Master's Thesis

Conversational Semantic- and Knowledge-guided Graph Convolutional Network for Multimodal Emotion Recognition

Fu Yahui

Supervisor Okada Shogo

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

June, 2021

Abstract

Emotion recognition in conversation (ERC) has received significant attention in recent years and become a new frontier of natural language processing research due to its widespread applications in diverse areas, such as social media, health care, education, and artificial intelligence interactions. Therefore, the effective and scalable conversational emotion recognition algorithms are of great significance.

It is challenging to enable machines to understand emotions in conversations, as humans often rely on the contextual interaction and commonsense knowledge to express emotion. Therefore, both context and incorporating external commonsense knowledge are essential for the task of ERC. Graph convolutional neural network (GCN) technologies have been widely applied in the contextual information extraction due to its ability in learning complex structures. Most studies which utilize GCN only consider the syntactic 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. Commonsense knowledge is the basis for understanding contextual dialogues and generating appropriate responses in human-robot interaction, however, it has not well been explored for ERC. Previous studies either focused on extracting features from a single sentence and ignored contextual semantics, or only considered semantic information when constructing the graph,ignoring the relatedness between the tokens. We hypothesize that both semantic contexts and commonsense knowledge are essential for machine to analyze emotion in conversations.

To further tackle the above problems, we propose a new multimodal Semantic- and Knowledge-guided Graph Convolutional Network (ConSK-GCN) to effectively structure the semantic-sensitive and knowledge-sensitive contextual dependence in each conversation. On one hand, we construct models capturing the contextual interaction and intradependence of the interlocutors via a conversational semantic-guided 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 and generate appropriate responses to enrich the semantic meaning of the tokens in the utterance via a conversational knowledge-guided GCN (ConK-GCN).

Furthermore, we introduce an affective lexicon into knowledge graph construction to enrich the emotional polarity of each concept. 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. In addition, we focus on multimodal emotion recognition using the acoustic and textual representations, because both text and prosody convey emotions when communicating in conversations.

Experimentation on the multimodal corpus IEMOCAP and MELD show that our methodology could effectively utilize the contextual dependence of the utterances in a conversation for emotion recognition. Moreover, detailed evaluation indicates that our approach is superior than several state-of-the-art baselines in both uni-modality and multi-modality emotion recognition.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research background

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)[3], convolutional neural network (CNN)[4], and recurrent neural network (RNN)[5] are the most commonly used architectures for emotion recognition and usually achieve competitive results.

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 non-conversation 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 [6] 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 [7]. To mitigate this issue, some variants like AIM [8] and DialogueRNN [9] integrate with an attention mechanism that can dynamically focus on the most relevant contexts. However, this attention mechanism does not consider the relative position of

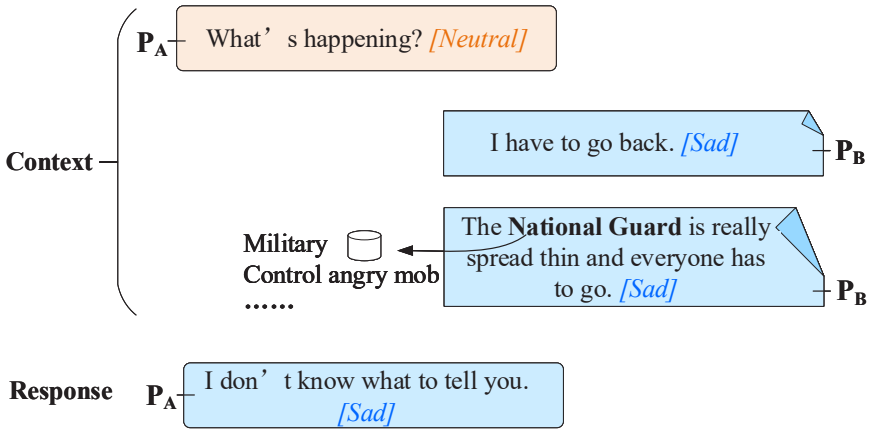


Figure 1.1: An example conversation with annotated labels from the IEMOCAP dataset.

the target and context utterances, which is important for modeling how past utterances influence future utterances and vice versa. DialogueGCN [10] and ConGCN [11] 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. As the emotion of the target utterance is usually strongly affected by the nearby utterances and relational edges in the graph would help in capturing the inter-dependence and intra-dependence among the speakers in play. 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. We can see from figure 1 that, 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". Therefore, the implicit emotion in the third utterance can be inferred more easily via its enriched meaning. However, in the literature, only a limited number of studies have explored the incorporation of context and commonsense knowledge via GCN

for the ERC task.

1.2 Problem statement

Current research considers utterances as independent entities alone, but ignores the inter-dependence and relations among the utterances in a dialogue. However, contextual dependence is significant for sentiment analysis. Conversational emotion analysis utilizes the relation among utterances to track the user's emotion states during conversation, it is important but challenging to effectively model the interaction of different speakers in the conversational dialogue.

Previous studies either use LSTM-based methods for sequential encoding or apply GCN-based architecture to extract neighborhood contextual information. LSTM-based methods have the issue of sequence propagation, which may not perform good on long-term context extraction. To address the long-term propagation issue, some state-of-the-arts adopt neighborhood-based graph convolutional networks to model conversational context for emotion detection and have a good performance, due to the relational edges modeling, which represents the relevance between the utterances. However, for the utterances that the emotional polarity of which are difficult to distinguish, it is difficult to correctly detect its emotion if only take the semantics of the utterance into account.

1.3 Research motivation

In the task of emotion recognition in conversation (ERC), both intra-dependence and inter-dependence of the interlocutors are significant to model the dynamic interaction and capture the emotion changes in each turn. Graph neural networks have been shown effective performance at several tasks due to their rich relational structure and can preserve global structure information of a graph in graph embeddings. The neighborhood-based structure of GCN is a suitable architecture to extract the contextual information of the interaction of both inter-speaker and intra-speaker.

The information conveyed by the semantics of the context are not enough for emotion detection, especially for the small-scale database and implicit emotional texts. 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.

1.4 Research objective

The objective of this thesis is to propose a new multimodal Semantic- and Knowledge-guided Graph Convolutional Network (ConSK-GCN) to effectively structure the semantic-sensitive and knowledge-sensitive contextual dependence in each conversation.

To further tackle the above problems, on the one hand, we construct the contextual inter-interaction and intradependence of the interlocutors via a conversational semantic-guided 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-guided GCN (ConK-GCN). Furthermore, we introduce an affective lexicon into knowledge graph construction to enrich the emotional polarity of each concept. To the end, 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.

1.5 Thesis organization

The organization of this thesis is generalized as belows:

Chapter 1:

We introduces the background of emotion recognition in conversations and illustrate the significance of efficient context construction as well as the necessity of incorporating external knowledge base to enrich the semantic meaning of each concept in dialogues. Then we elaborated on the existing problems in current research and put forward our motivation based on these problems. And also the objective of this thesis.

Chapter 2:

We first introduce related works based on single and multi-modalities for the task of emotion recognition. Then we describe the important factors in the task of emotion detection in conversation. Then the state-of-the-art approaches of incorporating knowledge base and graph convolutional neural network in the conversational emotion analysis are described to show the

effectiveness of these two methods.

Chapter 3:

We first introduce the corpus used in this study, then describe the preparation of the database and the extraction of multimodal features (text and audio, and multimodality fusion), further illustrate our proposed conversational semantic- and knowledge-guided graph convolutional network (ConSK-GCN) methodology in detail.

Chapter 4:

We describe the experimental setup in this thesis, then makes the comparisons with the state of the arts as well as the ablation studies based on both single modality and multi-modality on two different database. In this chapter, we make detailed analysis about the performance of our method and list several case studies to further illustrate the effectiveness of our proposed model.

Chapter 5:

In this part, we eventually make a conclusion about the contributions of this work and then give an outlook on future work.

Chapter 2

Literature Review

Emotion is inherent to humans and with the development of human-robot interaction, emotion understanding is a key part of human-like artificial intelligence. The primary objective of an emotion recognition system is to interpret the input signals from different modalities, and use them to analyze the emotion intention of the users in the conversation or social network. As shown in figure 2.1, which is one of the typical emotion recognition framework, the extracted and processed features of the selected modalities are used to determine emotions by applying appropriate classification or regression methods. Meanwhile, external knowledge, such as personality, age, gender and knowledge base are usually applied to enrich the meaning of each modality. Then the final decision is made by fusing different results.

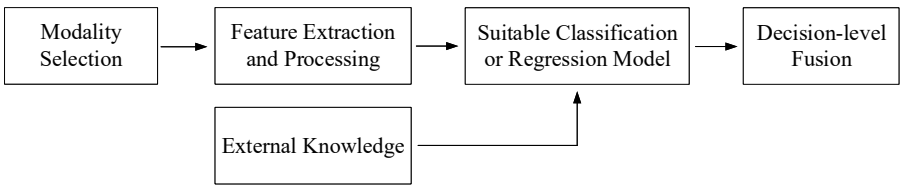


Figure 2.1: Typical emotion recognition framework.

2.1 Multimodal emotion recognition

In the literature, there are plenty of efforts focusing on different single modality or multi-modalities for emotion analysis, such as, physiological signals, facial expression, acoustic and textual features. In this section, we mainly introduce related works based on speech or text modality for the task of emotion recognition.

