitive context similarity within each conversation. We adopt the method proposed in [20] to compute the semantic similarity between two utterances, that is,

$$sim_{i,j} = 1 - \arccos\left(\frac{s_i^\top s_j}{\|s_i\| \, \|s_j\|}\right)/\pi \tag{4}$$

Then, the edge weights in the semantic graph is formulated as:

$$a^s_{i,j} = softmax(W_s[sim_{i-p}, ..., sim_{i+f}]) \tag{5}$$

where s_i, s_j denote the multimodal representation of i-th and j-th utterance in the same conversation respectively, and W_s is a trainable parameter matrix.

4.3. ConSK-GCN learning

We build our semantic- and knowledge-oriented graph as $G_{sk} = (V_s, E_{sk})$. To incorporate both knowledge-sensitive and semantic-sensitive contextual features, we leverage the addition of the edge weights of knowledge graph $(a^k_{i,j})$ and the edge weights of semantic graph $(a^s_{i,j})$ as our adjacency matrix E_{sk}, that is,

$$a_{i,j} = \omega_k a^k_{i,j} + (1-\omega_k) a^s_{i,j} \tag{6}$$

where ω_k is a model parameter balancing the impacts of knowledge and semantics on computing the contextual dependence in each conversation. Then, we feed the global contextual multimodal representations $\boldsymbol{s}$ and edge weights $a_{i,j}$ into a two-layer GCN [21] to capture local contextual information that is both semantic-aware and knowledge-aware:

$$h^{(1)}_i = \sigma\Big(\sum_{r\in\Re}\sum_{j\in N^r_i} \frac{a_{i,j}}{q_{i,r}} W^{(1)}_r s_j + a_{i,i} W^{(1)}_0 s_i\Big) \tag{7}$$

$$h^{(2)}_i = \sigma\Big(\sum_{j\in N^r_i} W^{(2)} h^{(1)}_j + W^{(2)}_0 h^{(1)}_i\Big) \tag{8}$$

where N^r_i denotes the neighboring indices of each node under relation $r \in \Re$, $\Re$ contains relations both in the canonical direction (e.g.*born_in*) and in the inverse direction (e.g.*born_in_inv*). $q_{i,r}$ is a problem-specific normalization constant that can either be learned or chosen in advance (such as $q_{i,r} = |N^r_i|$), and $W^{(1)}_r$,$W^{(1)}_0$,$W^{(2)}$,$W^{(2)}_0$ are model parameters, $\sigma(.)$ is the activation function such as ReLU.

5. EXPERIMENTATION

5.1. Experiment dataset

We evaluate our ConSK-GCN on a multimodal conversational dataset, namely Interactive Emotional Dyadic Motion Capture (IEMOCAP) [4]. We use 5531 utterances in 151 dialogues with four emotion categories with the distribution of 29.6% happy, 30.9% neutral, 19.9% anger, and 19.6% sad. In this paper, we use the first eight speakers from sessions 1-4 as the training set and use session five as the test set to perform speaker-independent emotion recognition. We choose cross-entropy as the cost function, Adam as the optimizer, and ReLU as the activation. We set the batch size and number of epochs to 32 and 100, respectively. The window sizes of the past and future contexts are all set to 10 because we have verified that window sizes of 8-12 show better performance. The learning rate is 0.00005 for multimodality and 0.0001 for unimodality training. And ω_k is set to 0.5 to balance the effect of knowledge and semantics.

5.2. Comparison methods

For a comprehensive evaluation, we compare our method with the current advanced approaches and with the results of the ablation studies. All of the experiments are trained on the utterance-level.

CNN[2]: A widely used architecture for both text and audio feature extraction with strong effective performance. We employ it to extract utterance-level textual and acoustic features; it does not contain contextual information.

LSTMs[3]: Adopted LSTM framework for unimodality and multimodality emotion recognition based on audio and text, without exploring context information.

bc-LSTM[5]: Utilized bidirectioinal LSTM network that takes as input the sequence of utterances in a video and extracts contextual unimodal and multimodal features by modeling the dependencies among the input utterances.

DialogueRNN[7]: Employed three GRUs to model the dynamics of the speaker states, the context from the preceding utterances and the emotion of the preceding utterances respectively. This method achieved state of the art in multimodal emotion recognition in conversations.

DialogueGCN[8]: Adopted GCN to leverage self and interspeaker dependence of the interlocutors to model conversational context for textual emotion recognition.

Table 1. Comparative experiments of different methods for unimodality (Text) emotion recognition. Acc.= Accuracy; Average (w)= Weighted average; bold font denotes the best performances.

