




Context- and Knowledge-Aware Graph Convolutional Network for Multimodal Emotion Recognition

Yahui Fu  and Shogo Okada , School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, 9231211, Japan

Longbiao Wang , Lili Guo , Yaodong Song, and Jiaying Liu , Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China

Jianwu Dang , Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China

This work proposes an approach for emotion recognition in conversation that leverages context modeling, knowledge enrichment, and multimodal (text and audio) learning based on a graph convolutional network (GCN). We first construct two distinctive graphs for modeling the contextual interaction and knowledge dynamic. We then introduce an affective lexicon into knowledge graph building to enrich the emotional polarity of each concept, that is the related knowledge of each token in an utterance. Then, we achieve a balance between the context and the affect-enriched knowledge by incorporating them into the new adjacency matrix construction of the GCN architecture, and teach them jointly with multiple modalities to effectively structure the semantics-sensitive and knowledge-sensitive contextual dependence of each conversation. Our model outperforms the state-of-the-art benchmarks by over 22.6% and 11% relative error reduction in terms of weighted-F1 on the IEMOCAP and MELD databases, respectively, demonstrating the superiority of our method in emotion recognition.

Emotion recognition in conversations (ERC) has attracted increasing attention because it is a necessary step for a number of applications, including social media threads (such as YouTube, Facebook, Twitter), human–computer interaction, and so on. Different from nonconversation cases, “context” is a vital component of ERC, which represents the previous dialog content of a target utterance. The intention and emotion of a target utterance are mostly affected by the surrounding context, as we can see from conversations in Figure 1. Therefore, it is important but challenging to effectively model the contextual dependence within conversations.

Recent studies mainly utilize recurrent neural networks or graph convolutional neural networks, such as the bc-LSTM,¹ DialogueRNN,² and DialogueGCN,³ to propagate contextual information. However, they only process the semantic information of the dialog. For implicit emotional utterances that do not contain clear emotional terms, and the words of which are relatively objective and neutral, such as utterance P_2U_2 in Figure 1, it is difficult to correctly distinguish the emotions when only considering the contextual semantics. Knowledge bases provide a rich source of background concepts, which can enhance understanding of a conversation. Both context modeling and commonsense knowledge are essential for a machine to analyze emotion in conversations.

Figure 1 shows two cases demonstrating the indispensability and superiority of knowledge and context modeling in ERC task, separately. We can see from conversation 1 that by incorporating an external knowledge base, the context of the term “National

1070-986X © 2022 IEEE

Digital Object Identifier 10.1109/MMUL.2022.3173430

Date of publication 10 May 2022; date of current version 23 September 2022.

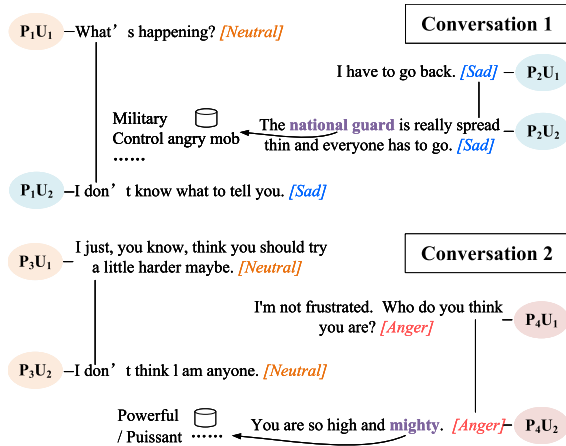


FIGURE 1. Two example conversations with annotated labels from the IEMOCAP dataset.⁴

Guard” in utterance P_2U_2 is enriched by associated concepts such as “Military” and “Control angry mob.” Thus, we can infer that the implicit emotion in utterance P_2U_2 should be “sad” via its enriched meaning.

In conversation 2, utterance P_4U_2 conveys the emotion of sarcasm. It is challenging to detect the emotion of such an utterance as the semantic meaning itself as well as the external knowledge “powerful, puissant” of “mighty” are positive, but we can infer the true emotion should be negative based on the context reasoning. Therefore, the exploration of the balance between context modeling and commonsense knowledge is of great importance to the task of ERC. However, to our best knowledge, it has not been well explored in the literature.

To further tackle the above problems, we propose a new conversational semantic- and knowledge-guided graph convolutional network (ConSK-GCN) with both text and audio modalities to effectively structure the context- and knowledge-sensitive dependence in each conversation. The contributions of this article can be summarized as follows:

- ▶ We propose a ConSK-GCN to leverage commonsense knowledge, affective lexicon, and context modeling by incorporating them into the new adjacency matrix construction of a new GCN architecture. And the enriched relational edges in the graph structure help in capturing the true emotion and intention of the interlocutors in the dialog.
- ▶ We validate our model on two benchmark databases, IEMOCAP and MELD, using text and audio modalities. The extensive experimental results demonstrate the effectiveness of our model in

comparison with state-of-the-art methods for the IEMOCAP and MELD databases, respectively.