Table 2.1: Auditory-based feature set

LLDs (16×2)	MFCC(1-12): Mel Frequency Cepstral Coefficient, RMS Energy(1): root mean square frame energy, F0(1): fundamental frequency, ZCR(1): zero-crossing-rate from the time signal, HNR(1): harmonics-to-noise ratio by autocorrelation function
Functionals(12)	Max, min, mean, range, standard deviation, kurtosis, skewness, offset, slope, MSE, absolute position of min/max

2.1.1 Acoustic modality

Verbal communication aids in recognizing the emotional state of the communicating person effectively, as speech is one of the most natural ways to express ourselves and to grasp emotion and content of interlocutors. Speech emotion recognition (SER) has been around for more than two decades[12] and it has applications in many applications, such as human-computer interaction[13], robots[14], psychological assessment[15] and so on. However, SER is still a challenging task. One of the difficulties is how to extract effective acoustic features. There are two kinds of most used acoustic features in SER: (1) auditory-based features, such as Mel Frequency Cepstral Coefficient (MFCC), F0, zero-crossing-rate (ZCR), energy; (2) spectrogram-based deep acoustic features.

The auditory-based features are selected based on human auditory perception, which can be extracted by the openSMILE [16] tool with 384 dimensions proposed in [17]. The selected 16 low-level descriptors (LLDs) and their first-order derivatives are the basic features, and then 12 functionals are applied to these basic features, as shown in table 2.1.

There exits several problems in extracting auditory-based features manually, such as it's time-consuming and producing a limited number of feature categories [18]. With the development of deep learning, there is a trend in the field of speech processing to use Convolutional neural networks (CNNs) directly on spectrograms to extract deep acoustic features [19], and then applied the Bidirectional Long Short-Term Memory (BLSTM) to recognize emotions. The CNN-BLSTM model [18, 20] has been widely adopted for SER at present and has shown good performance.

2.1.2 Text modality

Text emotion recognition has emerged as a prevalent research topic that can make some valuable contributions in social media applications like Facebook,

Twitter and Youtube. It is significant to extract effective textual features for emotion recognition but still a challenging task.

In the traditional studies, distributed representations or pre-trained embeddings are playing important roles in state-of-the-art sentiment analysis systems. For example, predictive methods Word2Vec [21] and Glove [22], which can capture multi-dimensional word semantics. Beyond word-semantics, there has been a big efforts toward End-to-End neural network models [23] and achieved better performance by fine-tuning the well pre-trained models such as ELMO [24] and BERT [25].

To enrich the affective information into training, [3, 26, 27, 28, 29] introduced lexical resources to enrich previous word distributions with sentiment-informative features, as lexical values are intuitively associated with the word's sentiment polarity and strength. Especially, [26] propose a sentiment similarity-oriented attention (SSOA) mechanism, which uses the label embeddings and the valence value from affective lexicon to guide the network to extract emotion-related information from input sentences. [28] proposed a lexicon-based supervised attention model to extract sentiment-enriched features for document-level emotion classification. Similarly, [29] introduced a kind of affect-enriched word distribution, which was trained with lexical resources on the Valence-Arousal-Dominance dimensions. These studies demonstrate the effectiveness of sentiment lexicons in emotion recognition.

2.1.3 Multi modality

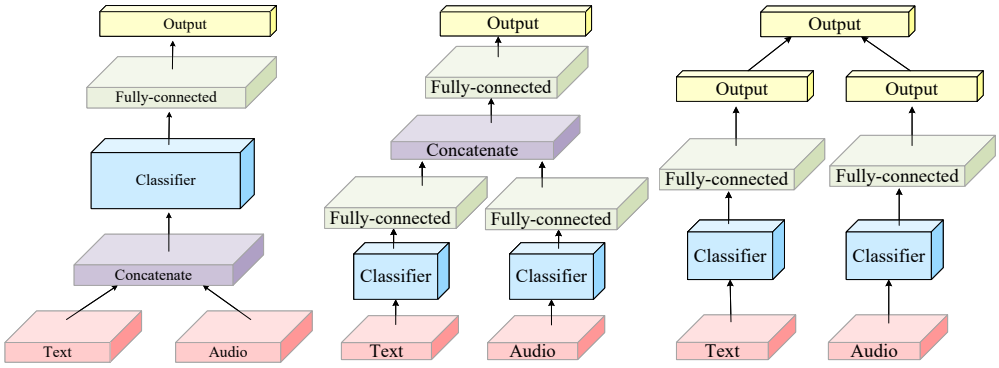
To detect the emotions in utterances, humans often consider both the textual meaning and prosody. Moreover, people tend to use specific words to express their emotion in spoken dialog, for example the use of swear words [30]. Along with the speech and text modalities, other visible cues such as gestures, facial expressions are also helpful to detect accurate emotions. Visual content plays an important role in emotion detection as facial expression is the straight way to express emotions and provide meaningful emotion-specific information. For example, when a person is angry, he frowns; when a person is sad, the corners of his mouth will fall. Recent researches such as [31] have demonstrated the effectiveness of facial representation in the emotion recognition task.

Whether these expressions are sufficient to identify emotions or not? However, it is much difficult to identify the emotion if people are good at concealing their emotions, such as sarcasm. Physiological response can depicts person's true reaction during an emotion, it will be useful for a large set of applications, if the affective state of user is available. The most common emotion representation model is the dimensional model which provides

arousal and valence values for a given range [32]. There exits various types of signals, which are used to record the bio-signals produced by the human's system, such as Electroencephalogram (EEG), Electrodermal Activity (EDA), Electromyography (EMG), Blood Volume Pulse (BVP) and so on. [33] applied physiological signals in the multimodal fusion framework for emotion detection. [34] integrated physiological and speech signals in the emotion recognition task.

A multimodal structure is thus necessary for using both the text and audio as input data[35]. The current research such as [3, 5, 36] on pattern recognition also shows that the use of multimodal features increases the performance compared to single modality. To accurately recognize human emotions, one of the challenges is the extraction of effective features from input data, while another is the fusion of different modalities.

There are three major fusion strategies [37] as shown in Figure 2.2: data/information fusion (low-level fusion), feature fusion (intermediate-level fusion), and decision fusion (high-level fusion). Data fusion combines several sources of raw data to produce new raw data that is expected to be more informative and synthetic than the inputs [37]. In intermediate-level feature fusion, data from each modality is first input to the best performing uni-modal networks which learn intermediate embeddings. The intermediate weights from these uni-modal networks are then concatenated and feed into another network such as fully connected layer to capture interactions between modalities [38]. Decision fusion uses a set of classifiers to provide a unbiased and more robust result. The outputs of all the classifiers are merged together by various methods to obtain the final output.



(a) Data fusion (b) Feature fusion (c) Decision fusion

Figure 2.2: Main fusion strategies multimodaltiy.

There are two typical methods for the annotation of emotion label, one

is the self annotation and another is the third-party's annotation. In most of the previous emotional corpus collections, the subjects are asked to express a given emotion, which is later used as the emotional label. A drawback of this approach is that it is not guaranteed that the recorded utterances reflect the target emotions. Additionally, a given display can elicit different emotional percepts. Therefore, the annotation method which based on agreements derived from subjective emotional evaluations by the third-party evaluator has been widely used in recent researches, such as [39][40]. In this thesis, my challenge is to predict the emotion label annotated by third party coders.

2.2 Emotion recognition in conversations

Due to the growing availability of public conversational data, emotion recognition in conversation (ERC) has gained more attention from the NLP community [6, 10, 9, 11]. ERC can be used to analyze conversations that happen on social media to mine emotion and opinion, rather than single utterance. It can also aid in analyzing contextual information in real times, which can be instrumental in human-robot interaction, interviews, and more [1].

2.2.1 Variables in conversations

Unlike utterance-level emotion recognition tasks, ERC relies on context architecture and modeling the contextual interaction of interlocutors. Poria et al classified conversations into two categories: task oriented and non-task oriented (chit-chat), meanwhile, factors such as topic, intent and speaker personality play the important role in the conversational interaction, as illustrated in Figure 2.3, in which grey and white circles represent hidden and observed variables, P represents personality, U represents utterances, S represents interlocutor state, I represents interlocutor intent, E represent emotion and *Topic* represents topic of the conversation [1]. It is a typical theoretical structure of dynamic interaction in conversation. Taking consideration of these factors would help modeling the discourse structure of the conversation and capture the true emotion and intention of the interlocutors.

For example, [41] exploited speaker identification as an auxiliary task to enhance the utterance representation in conversations. Topic modeling based on the subject's responses is significant to exploit global and time-varying statistics [42]. Genevieve Lam et al proposed a novel method that incorporated a data augmentation procedure based on topic modelling using transformer to capture contextual representations of text modality, and adopted 1D convolutional neural network (CNN) based on Mel-frequency

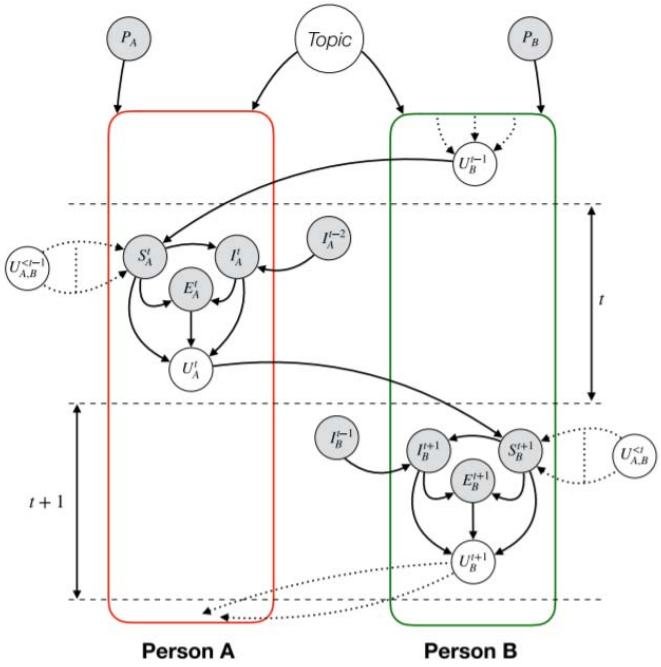


Figure 2.3: Interaction among different controlling variables during a dyadic conversation between speakers [1].

spectrogram to extract deep acoustic features. To capture contextual information from target utterances' surroundings in the same video, [6] proposed a LSTM-based model called bidirectional contextual long short-term memory (bc-LSTM), which are two unidirectional LSTMs stacked together having opposite directions. Therefore, the information from utterances occurring before and after itself can be captured. [9] applied three gated recurrent units (GRU) [43] to track the update of global context ,emotion and speaker state respectively. [44] proposed a new interaction-aware attention network (IAAN) that integrated contextual information in the learned acoustic representation through an attention mechanism. [45] came up with a deep neural architecture, incorporated with conversational memory network, which leverages contextual information from the conversation history. Such memories are merged using attention-based hops to capture inter-speaker dependencies. Studies such as [6, 9, 42, 45] are conducted based on multimodal representations, the results of these studies demonstrate that multimodal systems outperform the unimodal variants.