	Models	Neutrality		Anger		Happiness		Sadness		Average(W)	
		Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)
Baselines	CNN [2]	59.11	59.50	77.06	65.17	64.03	69.36	62.04	60.68	63.90	64.02
	LSTMs [3]	72.92	74.97	70.00	65.93	55.20	56.42	63.67	61.30	64.38	64.42
	bc-LSTM [5]	76.04	67.51	75.88	72.88	67.65	75.51	67.35	70.06	71.31	71.60
	DialogueRNN [7]	**81.51**	73.73	66.47	74.10	86.43	87.82	72.24	77.29	79.37	79.50
	DialogueGCN [8]	74.22	74.32	77.06	76.61	87.56	88.66	85.31	83.60	81.57	81.55
Ablation Studies	ConS-GCN	76.04	74.68	77.65	77.65	87.33	88.74	83.27	83.27	81.71	81.79
	ConK-GCN	75.52	75.23	77.65	77.88	86.65	88.05	86.12	84.06	81.87	81.90
Proposed	ConSK-GCN	74.48	**75.66**	**80.00**	**78.84**	**87.78**	**88.79**	**89.39**	**86.39**	**82.92**	**82.89**

Table 2. Comparative experiments of different methods for multimodality (Text+Audio) emotion recognition.

	Models	Neutrality		Anger		Happiness		Sadness		Average(W)	
		Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)
Baselines	LSTMs [3]	69.53	63.95	73.53	73.10	66.74	73.75	70.61	67.98	69.30	69.50
	bc-LSTM [5]	79.95	70.49	78.82	77.91	70.14	78.58	73.88	75.73	75.10	75.42
	DialogueRNN [7]	**86.20**	76.53	84.71	83.72	79.64	86.38	75.10	80.35	81.47	81.78
Ablation Studies	ConS-GCN	78.91	77.79	85.29	83.57	90.72	90.52	78.78	82.13	83.96	83.97
	ConK-GCN	75.78	77.70	**88.82**	**85.31**	89.37	90.08	**86.53**	84.46	84.53	84.49
Proposed	ConSK-GCN	78.13	**79.89**	87.06	84.33	**93.67**	**91.90**	82.86	**84.76**	**85.82**	**85.74**

ConS-GCN: Consider the semantic-sensitive contextual dynamics in the range of past p and future f window size based on semantic graph.

ConK-GCN: We replace the semantic graph by knowledge graph, which explores the contextual dynamics based on concept relatedness in conversations.

ConSK-GCN: Integrating ConS-GCN and ConK-GCN jointly to leverage the semantic and knowledge contribution to construct the new adjacency matrix of ConSK-GCN.

5.3. Experimental results and analysis

5.3.1. Comparation in unimodality

Table 1 indicates the performance of both state-of-the-art and our ablation studies for emotion recognition based on text modality. From this table, we observe that, the methods that consider the context are much more effective than the methods that do not, demonstrating the significance of context modeling. Among all of the baselines, "DialogueGCN" shows the best performance because it extracts information of the neighborhood contexts based on the graph convolution network, and the emotion of the target utterance is usually strongly influenced by nearby context.

According to the emotion theory introduced in [22] that the Valence-Arousal space depicts the affective meanings of linguistic concepts. We believe that both *Anger* and *Happiness* are explicit emotions in linguistic features with positive arousal, which are also contagious in the context. Therefore, the information extracted both through "ConS-GCN" and "ConK-GCN" that based on context construction affect similar for recognizing them. By contrast, *Sadness* is relatively implicit in linguistic characteristics with negative valence and negative arousal. Compared to "ConS-GCN", "ConK-GCN" have a significant improvement in *Sadness* detection, and we observe that the recognition accuracy has increased by almost 3% as shown in Table 1, while it shows a more significant increase by nearly 8% in Table 2. This illustrates the effectiveness of constructing knowledge graph for contextual features extraction in the ERC task, particularly in the analysis of implicit emotional utterances.

Encouragingly, the comparison shows that our proposed "ConSK-GCN" performs better than all of the baseline approaches, with improvement of at least 1.3% in terms of average accuracy and F1. Furthermore, "ConSK-GCN" also performs better than baselines and ablation studies for each emotion detection in terms of F1. These results indicate that the knowledge-aware contexts and semantic-aware contexts are complementary for extracting efficient contextual features.

5.3.2. Comparation in multimodality

Table 2 describes the performance of various approaches for emotion recognition based on text and audio modalities. An examination of the results presented in this table shows that compared with the multimodal baselines, our proposed "ConSK-GCN" method displays the best performance with near 4% improvement in terms of both average accuracy and F1. This result highlights the importance of integrating semantic-sensitive and knowledge-sensitive contextual information for emotion recognition.