- ▶ We demonstrate the efficiency and superiority of our method for emotion recognition in cases with both sufficient and insufficient context through learning with the IEMOCAP and MELD databases.

This study significantly expands upon⁵ by adding a second dataset MELD, which includes small group interactions by more than two interlocutors and demonstrates that the techniques improve classification performance for most of the variables in the MELD dataset. Furthermore, we continue to explore the balance between commonsense knowledge and context semantics in ERC, and we show the versatility of incorporating external knowledge by studying the two datasets.

RELATED WORK

Emotion Recognition With Multiple Modalities

People express emotions through multiple modalities, which include but are not limited to acoustic, textual, facial presentation. Previous studies like^{1,2,6-8,10} have demonstrated the effectiveness of multimodal learning for emotion recognition, which takes advantage of the complementary information of various modalities to enhance the understanding of emotions. In this article, we utilize both text and audio modalities for emotion recognition.

Emotion Recognition With Context

Previous research works used convolutional neural network (CNN)¹¹ and long short-term memory networks (LSTMs)¹⁰ for nonconversational feature extraction from diverse modality, while ignoring contexts. For the task of ERC, RNN-based methods such as bc-LSTM¹ and DialogueRNN² propagated contextual information to the utterances and process the constituent utterances of a dialog in sequence. DialogueGCN³ utilized a relational graph convolutional neural network⁹ to model the contextual dependence and achieved a new state of the art, proving the effectiveness of GCN in context structure. In this article, we further explore on the GCN architecture to encode the context and knowledge interaction of sentences.

Emotion Recognition With Knowledge Base

Commonsense knowledge bases help in grounding text to real entities, factual knowledge, and context-

specific commonsense concepts, and have attracted increasing attention in several research areas. For example,¹² augmented end-to-end dialog systems with commonsense knowledge.¹³ made use of a knowledge base by concatenating concept embedding and word embedding as the input to a transformer architecture, but ignored context modeling. In this article, we incorporate external knowledge into context construction.

PROPOSED METHOD

Definition and Notation

Context- and knowledge-aware multimodal emotion recognition: the goal is to predict the emotion label y of a conversation sample using the proposed ConSK-GCN method ($f(\cdot)$). The input to our approach are distinct modalities (audio μ_a , text μ_t), external knowledge concepts (C), and emotion polarities extracted from a lexicon (valence V , arousal A), that is,

$$y = f(\mu_a, \mu_t, C, V, A). \quad (1)$$

Multimodality

To better mine the information of the raw data and capture efficient contextual traits, we train separate networks to extract linguistic and acoustic features at the utterance level using emotion labels, and the details are described in the ‘‘Data Processing’’ section. Then, we concatenate the embeddings of these two modalities $\mu = [\mu_t; \mu_a]$, and utilize one bidirectional long short-term memory network (BLSTM) layer for sequence encoding to obtain the global contextual information, denoted as s .

Knowledge Retrieval

In this article, we utilize the ConceptNet commonsense knowledge base¹⁴ and the NRC_VAD emotion lexicon¹⁵ as the knowledge sources.

ConceptNet is a large-scale multilingual context-aware graph, which is designed to represent the general knowledge involved in understanding language. The nodes in ConceptNet are concepts and the edges represent relation. For example, as shown in Figure 2, ‘‘scholarship has synonym of bursary with confidence score of 0.741.’’ Therefore, we construct a knowledge graph based on the corresponding concepts extracted from the semantic dependence of each conversation.

The 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-value scores for VAD are on a scale of 0-1 for each dimension, respectively, corresponding to the degree from low to high.

Knowledge-Aware Graph Construction

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

Each node in the knowledge graph is embedded into an effective semantic space, named ‘‘ConceptNet Numberbatch,’’ that learns from both distributed semantics and ConceptNet.¹⁴ The tokens that are not included in ConceptNet are initialized by the ‘‘fast-Text’’¹⁶ method, which is a library for efficient learning of word representations. Formally, we denote the m th concept in the i th utterance as $c_{i,m}$ and the corresponding encoded embedding as $e_{i,m}$.

To obtain the edge weights, we first adopt l_2 norm to compute the emotion intensity $emo_{i,m}$ for each $c_{i,m}$, that is,

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

where $\|\cdot\|_2$ denotes the l_2 norm, M is a min-max normalization function, and $V(c_{i,m}) \in (0, 1)$, $A(c_{i,m}) \in (0, 1)$. In (2), we scale $V(c_{i,m})$ from -0.5 to 0.5 to denote the emotion polarity from negative to positive, and scale $A(c_{i,m})$ from 0 to 0.5 to describe the intensity of each emotion. For a concept not in the NRC_VAD, we set the value of $V(c_{i,m})$ and $A(c_{i,m})$ to 0.5 as a neutral score. Then, the affect-enriched knowledge embedding is represented as

$$k_{i,m} = emo_{i,m} e_{i,m}. \quad (3)$$

Furthermore, considering the past context window size of p and the future context window size of f , the edge weights in the knowledge-aware graph are defined as

$$\mathbf{a}_{i,j}^k = \sum_{m=1}^{n_i} \sum_{l=1}^{n_j} abs \left(\cos \left(\mathbf{k}_{i,m}^\top \mathbf{W}_k \left[\mathbf{k}_{j,1}, \dots, \mathbf{k}_{j,n_j} \right] \right) \right) \quad (4)$$

where n_i is the number of concepts in utterance i , and i ranges from 1 to n in each conversation, n_j is the number of concepts in utterance j , and $j = i - p, \dots, i + f$, and \mathbf{W}_k is a learnable parameters matrix.