2.2.2 Conversational context modeling

There are two important factors in emotional dynamics in dialog: *self* and *inter-personal dependencies* [40]. Self-dependency can be also understood as *emotional inertia* [1], which depicts the emotional affects that speakers have on themselves during a conversation. Meanwhile, inter-personal dependencies represent the emotional influences that the counterpart induces on a speaker/listener. As shown in the Figure 2.4, person A has the emotion inertia of being *neutral*. But the emotion of person B was largely affected by person A. As person B' emotion was *neutral* at the begin, after the response U_5 of person A, the emotion of person B was changed to *anger*. It is obvious that the semantic meaning of U_5 displeased person B. And we can also see that U_8 conveys the emotion of sarcasm. It is challenging to detect the emotion of this utterance as the semantic meaning of itself is positive, but the true meaning should be negative, therefore, context modeling is essential to capture the real intention and emotion of this kind of utterances.

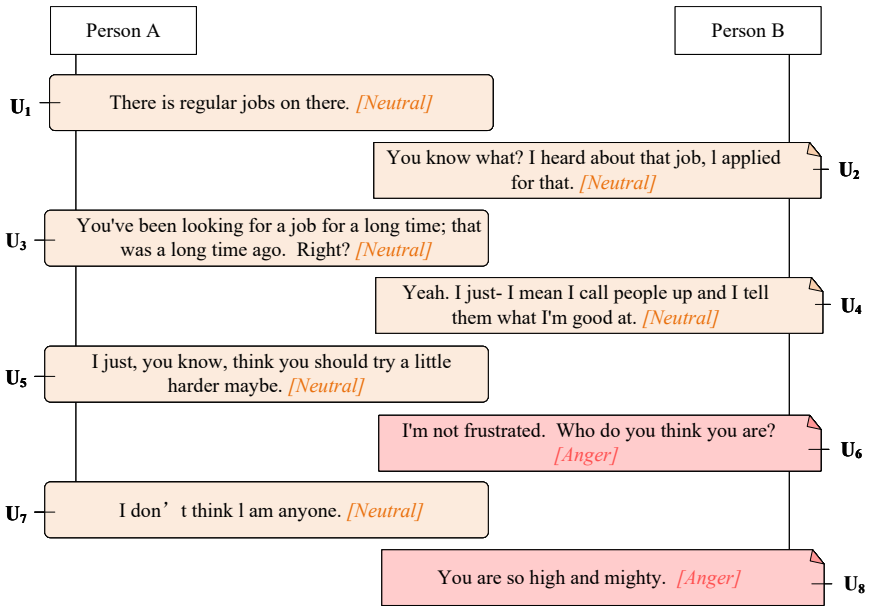


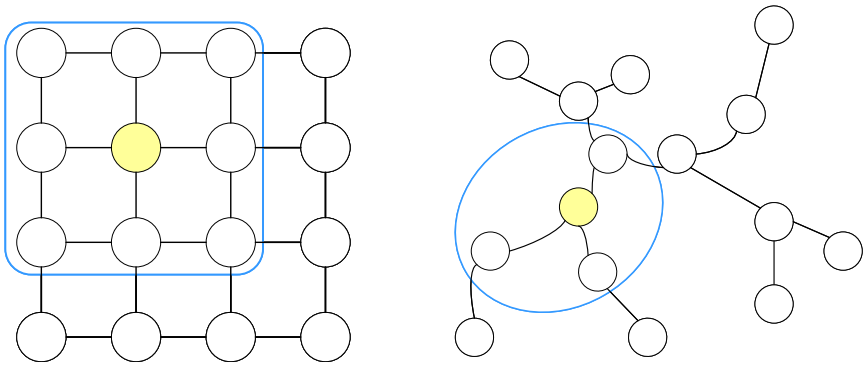
Figure 2.4: An example conversation from the IEMOCAP dataset

We assume that the surrounding utterances affect most for the target response, however, not only the contextual information from the local but also the distant conversational history are important for context modeling, especially in the situation that speaker refer to the topic and information from the distant context. Therefore, how to model the contextual sequence and

chose the most useful information in a conversation is a difficult but indispensable task. To further tackle these problems, deep neural networks such as RNN-based architecture [9], memory network [45], attention mechanism [44] has been widely used in previous researches.

2.3 Graph convolutional neural network

With the development of deep neural networks, the researche on pattern recognition and data mining has been a significant and popular topic. Methods such like CNN [4] has been widely used in the euclidean structure (e.g., images, text, and videos). Taking image data as an example, it can be considered as the regular grid in the euclidean space, and CNN is able to exploit the shift-invariance, local connectivity, and compositionality of image data [46]. Therefore, CNN can extract local deep meaningful features. However, there are many situations that data can not be displayed as euclidean structure, such as social network, e-commerce, information network, citation link, we can structure this kind of data in the form of graph, or non-euclidean architecture.



(a) 2-D convolution (b) Graph convolution

Figure 2.5: Euclidean Structure versus Non-Euclidean Structure.

Motivated by CNNs, RNNs, and other deep learning methods, new generalizations and definitions of important operations have been rapidly developed in the past few years to deal with the complexity of graph data. As shown in Figure 2.5, in (a), each pixel in an image can be taken as a node where neighbors are determined by the filter size. The 2-D convolution takes the weighted average of pixel values of the yellow node along with its neighbors. It is ordered and has a fixed size in the neighbors of a node. In (b), a graph convolution can be generalized from a 2-D convolution. An

image can be considered as a special case of graphs, where pixels are connected by adjacent pixels. Similar to 2-D convolution, the operation of graph convolution is taking the weighted average of yellow one's neighborhood information, however, different from the structure in (a), the neighbors of a node are unordered and variable in size [47].

There are several variances in graph neural networks, such as Recurrent GNNs (RecGNN) [48] , Convolutional GNNs (ConvGNNs) [49], Convolutional recurrent GNNs (GCRN) [50], Graph Autoencoders (GAEs) [51], and Spatial-Temporal GNNs (STGNNs) [52]. In our studies, we focus on the ConvGNNs , which generalize the operation of convolution from grid data to graph data. The main idea is to generate the representation of a node by aggregating its own features and surrounding features.

Convolutional graph neural networks have been widely used in the pattern recognition community. There are two categories of ConvGNNs, spectral-based and spatial-based. In spectral-based approaches, the properties of a graph are in relationship to the characteristic polynomial, eigenvalues, and eigenvectors of matrices associated with the graph, such as its adjacency matrix or Laplacian matrix. Spatial-based approaches extract the spatial features on the topological graph based on the neighbors of each vertex. GCN [49] bridged the gap between spectral-based and spatial-based approaches, spatial-based methods have developed rapidly due to its competitive advantages in efficiency, flexibility and generality [47]. As for the graph-based neural network model $f(X, A)$, the layer-wise propagation rule of a multi-layer Graph Convolutional Network (GCN) is displayed as following [49]:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{2.1}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix of the undirected graph with added self-connections. I_N is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{i,j}$ and $W^{(l)}$ is a layer-specific trainable weight matrix. $\sigma(\cdot)$ represents an activation function, such as the $ReLU(\cdot) = max(0, \cdot)$. $H^{(l)} \in R^{N \times D}$ is the matrix of activation in the l^{th} layer. $H^{(l)} = X$.

In the literature, GCN has been widely used in several works recently, such as text classification [53], aspect-level sentiment classification [54], emotion recognition in conversations [11], and have achieved competitive performance, where GCN is used to encode the syntactic structure of sentences.