Furthermore, compared with unimodality in Table 1, the

detection accuracy in *Neutrality*, *Anger* and *Happiness* have been improved by 3.65%, 7.06% and 5.89% respectively via the proposed "ConSK-GCN" with multimodality. These demonstrates the importance of integrating acoustic and linguistic features that are complementary in emotion recognition. However, there is an exception in *Sadness* detection that we assume is due to the negative valence and negative arousal emotion of *Sadness* so that similar to text features, the acoustic characteristics of *Sadness* are also implicit.

Utterances	Gold_label	ConS-GCN	ConSK-GCN	Knowledges
But if that can't happen, I'll just have to get out.	A	N ✗	A ✓	Escape, Difficulty
We will go out to dinner later this week.	H	N ✗	H ✓	A good time for socialization, Party
Being dishonest with him. It is the kind of thing that pays off.	S	N ✗	S ✓	Hurt someone else, Deceitful
Cool, if you want me to go with you, I will.	N	N ✓	S ✗	Unemotional, Chill, Unfriendly

Fig. 3. Visualization of several representative examples. Blue denotes the typical concept in each utterance.

5.4. Case study

To verify the effectiveness of external knowledge in conversational emotion recognition, we visualize several typical samples, as shown in Figure 3. We can observe that compared to "ConS-GCN", which only considers the semantics of context, our proposed "ConSK-GCN" that leverages both semantic- and knowledge-aware contextual information can effectively capture implicit emotional features, as shown in the utterance 1-3, that do not contain obvious emotional terms. However, our model misclassifies the *Neutrality* emotion of utterance 4; we attribute this result to the fact that the concept embeddings of the utterance are enriched by emotional knowledge, misleading the model and resulting in wrong detection.

6. CONCLUSION

In this paper, we proposed a new conversational semantic- and knowledge-oriented graph convolutional network (ConSK-GCN) for multimodal emotion recognition. In our approach, we construct the contextual interactions of inter- and intra-speaker via a conversational graph-based convolutional network based on multimodal representations. Then incorporate semantic graph and commonsense knowledge graph jointly to model the semantic-sensitive and knowledge-sensitive contextual dynamics. Comparative experiments on IEMOCAP show that our approach significantly outperforms the state of the art, illustrating the importance of both the semantic and commonsense knowledges in contextual emotion recognition. In our future work, we will employ our approach in multi-speaker conversations and model the speaker dynamics and emotion shifts for better emotion recognition.

7. REFERENCES

[1] E. Kim and J. W. Shin, "Dnn-based emotion recognition based on bottleneck acoustic features and lexical features," in *ICASSP*, 2019, pp. 6720–6724.

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

[3] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[4] C. Busso, M. Bulut, C. Lee, and et al., "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[5] S. Poria, E. Cambria, and et al, "Context-dependent sentiment analysis in user-generated videos," in *ACL*, 2017, pp. 873–883.

[6] J. Bradbury, S. Merity, C. Xiong, and et al., "Quasi-recurrent neural networks," *arXiv preprint arXiv:1611.01576*, 2016.

[7] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and et al., "Dialoguernn: An attentive rnn for emotion detection in conversations," in *AAAI*, 2019, vol. 33, pp. 6818–6825.

[8] D. Ghosal, N. Majumder, S. Poria, and et al., "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *EMNLP-IJCNLP*, 2019, pp. 154–164.

[9] D. Zhang, L. Wu, C. Sun, and et al., "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations.," in *IJCAI*, 2019, pp. 5415–5421.

[10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[11] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *AAAI*, 2019, vol. 33, pp. 7370–7377.

[12] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *AAAI*, 2018, pp. 4970–4977.

[13] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," in *EMNLP-IJCNLP*, 2019, pp. 165–176.

[14] T. Mikolov, K. Chen, G. Corrado, and et al., "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[15] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[16] L. Guo, L. Wang, J. Dang, and et al., "A feature fusion method based on extreme learning machine for speech emotion recognition," in *ICASSP*. IEEE, 2018, pp. 2666–2670.

[17] R. Speer and et al, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *AAAI*, 2017, pp. 4444–4451.

[18] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *ACL*, 2018, pp. 174–184.

[19] P. Bojanowski, E. Grave, and et al., "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[20] D. Cer, Y. Yang, S. Kong, and et al., "Universal sentence encoder for english," in *EMNLP*, 2018, pp. 169–174.

[21] M. Schlichtkrull, T. N. Kipf, P. Bloem, and et al, "Modeling relational data with graph convolutional networks," in *European Semantic Web Conference*, 2018, pp. 593–607.

[22] C. E. Osgood, "The nature and measurement of meaning," *Psychological bulletin*, vol. 49, no. 3, pp. 197–237, 1952.