Context-Aware Graph Construction

We build a context-aware graph $G_c = (V_c, E_c)$ based on the conversational semantic dependence, where V_c denotes a set of utterance nodes, and $E_c \subset V_c \times V_c$ is a set of relation/edge weights that represent the context-sensitive semantic similarity among the utterances.

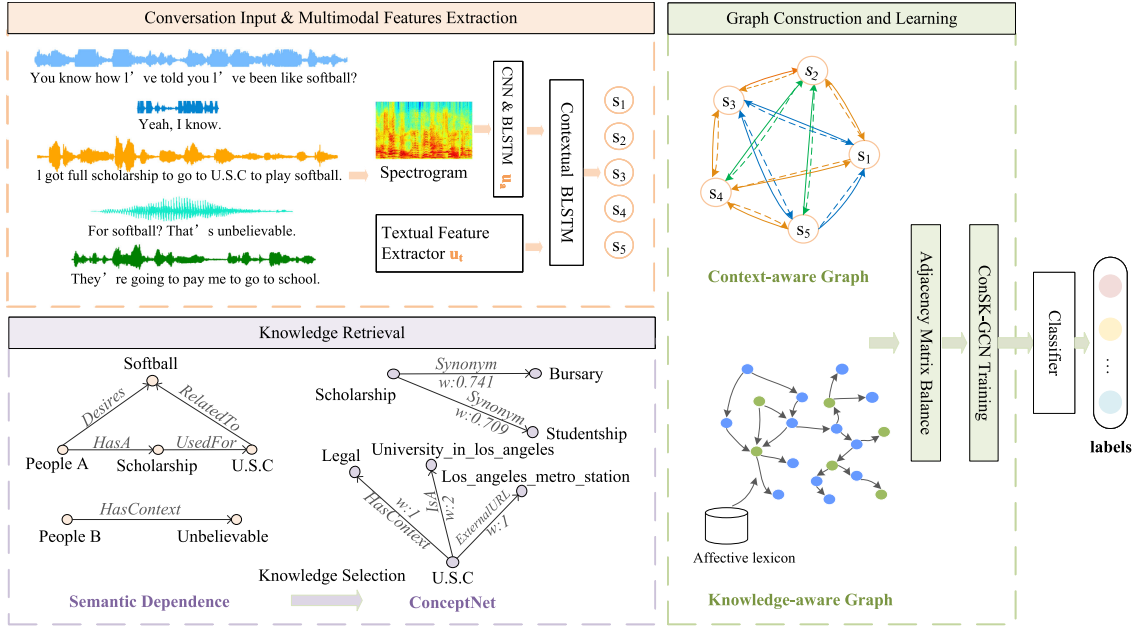


FIGURE 2. Overall architecture of our proposed ConSK-GCN approach for multimodal emotion recognition.

The node features in the context-aware graph are the multimodal representation s . To obtain the edge weights of the context-aware graph, we first compute the cosine similarity of two utterances, and then take the arccos to convert the cosine similarity into an angular distance, that is,

$$sim_{i,j} = 1 - \arccos\left(\frac{s_i^\top s_j}{\|s_i\| \|s_j\|}\right) / \pi. \quad (5)$$

Then, the edge weights in the context-aware graph is formulated as

$$a_{i,j}^c = \text{softmax}(\mathbf{W}_c [sim_{i,i-p}, \dots, sim_{i,i+f}]) \quad (6)$$

where s_j denote the multimodal representation of the i th and j th utterances in the same conversation, respectively, and \mathbf{W}_c is a trainable parameter matrix.

ConSK-GCN Training

We build our context- and knowledge-aware graph as $G_{ck} = (V_c, E_{ck})$. To model both context-sensitive and knowledge-dependence between utterances in each conversation, we leverage the addition of knowledge edge weights $a_{i,j}^k$ and contextual edge weights $a_{i,j}^c$ as the new edge weights $a_{i,j}$ to build the adjacency matrix \mathbf{E}_{ck} of ConSK-GCN, that is,

$$a_{i,j} = \omega_k a_{i,j}^c + (1 - \omega_k) a_{i,j}^k \quad (7)$$

where ω_k is a model parameter balancing the impacts of knowledge and contextual semantics on computing

the relational dependence in each conversation. Then, we feed the global contextual multimodal representations s and edge weights $a_{i,j}$ into a two-layer GCN⁹ to capture local contextual information that is both context-aware and knowledge-sensitive