Inspired by GCN which operates on local graph neighborhoods, [2] proposed the Relational Graph Convolutional Networks (R-GCNs) to extend GCNs to large-scale relational data, such as in knowledge graphs. As shown in figure 2.6, the representation of the surrounding nodes (blue) and self (red) are accumulated and then transformed based on every relation type,

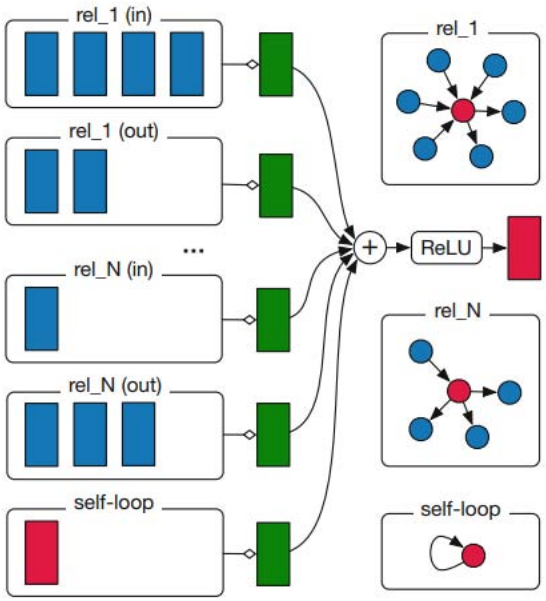


Figure 2.6: Diagram for computing the update of a single graph node/entity (red) in the R-GCN model proposed in [2].

then the result embedding (green) is gathered in a normalized sum and passed through an activation function. The directed and labeled multi-graph can be denotes as $G = (V, E, R)$, meanwhile, V is the nodes (entities) set, and E is the labeled edges (relations) set, and R represents the relation type which contains both *born_in* and *born_in_inv*. And the propagation model for calculating the forward-pass update of an entity with surrounding edges in a relational multi-graph is defined as:

$$h_i^{(l+1)} = \sigma(\sum_{r \in \Re} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)}) \tag{2.2}$$

where N_i^r denotes the set of neighbor indices of node i under relation $r \in R$. $c_{i,r}$ is a problem-specific normalization constant that can either be learned or chosen in advance.

$$W_r^{(l)} = \sum_{b=1}^{B} a_{rb}^{(l)} V_b^{(l)} \tag{2.3}$$

$$W_r^{(l)} = \bigoplus_{b=1}^{B} Q_{br}^{(l)} \tag{2.4}$$

However, to regular the weights of R-GCN layers, especially in highly multi-relational graph, Michael Schlichtkrull et al also came up with two approaches: *basis-* (Formula 2.3) and *block-diagonal-* decomposition (Formula 2.4). Where $V_b^{(l)} \in R^{d^{(l+1)} \times d^{(l)}}$ is the basis transformation, coefficient $a_{rb}^{(l)}$ depends on r. And $Q_{br}^{(l)} \in R^{(d^{(l+1)}/B) \times (d^{(l)}/B)}$. The block decomposition structure encodes an intuition that latent features can be grouped into sets of variables which are more tightly coupled within groups than across groups [2]. The R-GCN architecture has a competitive advantage in both link prediction and entity classification with relational data. Our graph convolution is closely related to this work.

2.4 Knowledge base in emotion recognition

The knowledge base has attracted increasing attention in several research areas such as open-domain dialogue systems [55], question answering systems [56], cross-domain sentiment analysis [57], aspect-based sentiment analysis [58], and emotion detection in conversations [59]. Commonsense knowledge bases help in grounding text to real entities, factual knowledge, and commonsense concepts. In particular, commonsense 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. [58] proposed a knowledge-guided capsule network, which incorporates syntactical and n-gram information as the prior knowledge to guide the capsule attention process in aspect-based sentiment analysis. [59] 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. We construct knowledge graph based on the selected concepts first. Then we apply our knowledge graph to guide the semantic edge weighting of GCN, which helps to capture significance context-sensitive information of conversations with both implicit and explicit emotional texts.

Chapter 3

Proposed Model

Human-computer interaction has become prevalent in various fields, especially for spoken dialogue systems and intelligent voice assistants. Emotions, which are often denoted as an individual's mental state associated with thoughts, feelings, and behavior, can significantly help the machine to understand the user's intention. Therefore, accurately distinguish user's emotions can enable great interactivity and improve user experiences.

Contextual dependence is significant for emotion recognition, as the intention and emotion of the target utterance are mostly affected by the surrounding contexts. Unlike traditional methods, which based on individual utterances, conversational emotion recognition utilizes the relation among utterances to track the user's emotion states during conversations. However, it's a challenging task to effectively model the interaction of different speakers in the conversational dialog. To solve this problem, previous studies such as [9][44] proposed the LSTM-base methods for sequential encoding of contexts. However, this kind of method has the issue of sequence propagation, which may not perform well on long-term context extraction, as the emotion effect to the target utterance from the long-distance may decrease or even vanish. [10] [11] [54] applied GCN-based architecture to extract neighborhood contextual information, which solve the issue of sequence propagation, and the result of these works also demonstrates that GCN are good at modeling both inter-interaction and intra-dependence of the user in a conversation, which are the important factors in the task of conversational emotion recognition. However, for implicit emotional texts that do not contain obvious emotional terms, it is difficult to correctly distinguish the emotion if only the semantics of the utterances are considered. Moreover, the lack of sufficient labeled public databases is still an issue. It's difficult to extract enough information for emotion recognition because of the small scale of samples.

Knowledge bases provide a rich source of background concepts related

by commonsense links, which can enhance the semantics and emotion polarity of one utterance by providing context-specific concepts. Therefore, to further tackle the above problems, we propose a new multimodal Semantic- and Knowledge-guided Graph Convolutional Network (ConSK-GCN) to effectively structure the semantic-sensitive and knowledge-sensitive contextual dependence in each conversation. In this capture, I will introduce my methodology based on four parts: database preparation, multimodal features extraction, knowledge retrieval and ConSk-GCN construction, as shown in the figure 3.1.

In detail, we will introduce the preparation of both text and audio data in section 3.1; The detailed description on how to extract textual and acoustic representations and how to fuse these two modalities will be claimed in section 3.2; In section 3.3, we focus on the external knowledge retrieval of both knowledge base and affective lexicon. Then we will introduce the construction of ConSK-GCN in section 3.4. In particular, on the one hand, we construct the contextual inter-interaction and intradependence of the interlocutors via a conversational semantic-guided GCN (ConS-GCN). On the other hand, we incorporate an external knowledge base that is fundamental to understand conversations and generate appropriate responses to enrich the semantic meaning of the tokens in the utterance via a conversational knowledge-guided GCN (ConK-GCN). Meanwhile, we introduce an affective lexicon into knowledge graph construction to enrich the emotional polarity of each concept. Then, 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.

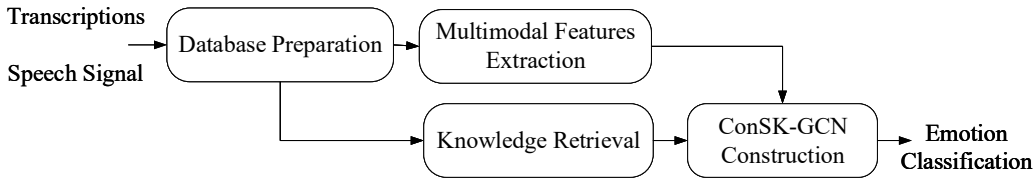


Figure 3.1: Overall architecture of our proposed ConSK-GCN approach for multimodal emotion recognition

3.1 Database preparation

We evaluate our ConSK-GCN on two conversational databases, namely *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) [39] and *Multimodal*

EmotionLines Dataset (MELD) [40]. Both these datasets are multimodal datasets containing text, audio and video modalities for each conversation. In our work, we focus on multimodal emotion recognition with the modality of text and audio. However, multimodal emotion recognition with all these three modalities is left as future work.

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 dialogues with four emotion categories with the distribution of 29.6% happiness, 30.9% neutral, 19.9% anger, and 19.6% sadness. 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.

MELD database was evolved from the *EmotionLines* database which is collected by Chen et al.[60]. *EmotionLines* was developed by crawling the dialogues from each episode in the popular sitcom *Friends*, where each dialogue contains utterances from multiple speakers. Poria et al. extend *EmotionLines* into around 13000 utterances from 1433 dialogues with the distribution of 46.95% neutral, 16.84% joy, 11.72% anger, 11.94% surprise, 7.31% sadness, 2.63% disgust, 2.61% fear. The data distribution in train, validation and test set are shown in Table 3.1. And the statistics of all the emotions are displayed in Table 3.2.

Table 3.1: Statistics of the IEMOCAP and MELD dataset

Dataset	Dialogues			Utterances			Classes
	Train	Val	Test	Train	Val	Test	
IEMOCAP	120		31	4290		1241	4
MELD	1039	114	280	9989	1109	2610	7

Table 3.2: Emotions distribution in IEMOCAP and MELD dataset

	IEMOCAP		MELD		
	Train/Val	Test	Train	Val	Test
Neutral	1325	384	4710	470	1256
Happiness/Joy	1195	442	1743	163	402
Anger	931	170	1109	153	345
Surprise	-	-	1205	150	281
Sadness	839	245	683	111	208
Disgust	-	-	271	22	68
Fear	-	-	268	40	50

Furthermore, to better mine the information of the raw data and capture efficient contextual traits, we prepare the text and audio data firstly. As for context construction, we first display the textual data of each dialogue in context sequence, and the sequence order of audio corresponds to the text, as shown in figure 3.2.

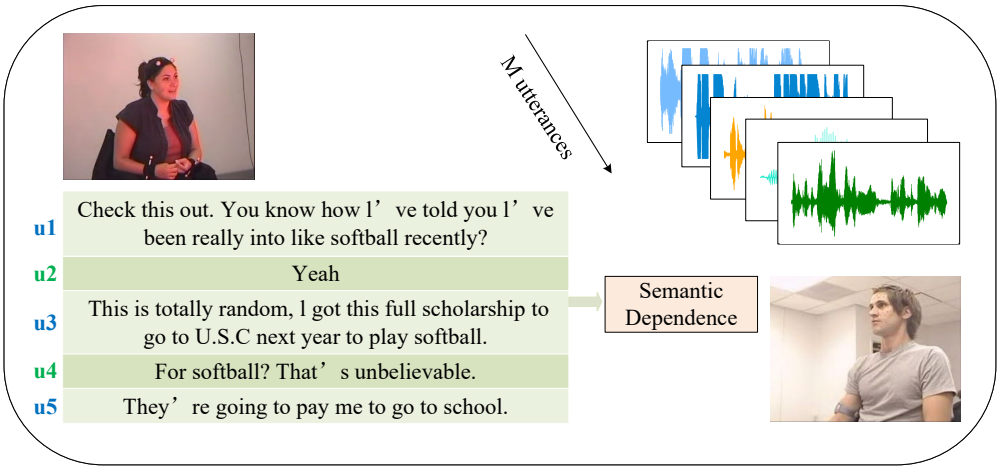


Figure 3.2: Architecture of database preparation

3.2 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, to initialize each modality, we train separate networks to extract linguistic and acoustic features at the utterance level with emotion labels.

3.2.1 Textual features

We employ different approaches to extract utterance-level linguistic features for IEMOCAP and MELD datasets based on the particular traits of these two datasets. Formally, the textual representation of an utterance is denoted as $\boldsymbol{\mu_t}$.

IEMOCAP:

To compare with the state-of-the-art approaches, we employ the traditional and most used convolutional neural network [4] to extract textual embeddings

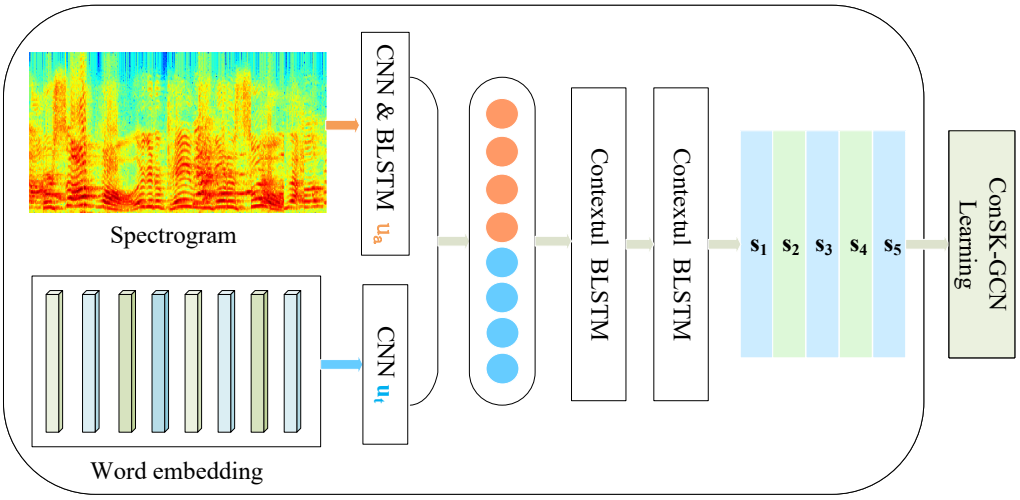


Figure 3.3: Architecture of multimodal features extraction

of the transcripts. First, we use the publicly available pretrained word2vec [61] 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 [62]. These are then concatenated and fed into two fully connected layers with 500 and 100 hidden nodes separately followed by the ReLU activation.

MELD:

The average utterance length and average turn length are 8.0 and 9.6 in the MELD database, which is 15.8 and 49.2 in IEMOCAP database [40]. The utterances in MELD are shorter and the context-dependence is not strong as in IEMOCAP. Therefore we consider that the approach mentioned above is insufficient to extract effective latent representations of the utterances in MELD. Considering that *BERT_BASE* [25] has shown the state-of-the-art performance in many NLP tasks, such as reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations, therefore we apply *BERT_BASE*, the model architecture of which is a multi-layer bidirectional Transformer encoder to initialize the textual representations. Firstly, we fine-tune the pre-trained *BERT_BASE* model with 12 Transformer blocks, 768 hidden sizes, 12 self-attention heads, and 110M total parameters based on the training samples, and use the test samples for emotion label prediction. Then, we take the representations of

both training and test samples from the penultimate dense layer as the context independent utterance level feature vectors.

3.2.2 Acoustic features

In this paper, we follow the audio preprocessing method introduced in [63]. Researchers have found that a segment speech signal that is greater than 250-*ms* includes sufficient emotional information [64]. As the average utterance length of IEMOCAP dataset is around 2-*s*, and it's about 3.6-*s* in MELD dataset [40]. Therefore, for IEMOCAP dataset, the time of each segment is set to 265-*ms* and the slide window is set to 25-*ms*, then the input spectrogram has the following *time* $\times$ *frequency*: 32 $\times$ 129. For MELD dataset, we apply a 2-*s* window size with a slide window of 1-*s* to transform an utterance into several segments, and the size of the spectrogram is 1874 $\times$ 129.

Two 2-dimensional CNNs are utilized to extract deep acoustic features from the segment-level spectrograms. We use convolutional filters of size (5,5) with 32 and 65 feature maps for each CNN layer. The window size of max-pooling is set as (4,4) followed by the ReLU activation. 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 $\boldsymbol{\mu_a}$.

3.2.3 Multimodal fusion

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 layer with 400 and 300 hidden units respectively. It's for sequence encoding to obtain the global utterance-level contextual information. Formally, we denote the context-aware multimodal representations as $\boldsymbol{s}$:

$$s_i = BiLSTM(s_{i(+,-)}, u_i) \tag{3.1}$$

where i=1,2,...,N, and N represents the number of samples, u_i and s_i are context-independent and sequential context-sensitive utterance-level representations respectively, and $s_{i(+,-)}$ means the forward and backward sequential information of utterance i.

3.3 Knowledge retrieval

In this paper, we utilize external commonsense knowledge base ConceptNet [65] and an emotion lexicon NRC_VAD [66] 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 represent the general knowledge involved in understanding language, improving natural language applications by 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 Figure 3.4, 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"*. For English, ConceptNet comprises 5.9M assertions, 3.1M concepts and 38 relations. 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.

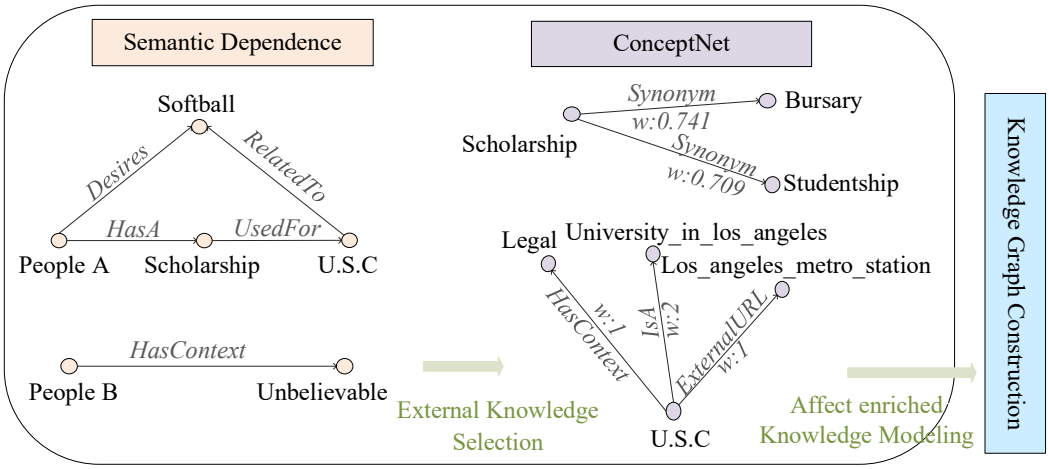


Figure 3.4: Architecture of external knowledge retrieval

3.4 ConSK-GCN construction

Figure 3.5 shows the architecture of our proposed ConSK-GCN approach for multimodal emotion recognition.

3.4.1 Knowledge graph construction

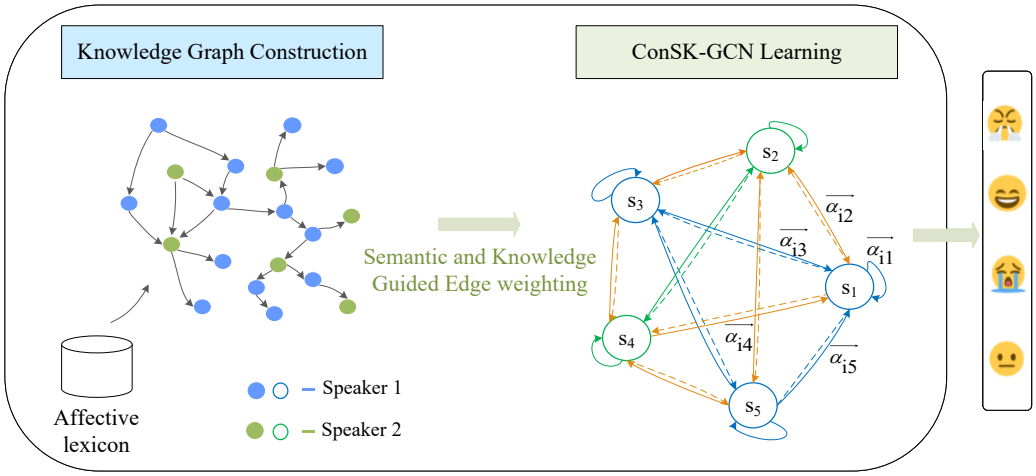


Figure 3.5: Architecture of ConSK-GCN construction

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 each concept $c_{i,m}$ in V_k, we retrieve the corresponding valence (V) and arousal (A) scores from NRC_VAD, respectively, where m ranges from 1 to n, and n is the number of concepts in each utterance.

Each node/concept in the knowledge graph is embedded into an effective semantic space, named *ConceptNet Numberbatch*, that learns from both distributional semantics and ConceptNet using a generalization of the "retrofitting" method [67]. The tokens that are not included in the ConceptNet are initialized by the "fastText" method [68], which is a library for efficient learning of word representations. Formally, we denote each encoded concept embedding as $C(c_{i,m})$.