$$\begin{aligned} \mathbf{h}_i^{(1)} &= \sigma \left(\sum_{r \in \mathfrak{R}} \sum_{j \in N_i^r} \frac{a_{i,j}}{q_{i,r}} \mathbf{W}_r^{(1)} s_j + \mathbf{a}_{i,i} \mathbf{W}_0^{(1)} s_i \right) \\ \mathbf{h}_i^{(2)} &= \sigma \left(\sum_{j \in N_i^r} \mathbf{W}^{(2)} \mathbf{h}_j^{(1)} + \mathbf{W}_0^{(2)} \mathbf{h}_i^{(1)} \right) \end{aligned} \quad (8)$$

where N_i^r denotes the neighboring indices of each node under relation $r \in \mathfrak{R}$, \mathfrak{R} 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, $\mathbf{W}_r^{(1)}$, $\mathbf{W}_0^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{W}_0^{(2)}$ are model parameters, and $\sigma(\cdot)$ is an activation function such as ReLU.

EXPERIMENTS

Databases

We evaluate our ConSK-GCN on two conversational databases, namely the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) dataset⁴ and the *Multimodal EmotionLines Dataset* (MELD).¹⁷ In our work, we focus on multimodal emotion recognition with text and audio modalities.

The IEMOCAP database contains videos of ten unique speakers acting in two different scenarios:

scripted and improvised dialog with dyadic interactions. We use 5531 utterances in 151 dialogs with four emotion categories: 29.6% happiness, 30.9% neutral, 19.9% anger, and 19.6% sadness. In this article, 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.

The MELD is developed by crawling the dialog from each episode in the popular sitcom *Friends*, where each dialog contains utterances from multiple speakers. There are around 13,708 utterances from 1433 dialogs with seven emotion categories: 46.95% neutral, 16.84% joy, 11.72% anger, 11.93% surprise, 7.31% sadness, 2.63% disgust, and 2.61% fear.

The average conversation length and average utterance length are 49.2 and 15.8, respectively, in the IEMOCAP database with two participants, and 9.6 and 8.0 in the MELD with many conversations having more than 5 participants,¹⁷ which means the majority of participants utter only a small number of utterances per conversation. Therefore, compared with IEMOCAP, the utterances in the MELD are shorter, and their context-dependence is weaker. We would like to explore the efficiency of our method in both cases with sufficient and insufficient context.

Data Processing

Textual Features

We employ different approaches to extract utterance-level linguistic features from IEMOCAP and MELD datasets based on the particular traits of these two datasets.

IEMOCAP: To compare with the state-of-the-art approaches, we employ a convolutional neural network (CNN)¹¹ to extract the textual embeddings of the transcripts. We use the publicly available word2vec to initialize the word vectors. And the number of convolutional filters is set to 3, 4, and 5 with 100 feature maps in each. These are then concatenated and fed into two fully-connected layers with 500 and 100 hidden nodes, followed by a ReLU activation function.

MELD: Considering the weaker context-dependence in the MELD, we apply BERT-Base, which has a multilayer bidirectional transformer encoder model architecture, to initialize the textual representations of the MELD. We take the representations of both training and test samples from the penultimate dense layer as the context independent utterance level feature vectors by fine-tuning on a trained BERT-Base.

Acoustic Features

In this article, we follow the audio preprocessing method introduced by Guo *et al.*¹⁸ We apply a 265 ms

window size with a 25 ms slide window to cut an utterance into several segments. Each speech segment is transformed to a spectrogram using a short time Fourier Transform (STFT), then each input spectrogram has the following $time \times frequency$: 32×129 .

For the MELD, we set the time of each segment to 2 s and the slide window to 1 s, then the size of each spectrogram is 1874×129 . We determined that a larger window size has a better performance, which we attribute to the fact that there is a lot of noise in sitcom dialog, and if utterances are cut using a small window size, many fragments will just be noise.

A CNN is utilized to extract deep acoustic features from the segment-level spectrograms. Then, the segment-level features are propagated into a 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.

Experimental Settings

For the IEMOCAP dataset, we use a batch size of 32 and train it for 100 epochs, whereas for the MELD, we change the number of epochs to 20. In the IEMOCAP dataset, 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. In the MELD, the window sizes of the past and future contexts are all set to 6. The learning rate is set to 0.0001 for both unimodality and multimodality training and ω_k is set to 0.6 for the MELD database to balance the effect of context and knowledge. In our conference version,⁵ w_k is set to 0.5 for the IEMOCAP database, which means equal balance. In this article, to determine an optimal balance, we further make a detailed analysis about the effect of w_k in the section “Effect of the Balancing Weight.” Based on the comparative results in Figure 4, we choose w_k to 0.3 in this article.

Method Comparison

For comprehensive evaluation, we compare our method with state-of-the-art (SOTA) approaches as well as ablation studies.