The edges in the knowledge graph represent the knowledge relatedness between the concepts. First, for the m_th concept in the i_th utterance, we adopt l_2 norm to compute the emotion intensity $emo_{i,m}$, following [59], that

is,

$$emo_{i,m} = min - max(\|[V(c_{i,m}) - 1/2, A(c_{i,m})/2]\|_2) \tag{3.2}$$

where $||.||_2$ denotes l_2 norm, $V(c_{i,m}) \in (0,1)$ and $A(c_{i,m}) \in (0,1)$ represent the corresponding valence and arousal score for the m_th concept $c_{i,m}$ in utterance i. For the 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, 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:

$$k_{i,m} = emo_{i,m} C(c_{i,m}) \tag{3.3}$$

$$a^k_{i,j} = \sum_{m=1}^{n_i} \sum_{m=1}^{n_j} abs\left(\cos(k^\top_{i,m} W_k \left[\, k_{j,1}, ..., k_{j,m}\right])\right) \tag{3.4}$$

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

3.4.2 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 [69] to compute the semantic similarity between two utterances, which is computed as the cosine similarity of two utterances first, and then employ *arccos* to convert the cosine similarity into an angular distance, that is,

$$sim_{i,j} = 1 - \arccos\left(\frac{s_i^\top s_j}{\|s_i\| \, \|s_j\|}\right)/\pi \tag{3.5}$$

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

$$a^s_{i,(i-p,...,i+f)} = softmax(W_s[sim_{i,i-p}, ..., sim_{i,i+f}]) \tag{3.6}$$

where $j = i - p, ..., i + f$. 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.

3.4.3 ConSK-GCN learning

We build our semantic- and knowledge-guided 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,

$$a_{i,j} = \omega_k a^s_{i,j} + (1 - \omega_k) a^k_{i,j} \tag{3.7}$$

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 [2] to capture local contextual information that is both semantic-aware and knowledge-aware:

$$h^{(1)}_i = \sigma(\sum_{r \in \Re} \sum_{j \in N^r_i} \frac{a_{i,j}}{q_{i,r}} W^{(1)}_r s_j + a_{i,i} W^{(1)}_0 s_i) \tag{3.8}$$

$$h^{(2)}_i = \sigma(\sum_{j \in N^r_i} W^{(2)} h^{(1)}_j + W^{(2)}_0 h^{(1)}_i) \tag{3.9}$$

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 $W^{(1)}_r \in R^{d_{h^{(1)}_i} \times d_{s_j}}$, $W^{(1)}_0 \in R^{d_{s_j} \times 1}$, $W^{(2)} \in R^{d_{h^{(1)}_i} \times d_{h^{(2)}_i}}$, $W^{(2)}_0 \in R^{d_{h^{(2)}_i} \times 1}$ are model parameters, $\sigma(.)$ is the activation function such as ReLU.

This stack of transformations, Eqs. (3.7) and (3.8), effectively accumulates normalized sum of neighborhood features and self-connected features. Then, the global contextual vectors $\boldsymbol{s}$ as well as the local neighborhood-based contextual vectors $\boldsymbol{h^{(2)}_i}$ are concatenated to obtain the final representations as following:

$$v_i = [s_i, h^{(2)}_i] \tag{3.10}$$

Furthermore, the utterance is classified using a fully connected network:

$$l_i = ReLU(W_l v_i + b_l) \tag{3.11}$$

$$P_i = softmax(W_p l_i + b_p) \tag{3.12}$$

$$\hat{y}_i = \underset{k}{argmax}(P_i[k]) \tag{3.13}$$

where k is the classes of each database, and $\hat{y}_i$ is the predicted emotion class.

We use categorical cross-entropy as well as L2-regularization to compute the loss (L) during the training, that is:

$$L = -\frac{1}{\sum_{s=1}^{M} d(s)} \sum_{j=1}^{M} \sum_{i=1}^{d(s)} logP_{i,j}[y_{i,j}] + \lambda \left\|\theta\right\|_2 \tag{3.14}$$

where M is the number of dialogues in each database, $d(s)$ is the number of utterances in dialogue s, $P_{i,j}$ is the probability distribution of emotion labels for utterance i in dialogue j, $y_{i,j}$ is the label of ground truth of utterance i in dialogue j. And λ is the L2-regularizer weight, θ is the set of all trainable parameters.

Chapter 4

Experimentation

4.1 Experimental setup

We choose ReLU as the activation and apply the method of stochastic gradient descent based on Adam [70] optimizer to train our network and all the hyperparameters are optimized by grid search. We set the batch size and number of epochs to 32 and 100, respectively. 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 dataset, 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 0.5 in both IEMOCAP and MELD databases to balance the effect of knowledge and semantics.

4.2 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[4]: 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[5]: Adopted LSTM framework for unimodality and multimodality emotion recognition based on audio and text, without exploring context information.

bc-LSTM[6]: 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[9]: 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[10]: Adopted GCN to leverage self and interspeaker dependence of the interlocutors to model conversational context for textual emotion recognition.

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.

4.3 Experimental results and analysis

4.3.1 Experiments on IEMOCAP

Comparation in unimodality

Table 4.1 indicates the performance of both state of the arts 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. In addition, "DialogueRNN" and "DialogueGCN" are both superior to "bc-LSTM", highlighting the importance of encoding speaker-level context while "bc-LSTM" only encodes sequential context. 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 [71] 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

Table 4.1: The accuracy-score (%) of comparative experiments of different methods for unimodality (Text) emotion recognition. Average (w)= Weighted average; bold font denotes the best performances.

	Models	Neutrality	Anger	Happiness	Sadness	Average (W)
Baselines	CNN	59.11	77.06	64.03	62.04	63.90
	LSTMs	72.92	70.00	55.20	63.67	64.38
	bc-LSTM	76.04	75.88	67.65	67.35	71.31
	DialogueRNN	**81.51**	66.47	86.43	72.24	79.37
	DialogueGCN	74.22	77.06	87.56	85.31	81.57
Ablation Studies	ConS-GCN	76.04	77.65	87.33	83.27	81.71
	ConK-GCN	75.52	77.65	86.65	86.12	81.87
Proposed	ConSK-GCN	74.48	**80.00**	**87.78**	**89.39**	**82.92**

Table 4.2: The F1-score (%) of comparative experiments of different methods for unimodality (Text) emotion recognition.

	Models	Natural	Anger	Happiness	Sadness	Average (W)
Baselines	CNN	59.50	65.17	69.36	60.68	64.02
	LSTMs	74.97	65.93	56.42	61.30	64.42
	bc-LSTM	67.51	72.88	75.51	70.06	71.60
	DialogueRNN	73.73	74.10	87.82	77.29	79.50
	DialogueGCN	74.32	76.61	88.66	83.60	81.55
Ablation Studies	ConS-GCN	74.68	77.65	88.74	83.27	81.79
	ConK-GCN	75.23	77.88	88.05	84.06	81.90
Proposed	ConSK-GCN	**75.66**	**78.84**	**88.79**	**86.39**	**82.89**

"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 4.1, while it shows a more significant increase by nearly 8% in Table 4.3. 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.

Comparation in multimodality

Table 4.3 and 4.4 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.

Table 4.3: The accuracy-score (%) of comparative experiments of different methods for multi-modality (Text+Audio) emotion recognition.

	Models	Neutrality	Anger	Happiness	Sadness	Average(W)
Baselines	LSTMs	69.53	73.53	66.74	70.61	69.30
	bc-LSTM	79.95	78.82	70.14	73.88	75.10
	DialogueRNN	**86.20**	84.71	79.64	75.10	81.47
Ablation Studies	ConS-GCN	78.91	85.29	90.72	78.78	83.96
	ConK-GCN	75.78	**88.82**	89.37	**86.53**	84.53
Proposed	ConSK-GCN	78.13	87.06	**93.67**	82.86	**85.82**

Table 4.4: The F1-score (%) of Comparative experiments of different methods for multimodality (Text+Audio) emotion recognition.

	Models	Natural	Anger	Happiness	Sadness	Average(W)
Baselines	LSTMs	63.95	73.10	73.75	67.98	69.50
	bc-LSTM	70.49	77.91	78.58	75.73	75.42
	DialogueRNN	76.53	83.72	86.38	80.35	81.78
Ablation Studies	ConS-GCN	77.79	83.57	90.52	82.13	83.97
	ConK-GCN	77.70	**85.31**	90.08	84.46	84.49
Proposed	ConSK-GCN	**79.89**	84.33	**91.90**	**84.76**	**85.74**

Furthermore, compared with unimodality in Table 4.1, 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.

4.3.2 Experiments on MELD

Comparation with the state of the art

Table 4.5 and 4.6 depict the experimental comparations between our model and previous works in emotion recognition with both unimodality and multimodality. We can see from both table 4.5 and 4.6 that, our model which constructs both knowledge-sensitive and semantics-sensitive contexts has a better performance with more than 5.7% than the state of the arts in terms of weighted average f1-score in both unimodal and multimodal emotion recognition. However, comparing to IEMOCAP database, the effectiveness by multimodal fusion in MELD is limited. We think it's because Friends TV series include multiple audio sources, not only speaker's voice, but other audio source, so it is difficult to detect acoustic features clearly. Moreover, the average conversation length in IEMOCAP is 49.2, with only two participants in IEMOCAP. And the average conversation length in MELD is 9.6, with many conversations having more than 5 participants, which means majority of the participants only utter a smaller number of utterances per conversation, there are more noises when detect the target speaker's emotion in MELD.

Table 4.5: Comparative experiments of different methods for unimodality (Text) emotion recognition. F1-score (%) is used as the evaluation metric. W= Weighted average.

Models	Neutral	Anger	Disgust	Joy	Surprise	Sadness	Fear	W-F1
CNN [4]	67.3	12.2	0.0	32.6	45.1	19.6	0.0	45.5
LSTMs [5]	67.6	12.3	0.0	36.0	45.7	17.2	0.0	46.0
bc-LSTM [6]	77.0	38.9	0.0	45.8	47.3	0.0	0.0	54.3
DialogueRNN [9]	73.7	41.5	0.0	47.6	44.9	23.4	5.4	55.1
DialogueGCN [10]	-	-	-	-	-	-	-	58.1
ConS-GCN	77.0	50.3	**2.9**	58.8	59.1	35.8	0.0	62.0
ConK-GCN	**80.0**	51.6	0.0	56.3	58.1	35.1	**13.7**	61.9
ConSK-GCN (Ours)	78.1	**54.1**	0.0	**61.1**	**61.0**	**36.9**	10.5	**63.8**

However, the data ratio of disgust only accounts for 2.63% in MELD database, while the percentage of fear is around 2.61%, therefore it is difficult to accurately distinguish these two emotions in ERC task. The task for emotion detection with small data, which may depends on specific emotional characteristics, is left as future work.