CNN¹¹: A widely used architecture for both text and audio feature extraction with effective performance. We employ it to extract utterance-level textual and acoustic features; it does not contain contextual information.

BERT-Base¹⁹: A bidirectional encoder with 12-layer transformer blocks, which obtains advanced results

Utterance	Gold_Label	ConS-GCN ($w_k=1$)	ConK-GCN ($w_k=0$)	ConSK-GCN ($w_k=0.5$)	Knowledge in ConcepNet
1 I got into college .	H	H ✓	H ✓	H ✓	Gather good minds/ Find oneself 😊
2 But if that can not happen, I will just have to get out .	A	N ✗	A ✓	A ✓	Escape/ Difficulty 😞
3 Being dishonest with him. It is the kind of thing that pays off.	S	N ✗	S ✓	S ✓	Hurt someone else/ Deceitful 😏
4 Have a good day.	N	N ✓	H ✗	N ✓	Favorable/ Satisfactory 😊

FIGURE 3. Visualization of several representative examples in the IEMOCAP dataset corpus. Blue denotes the typical token in each utterance.

for sentence-level sentiment analysis. This approach does not aim to deal with context.

LSTMs¹⁰: A framework for unimodality and multi-modality emotion recognition based on audio and text, without exploring context information.

bc-LSTM¹: A bidirectional LSTM network takes the sequence of utterances in a video as input and extracts contextual unimodal and multimodal features by modeling the dependencies among the input utterances.

DialogueRNN²: Three GRUs to model the dynamics of the speaker states, the context from the preceding utterances and the emotion of the preceding utterances. This method achieved SOTA results in multimodal emotion recognition in conversations.

DialogueGCN³: Leverages self and interspeaker dependence of the interlocutors to model the conversational context for textual emotion recognition. This method only considers context modeling.

ConS-GCN: Considers the semantics-sensitive dependence between utterances in the adjacency matrix construction of the GCN based on the context-aware graph with w_k set to 1 in (7).

ConK-GCN: Considers the knowledge-sensitive dependence between utterances in the adjacency matrix construction of the GCN based on the knowledge-aware graph with w_k set to 0 in (7).

ConSK-GCN: Considers both semantics- and knowledge-sensitive contextual dependence between utterances in the adjacency matrix construction of the GCN with w_k set to 0.3 for the IEMOCAP database and 0.6 for the MELD in (7).

Experimental Results on the IEMOCAP Database

Comparison With SOTA Methods

Table 1 shows the performance of comparative experiments with SOTA methods. From this table, we observe that methods that consider the context are much more effective than methods without context,

demonstrating the significance of context modeling. Encouragingly, the comparison shows that our proposed “ConSK-GCN” performs better than all of the baseline approaches, with a relative error reduction of over 22.6% in terms of both weighted-accuracy and weighted-F1. The results indicate the effectiveness of our model that incorporates external knowledge and contextual semantics for emotion detection in conversations.

Ablation Studies

In order to further demonstrate the contribution of knowledge and contextual semantics to emotion recognition, we design ablation studies as shown in Table 2.

We first discuss the efficiency of the proposed “ConSK-GCN” with single text modality. Compared to “ConS-GCN” and “ConK-GCN,” “ConSK-GCN” has an improvement of at least “1.7%,” “4.3%,” and “0.4%” relative error reduction in terms of the weighted-F1 for detecting *neutrality*, *anger*, and *happiness*. Specifically, the improvement in *happiness* detection is not

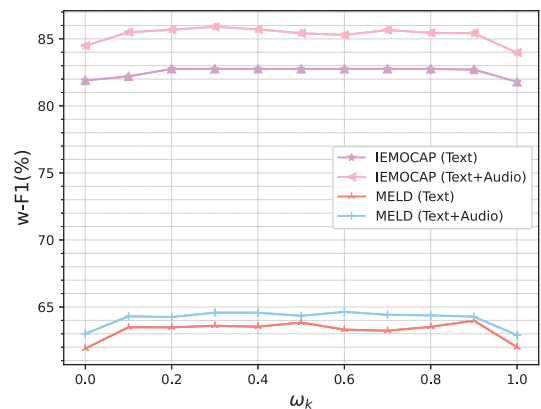


FIGURE 4. Effect of the balancing weight (ω_k) for emotion recognition with various modalities in the IEMOCAP and MELD databases.

TABLE 1. Comparative experiments with SOTA methods. acc.= accuracy; average (w)= weighted average.