Table 4.6: Comparative experiments of different methods for multimodality (Text+ Audio) emotion recognition.

Models	Neutral	Anger	Disgust	Joy	Surprise	Sadness	Fear	W-F1
LSTMs [5]	68.1	31.4	0.0	34.5	44.9	7.24	0.0	47.6
bc-LSTM [6]	76.4	44.5	0.0	49.7	48.4	15.6	0.0	56.8
DialogueRNN [9]	73.2	45.6	0.0	53.2	51.9	24.8	0.0	57.0
ConS-GCN	77.7	52.2	0.0	60.4	58.9	37.0	0.0	62.9
ConK-GCN	77.5	52.6	0.0	60.9	**62.0**	33.3	0.0	63.0
ConSK-GCN (Ours)	**78.8**	**53.4**	0.0	**63.2**	60.1	**38.9**	0.0	**64.3**

Ablation Studies

To further research and validate the performance of the proposed model, the comparative confusion matrices of classification results are shown in Figure 4.1, 4.2 and 4.3 separately.

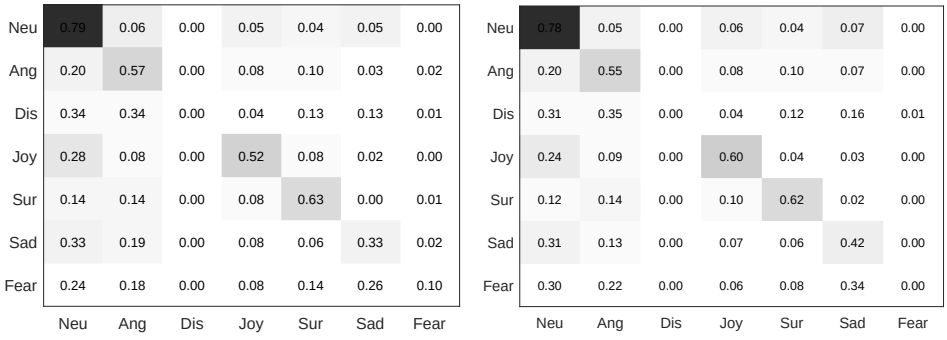
Compared with "ConS-GCN" and "ConK-GCN", the results shown in "ConSK-GCN" indicate that the knowledge-aware contexts and semantic-aware contexts are complementary for extracting efficient contextual features for better emotion recognition. There are two exceptions about *anger* and *surprise*, the detection rate of which is not highest in "ConSK-GCN", however, the false detection rate in "ConS-GCN" and "ConK-GCN" are also both far higher than "ConSK-GCN", which means more samples of *anger* and *surprise* have been misclassified.

(a) Unimodality (b) Multimodality

Figure 4.1: Confusion matrix of the proposed ConS-GCN.

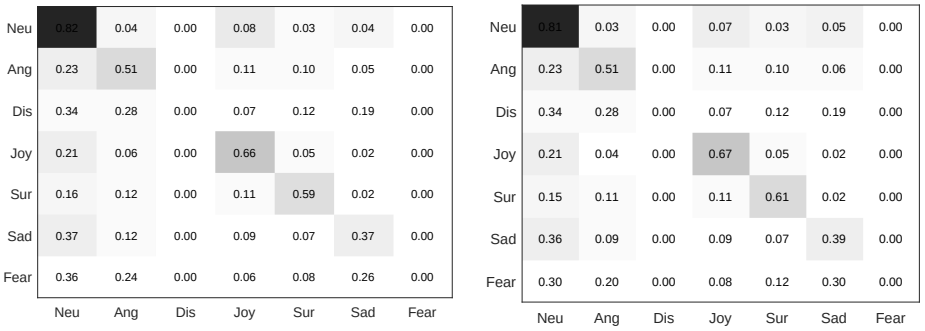
We can see from Figure 4.3 that, compared with (a), the results shown in (b) indicate that multimodality helps to improve the accuracy of emotion detection in conversations. The results demonstrate the importance of inte-

grating acoustic and linguistic features that are complementary in emotion recognition.



(a) Unimodality (b) Multimodality

Figure 4.2: Confusion matrix of the proposed ConK-GCN.



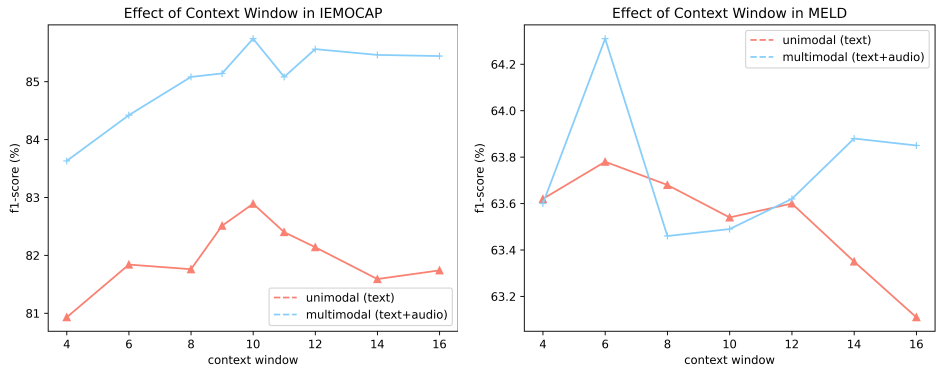
(a) Unimodality (b) Multimodality

Figure 4.3: Confusion matrix of the proposed ConSK-GCN.

4.4 Effect of Context Window

The accuracy of emotion detection in conversation varies with the context window. From Figure 4.4 (a), we can see that window sizes of 8-12 show better performance, and it reaches the peak when the past and future contexts are all set to 10 in the IEMOCAP dataset.

In the MELD dataset, we can conclude from the Figure 4.4 (b) that the window sizes of the past and future contexts are all set to 6 have the best performance, we think it is because the average conversation length is only 9.6 in MELD.



(a) IEMOCAP (b) MELD

Figure 4.4: Effect of context window for emotion recognition in different datasets.

4.5 Case study

To verify the effectiveness of external knowledge and semantic construction in conversational emotion recognition, we visualize several typical samples, as shown in Figure 4.5.

We can observe that compared to "ConS-GCN", which only considers the semantics of context, our proposed "ConK-GCN" and "ConSK-GCN" that take the advantages of external knowledge can effectively capture implicit emotional characteristics, as shown in utterance 1-3. We can see from utterance 4 that, in some cases, the modeling of semantics-sensitive or knowledge-sensitive context alone is not sufficient to accurately distinguish the emotion, but it's helpful when leveraging these two factors together.

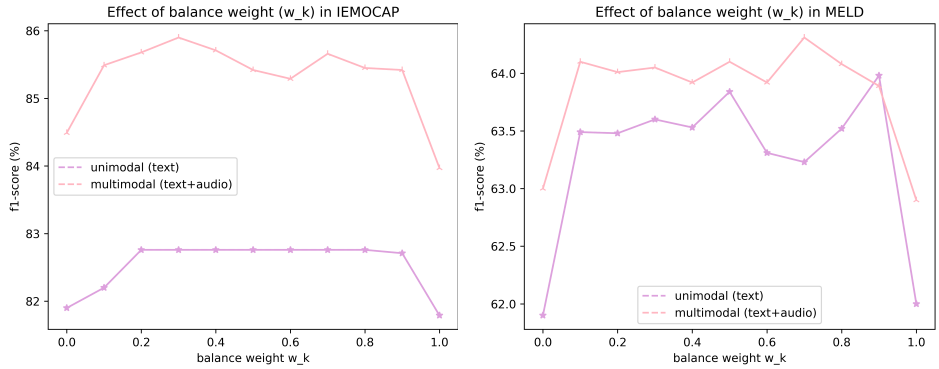
Our model misclassifies the *Neutrality* emotion of utterance 5; we attribute this result to the fact that the concept embeddings of the utterance are enriched by emotional knowledge, misleading the model and resulting in wrong detection, for example, "cool" in utterance 5 represents modal particle with no actual meaning, while it has several related implications such as "unemotional", "chill", and "unfriendly" with negative orientation in knowledge bases, which leads to the false detection.

Cases in utterance 6-7 and the cases in utterance 8-9 are in the same situation with opposite results, where "ConS-GCN" weights more than "ConK-GCN" in "ConSK-GCN" learning, but information in "ConS-GCN" oriented to wrong direction in utterance 6-7, vice verse in utterance 8-9. External knowledge, sometimes it can enrich the implicit concepts with helpful implications, however, emotion understanding is a challenging task as it not only

depends on semantic understanding but also contextual reasoning, it is important to make a balance between them. And the impact of balance weight between contextual semantics and external knowledge will be explained in the next section 4.6.

4.6 Effect of w_k

In order to find an optimal balance between knowledge weight and semantic weight in our ConSK-GCN learning, we make one pair of comparative experiments, that is unimodality and multimodality separately based on IEMOCAP and MELD databases.



(a) IEMOCAP (b) MELD

Figure 4.6: Effect of balance weight (w_k) for emotion recognition in different dataset.