Models	Modality	Neutrality		Anger		Happiness		Sadness		Average(w)	
		Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)
CNN [11]	Text	59.11	59.50	77.06	65.17	64.03	69.36	62.04	60.68	63.90	64.02
BERT-Base [19]	Text	75.00	78.26	75.29	68.09	60.41	60.82	70.20	69.92	68.90	69.01
DialogueGCN [3]	Text	74.22	74.32	77.06	76.61	87.56	88.66	85.31	83.60	81.57	81.55
LSTMs [10]	Text	72.92	74.97	70.00	65.93	55.20	56.42	63.67	61.30	64.38	64.42
	Text+Audio	69.53	63.95	73.53	73.10	66.74	73.75	70.61	67.98	69.30	69.50
bc-LSTM [1]	Text	76.04	67.51	75.88	72.88	67.65	75.51	67.35	70.06	71.31	71.60
	Text+Audio	79.95	70.49	78.82	77.91	70.14	78.58	73.88	75.73	75.10	75.42
DialogueRNN [2]	Text	81.51	73.73	66.47	74.10	86.43	87.82	72.24	77.29	79.37	79.50
	Text+Audio	86.20	76.53	84.71	83.72	79.64	86.38	75.10	80.35	81.47	81.78
Ours	Text	74.48	75.66	80.00	78.84	87.78	88.79	89.39	86.39	82.92	82.89
	Text+Audio	82.03	80.67	87.65	84.18	88.91	91.18	84.90	85.77	85.82	85.90

Bold font denotes the best performances.

significant. According to the emotion theory introduced by Osgood,²⁰ the Valence-Arousal space depicts the affective meanings of linguistic concepts. We believe that *happiness* is an emotion with explicit linguistic features with positive valence and positive arousal, which is also contagious to the context, therefore, the efficiency of external knowledge and context modeling is similar. By contrast, *sadness* is relatively implicit in linguistic characteristics with negative valence and negative arousal. Therefore, we can see that the improvement for *sadness* is more significant (“ConSK-GCN” has an increase of 14.6% relative error reduction in terms of weighted-F1 compared with “ConK-GCN” which is 18.7% compared to “ConS-GCN”). These illustrate the effectiveness of incorporating knowledge with semantics for contextual feature extraction in conversation, particularly for implicit emotional utterances.

Compared with single text modal, the detection performance in *neutrality*, *anger*, and *happiness* improved by 20.6%, 25.2%, and 21.3% relative error reduction in terms of weighted-F1, respectively, using the proposed “ConSK-GCN” with both text and audio modalities, which demonstrates the importance of integrating complementary acoustic and linguistic features into emotion recognition. However, there is

an exception in *sadness* detection that we assume because, similar to text features, the acoustic characteristics of *sadness* are also implicit, which decrease the overall performance.

Experimental Results on the MELD

Table 3 further quantifies the efficiency of our model on the MELD. For the text modality, “DialogueGCN” performs best among all of the baselines with a 58.10% weighted-F1. And our proposed “ConSK-GCN” shows the best performance with a 13.7% relative error reduction compared to “DialogueGCN.” For the text and audio modalities, the improvement over the SOTA “DialogueRNN” is an 11% relative error reduction, both of which results highlight the importance of integrating semantic-sensitive and knowledge-sensitive contextual information into emotion recognition.

Furthermore, compared to the IEMOCAP dataset, the context-dependence is weaker in the MELD, as discussed in the “Databases” section. For text modality, compared with the ablation studies, the proposed “ConSK-GCN” has an improvement of over 4.9%, 5.1%, 5.7%, 4.5%, and 1.8% relative error reduction in terms of the weighted-F1 for *neutrality*, *anger*, *joy*, *surprise*, and *sadness* detection, respectively,

TABLE 2. Ablation studies of the proposed method.

Models	Modality	Neutrality		Anger		Happiness		Sadness		Average(w)	
		Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)	Acc.(%)	F1(%)
ConS-GCN	Text	76.04	74.68	77.65	77.65	87.33	88.74	83.27	83.27	81.71	81.79
	Text+Audio	78.91	77.79	85.29	83.57	90.72	90.52	78.78	82.13	83.96	83.97
ConK-GCN	Text	75.52	75.23	77.65	77.88	86.65	88.05	86.12	84.06	81.87	81.90
	Text+Audio	75.78	77.70	88.82	85.31	89.37	90.08	86.53	84.46	84.53	84.49
ConSK-GCN ($w_k=0.5$)	Text	74.48	75.66	80.00	78.84	87.78	88.79	89.39	86.39	82.92	82.89
	Text+Audio	78.13	79.89	87.06	84.33	93.67	91.90	82.86	84.76	85.82	85.74
ConSK-GCN ($w_k=0.3$)	Text	74.48	75.66	80.00	78.84	87.78	88.79	89.39	86.39	82.92	82.89
	Text+Audio	82.03	80.67	87.65	84.18	88.91	91.18	84.90	85.77	85.82	85.90

TABLE 3. Comparative experiments of different methods for unimodality (text) and multimodality (text+audio) emotion recognition. F1-Score (%) is used as the evaluation metric. acc.=accuracy, w= weighted average.

	Models	Modality	Neutrality	Anger	Disgust	Joy	Surprise	Sadness	Fear	w-F1	w-Acc.
Baselines	CNN [11]	Text	67.30	12.20	0.0	32.61	45.13	19.63	0.0	45.45	52.15
	BERT-Base [19]	Text	77.96	43.97	0.0	53.52	52.50	0.0	0.0	57.22	62.99
	DialogueGCN [3]	Text	-	-	-	-	-	-	-	58.10	-
	LSTMs [10]	Text	67.59	12.32	0.0	35.99	45.68	17.19	0.0	45.99	52.91
		Text+Audio	69.00	20.37	0.0	36.60	44.82	8.98	0.0	47.08	53.56
	bc-LSTM [1]	Text	71.63	42.06	21.69	54.31	48.15	26.92	7.75	56.44	-
		Text+Audio	76.67	43.39	23.66	54.48	51.04	24.34	9.38	59.25	-
	DialogueRNN [2]	Text	75.75	40.59	2.04	50.27	49.38	24.19	8.93	57.03	-
		Text+Audio	77.44	43.65	7.89	54.40	52.51	34.59	11.68	60.25	-
	Ablation Studies	ConS-GCN	Text	76.96	50.28	2.86	58.80	59.13	35.75	0.0	62.03
Text+Audio			77.70	52.16	0.00	60.40	58.86	36.95	0.00	62.87	64.29
ConK-GCN		Text	76.96	51.64	0.00	56.34	58.13	35.05	13.70	61.85	63.10
		Text+Audio	77.53	52.59	0.00	60.94	61.99	33.33	0.0	62.98	64.18
Proposed	ConSK-GCN	Text	78.08	54.10	0.00	61.13	60.95	36.89	10.53	63.84	64.90
	($w_k=0.6$)	Text+Audio	79.01	53.52	0.00	63.54	61.43	39.23	0.00	64.63	66.40

the improvement of which is more significant than for the IEMOCAP dataset. This illustrates the superiority and potentiality of our model to emotion detection with insufficient context. In addition, the data ratio of "Disgust" only accounts for 2.63% in the MELD, while the percent of "Fear" is around 2.61%. It is difficult to accurately distinguish genuine emotions for such a small amount of data, which, depending on specific emotional characteristics, is left as future work.

However, the effectiveness of multimodal fusion for the MELD is more limited than for the IEMOCAP database. We think this is because many conversations on the *Friends* TV series include more than 5 participants so there are multiple audio sources, not only the speaker's voice, when detecting the acoustic signal of the target speaker. Therefore, the denoising process should be adopted to extract clean acoustic features, which is left for future work.

EFFECT OF THE BALANCING WEIGHT

In order to determine an optimal balance between knowledge and contextual semantics in our ConSK-GCN learning, we verify the influence of w_k based on single (Text) and multimodal (Text+Audio) on the IEMOCAP and MELD databases. We can conclude from Figure 4 that leveraging knowledge-aware and semantic-aware contextual construction together contributes significantly to emotion recognition in conversations, as the F1-score when leveraging knowledge and semantics together (w_k ranges from 0.1 to 0.9) increased more than only one or the other (w_k equal to 0 or 1). However, the effect of various values of w_k (0.1 to 0.9) on emotion detection is not conspicuous, as

they give similar results when the difference does not exceed 1%. Therefore, we choose a relatively optimal value of w_k to be 0.3 and 0.6 for the IEMOCAP and MELD databases, respectively, to balance the effect of knowledge and contextual semantics.

CASE STUDY

To verify the effectiveness of external knowledge and context modeling for emotion recognition, we visualize several typical samples, as shown in Figure 3.

Case 1 demonstrates the efficiency of both context modeling and commonsense knowledge alone for emotion detection in a conversation. Cases 2 and 3 illustrate the superiority of commonsense knowledge. In detail, compared to "ConS-GCN," which only considers the contextual semantics, our proposed "ConK-GCN" and "ConSK-GCN" that take advantage of external knowledge, can effectively capture implicit emotional characteristics. Case 4 shows that sometimes the emotional polarity of knowledge concepts are rather misleading in neutral expressions but incorporating them with context can alleviate this phenomenon.

CONCLUSION

In this work, we proposed a new conversational semantic- and knowledge-guided graph convolutional network (ConSK-GCN) for multimodal emotion recognition. In our approach, we teach context-aware and knowledge-aware graphs jointly using a GCN based on multimodal representations to help in capturing the true emotion and intention of the interlocutors in a conversation. Comparative experiments on the IEMOCAP and MELD databases show that our approach significantly outperforms the SOTA, illustrating the complementary effect of context modeling and commonsense knowledge of

the conversational emotion recognition. In addition, we plan to further explore the emotion cue of interlocutors and multimodal fusion strategies for more accurate emotion recognition in conversation.

REFERENCES

1. S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883, doi: [10.18653/v1/P17-1081](https://doi.org/10.18653/v1/P17-1081).
2. N. Majumder et al., "Dialoguernn: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 6818–6825, doi: [10.1609/aaai.v33i01.33016818](https://doi.org/10.1609/aaai.v33i01.33016818).
3. D. Ghosal, N. Majumder, S. Poria, and A. Gelbukh., "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 154–164, doi: [10.18653/v1/D19-1015](https://doi.org/10.18653/v1/D19-1015).
4. C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008, doi: [10.1007/s10579-008-9076-6](https://doi.org/10.1007/s10579-008-9076-6).
5. Y. Fu et al., "Consk-GCN: Conversational semantic and knowledge-oriented graph convolutional network for multimodal emotion recognition," in *Proc. Int. Conf. Multimedia Expo.*, 2021, pp. 1–6, doi: [10.1109/ICME51207.2021.9428438](https://doi.org/10.1109/ICME51207.2021.9428438).
6. P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017, doi: [10.1109/JSTSP.2017.2764438](https://doi.org/10.1109/JSTSP.2017.2764438).
7. N. Li, B. Liu, Z. Han, Y.-S. Liu, and J. Fu, "Emotion reinforced visual storytelling," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 297–305, doi: [10.1145/3323873.3325050](https://doi.org/10.1145/3323873.3325050).
8. T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14234–14243, doi: [10.1109/CVPR42600.2020.01424](https://doi.org/10.1109/CVPR42600.2020.01424).
9. M. Schlichtkrull et al., "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.*, 2018, pp. 593–607, doi: [10.1007/978-3-319-93417-4_38](https://doi.org/10.1007/978-3-319-93417-4_38).
10. S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," 2018, *arXiv:1804.05788*.
11. Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.
12. T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 4970–4977, 2018.
13. P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 165–176, doi: [10.18653/v1/D19-1016](https://doi.org/10.18653/v1/D19-1016).
14. R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4444–4451.
15. S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 174–184, doi: [10.18653/v1/P18-1017](https://doi.org/10.18653/v1/P18-1017).
16. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017, doi: [10.1162/tacl_a_00051](https://doi.org/10.1162/tacl_a_00051).
17. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536, doi: [10.18653/v1/P19-1050](https://doi.org/10.18653/v1/P19-1050).
18. L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2666–2670, doi: [10.1109/ICASSP.2018.8462219](https://doi.org/10.1109/ICASSP.2018.8462219).
19. J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2018, pp. 4171–4186.
20. C. E. Osgood, "The nature and measurement of meaning," *Psychol. Bull.*, vol. 49, no. 3, pp. 197–237, 1952, doi: [10.1037/h0055737](https://doi.org/10.1037/h0055737).

YAHUI FU is currently working toward the Ph.D. degree with the Department of Intelligence Science and Technology, Kyoto University, Kyoto, 9231211, Japan, and also with Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China. She received the M.S. degree from both Tianjin University, Tianjin, China, and the Japan Advanced Institute of Science and Technology, Ishikawa, Japan. Her research interests include multimodal emotion recognition and spoken dialogue systems. Contact her at fu.yahui.64p@st.kyoto-u.ac.jp.

SHOGO OKADA is an associate professor with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, 9231211, Japan. He received the Ph.D. degree from the Tokyo Institute of Technology. His research interests include multimodal interaction modeling, human dynamics analysis based on machine learning, pattern recognition, and data mining techniques. Contact him at okada-s@jaist.ac.jp.

LONGBIAO WANG is a professor with the College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China. He received the Ph.D. degree from the Toyohashi University of Technology, Toyohashi, Japan. His research interests include robust speech recognition, speaker recognition, and acoustic signal processing. He is a member of the IEEE. Contact him at longbiao_wang@tju.edu.cn.

LILI GUO is a lecturer with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221008, China. She received the Ph.D. degree from Tianjin University, Tianjin, China. Her research interests are in the fields of emotion recognition, deep learning, and acoustic signal processing. Contact her at liligu@cumt.edu.cn.

YAODONG SONG is currently working toward the M.S. degree with the College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China. He received the B.S. degree from the North China Institute of Aerospace Engineering, Langfang, China. His research interests include emotion recognition and deep learning. Contact him at songyaodong@tju.edu.cn.

JIAXING LIU is currently working toward the Ph.D. degree with the College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China. He received the M.S. degree from the Tianjin University of Technology, Tianjin, China. His research interests include speech emotion recognition and multimodal emotion recognition. Contact him at jiaxingliu@tju.edu.cn.

JIANWU DANG is a professor with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, 9231211, Japan, and also with the Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China. He received the Ph.D. degree from Shizuoka University, Japan. His research interests include speech production, speech recognition, and spoken language understanding. Contact him at jdang@jaist.ac.jp.

Computing in Science & Engineering

The computational and data-centric problems faced by scientists and engineers transcend disciplines. There is a need to share knowledge of algorithms, software, and architectures, and to transmit lessons-learned to a broad scientific audience. *Computing in Science & Engineering (CISE)* is a cross-disciplinary, international publication that meets this need by presenting contributions of high interest and educational value from a variety of fields, including physics, biology, chemistry, and astronomy. *CISE* emphasizes innovative applications in cutting-edge techniques. *CISE* publishes peer-reviewed research articles, as well as departments spanning news and analyses, topical reviews, tutorials, case studies, and more.

Read *CISE* today! www.computer.org/cise



IEEE
COMPUTER
SOCIETY