We can conclude from Figure 4.6 that, both knowledge-aware and semantic-aware contextual construction are important for emotion recognition in conversation, as the f1-score of leveraging knowledge and semantics together (w_k ranges from 0.1 to 0.9) increased dramatically than single (w_k equal to 0 or 1). However, it seems that the effect of different balance weights (0.1 to 0.9) on emotion detection is not conspicuous, because in Figure 4.6 (a) and (b), the difference in the f1-score of different balance weight does not exceed 1%. Therefore, we set the balance weight w_k to 0.5 in both IEMOCAP and MELD databases to balance the effect of knowledge and semantics.

	Utterances	Gold_label	ConS-GCN (w_k=1)	ConK-GCN (w_k=0)	ConSK-GCN (w_k=0.5)	Knowledges in ConcepNet
1	I'll **get out**. I'll go get married and live someplace else. I don't know, maybe New York.	A	N ✗	A ✓	A ✓	Escape, Difficulty
2	So I was one of the first ones? That makes me feel so **important**.	H	N ✗	H✓	H ✓	Imperative, Friends, Benefits, to be good
3	Being **dishonest** with him. It is the kind of thing that pays off.	S	N ✗	S ✓	S ✓	Hurt someone else, Deceitful
4	We will go out to **dinner** later this week.	H	N ✗	N ✗	H ✓	A good time for socialization, Party
5	**Cool**, if you want me to go with you, I will.	N	N ✓	S ✗	S ✗	Unemotional, Chill, Unfriendly
6	I think **infantry**. I'm not sure.	S	N✗	S ✓	N✗	Trench, Artillery-battalion, Colour_sergeant
7	Wouldn't you? Oh, come on. you just tell me. You would make an **exception** for me.	A	N ✗	A ✓	N ✗	Unhandled, Exclusion, Unlike
8	Well, just don’ t **give up**, something might be around the corner tomorrow.	N	N ✓	S ✗	N✓	Reach an impasse, Capitulate
9	Have a **good** day.	N	N ✓	H ✗	N✓	Favorable, Satisfactory

Figure 4.5: Visualization of several representative examples. Blue denotes the typical concept in each utterance.

Chapter 5

Conclusion

5.1 Summary

In this thesis, we proposed a new conversational semantic- and knowledge-guided 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 both IEMOCAP and MELD databases 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 multispeaker conversations and model the speaker dynamics and emotion shifts for better emotion recognition.

5.2 Contribution

This thesis proposed a new GCN-based model for multimodal emotion recognition, which incorporating external knowledge and affective lexicon into semantic understanding. Experiments on two databases demonstrate that the proposed methodology can effectively improve the accuracy of emotion detection in conversation, especially for the document with implicit emotion expression. Knowledge base enriched the semantics of each utterance in conversation with several related concepts, and affective lexicon enhance the emotion polarity of each concept in the conversation. Moreover, this technology can be applied as an important part of the human-robot system to

enhance emotional interaction and improve user experience.

5.3 Future works

This thesis integrates audio modality and text modality for emotion recognition. Experimental results demonstrate that multimodal representations can help to increase the accurate detection of emotion in conversations. However, human language prossesses not only spoken words and tone of voice but also facial attributes. Visual characteristic is one of the significant factors in emotion detection and further research of this modality in left as the remaining work.

Furthermore, modality alignment is a challenging but important process in the task of multimodal emotion recognition. However, the heterogeneities across modalities increase it's difficulty. For example, variable receiving frequency of audio and vision streams leads to different receptors, which makes it difficult to obtain optimal mapping between them. The face with a pair of frowning eyebrows may relate to a negative word spoken in the past. In our architecture, we just concatenate the acoustic and linguistic representations, with no modality alignment, which is outside the scope of this thesis, and should be further researched in the future work.

Bibliography

[1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[2] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European Semantic Web Conference*, pp. 593–607, 2018.

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

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

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

[6] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *ACL*, pp. 873–883, 2017.

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

[8] J. Zhao, S. Chen, J. Liang, and Q. Jin, "Speech emotion recognition in dyadic dialogues with attentive interaction modeling.," in *INTERSPEECH*, pp. 1671–1675, 2019.

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

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

[11] 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*, pp. 5415–5421, 2019.

[12] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[13] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[14] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, pp. 537–541, IEEE, 2010.

[15] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[17] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[18] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75798–75809, 2019.

[19] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[20] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms.," in *Interspeech*, pp. 1089–1093, 2017.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[23] G. I. Winata, A. Madotto, Z. Lin, J. Shin, Y. Xu, P. Xu, and P. Fung, "Caire_hkust at semeval-2019 task 3: Hierarchical attention for dialogue emotion classification," *arXiv preprint arXiv:1906.04041*, 2019.

[24] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] Y. Fu, L. Guo, L. Wang, Z. Liu, J. Liu, and J. Dang, "A sentiment similarity-oriented attention model with multi-task learning for text-based emotion recognition," in *International Conference on Multimedia Modeling*, pp. 278–289, Springer, 2021.

[27] O. Araque, G. Zhu, and C. A. Iglesias, "A semantic similarity-based perspective of affect lexicons for sentiment analysis," *Knowledge-Based Systems*, vol. 165, pp. 346–359, 2019.

[28] Y. Zou, T. Gui, Q. Zhang, and X.-J. Huang, "A lexicon-based supervised attention model for neural sentiment analysis," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 868–877, 2018.

[29] S. Khosla, N. Chhaya, and K. Chawla, "Aff2vec: Affect–enriched distributional word representations," *arXiv preprint arXiv:1805.07966*, 2018.

[30] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Seventh international conference on spoken language processing*, 2002.

[31] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.

[32] G. Sharma and A. Dhall, "A survey on automatic multimodal emotion recognition in the wild," in *Advances in Data Science: Methodologies and Applications*, pp. 35–64, Springer, 2021.

[33] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, 2014.

[34] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "Emotion recognition involving physiological and speech signals: a comprehensive review," *Recent advances in nonlinear dynamics and synchronization*, pp. 287–302, 2018.

[35] Y. Gu, S. Chen, and I. Marsic, "Deep mul timodal learning for emotion recognition in spoken language," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5079–5083, IEEE, 2018.

[36] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018, p. 2225, NIH Public Access, 2018.

[37] B. V. Dasarathy, *Decision fusion*, vol. 1994. IEEE Computer Society Press Los Alamitos, 1994.

[38] J. Williams, R. Comanescu, O. Radu, and L. Tian, "Dnn multimodal fusion techniques for predicting video sentiment," in *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, pp. 64–72, 2018.

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

[40] 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

[41] J. Li, M. Zhang, D. Ji, and Y. Liu, "Multi-task learning with auxiliary speaker identification for conversational emotion recognition," *arXiv e-prints*, pp. arXiv–2003, 2020.

[42] G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3946–3950, IEEE, 2019.

[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[44] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6685–6689, IEEE, 2019.

[45] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2018, p. 2122, NIH Public Access, 2018.

[46] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[47] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, 2020.

[48] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel, "Gated graph sequence neural networks.," in *ICLR*, 2016.

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

[50] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, "Structured sequence modeling with graph convolutional recurrent networks," in *International Conference on Neural Information Processing*, pp. 362–373, Springer, 2018.

[51] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, "Adversarially regularized graph autoencoder for graph embedding," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2609–2615, 2018.

[52] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 5308–5317, 2016.

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

[54] J. Zhou, J. X. Huang, Q. V. Hu, and L. He, "Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification," *Knowledge-Based Systems*, vol. 205, p. 106292, 2020.

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

[56] T. Mihaylov and A. Frank, "Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge," in *ACL*, pp. 821–832, 2018.

[57] D. Ghosal, D. Hazarika, A. Roy, N. Majumder, R. Mihalcea, and S. Poria, "KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis," in *ACL*, 2020.

[58] B. Zhang, X. Li, X. Xu, K.-C. Leung, Z. Chen, and Y. Ye, "Knowledge guided capsule attention network for aspect-based sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2538–2551, 2020.

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

[60] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

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

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

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

[64] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3677–3681, IEEE, 2013.

[65] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *AAAI*, pp. 4444–4451, 2017.

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

[67] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," *arXiv preprint arXiv:1411.4166*, 2014.

[68] 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.

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

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

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

Acknowledgement

I would like to express my sincere thanks to my supervisor, professor Okada Shogo. Professor Okada has been giving me many suggestions on my research work. Professor Okada has been always supportive of my idea and helps me refine and improve the idea to make it more feasible, which gives me a lot of motivation in my research, without his guidance, this work could not be completed. I would also like to thank Mr. Zhou Di and Mr. Wei Wenqing, who has given me a lot of help during my stay in JAIST so that I can adapt well to the life of JAIST.

At the same time, l would like to express my thanks to my supervisor at Tianjin University, professor Wang Longbiao, thank him for his guidance and support. I would also like to express my appreciation for the opinions and suggestions from my group mates, Mrs. Guo Lili, Mr. Liu Jiaxing, Mr. Gao Yuan, Mr. Song Yaodong. Especially, give my heartfelt thanks to Mrs. Guo Lili, who has not only to help me refine my idea but also help me modify my papers.

Moreover, thanks to Professor Dang Jianwu and the cooperative education program of Tianjin University and JAIST for giving me an opportunity to experience different cultures and research patterns in China and Japan. In the addition, l would like to thank my parents and friends, who have given me much support in my life and study.

Publications

- Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu and Jianwu Dang, "ConSK-GCN: Conversational Semantic- and Knowledge-guided Graph Convolutional Network for Multimodal Emotion Recognition", IEEE International Conference on Multimedia and Expo, 2021. (peer-reviewed, accepted)
- Yahui Fu, Lili Guo, Longbiao Wang, Zhilei Liu, Jiaxing Liu and Jianwu Dang, "A Sentiment Similarity-Oriented Attention Model with Multi-task Learning for Text-Based Emotion Recognition." International Conference on Multimedia Modeling. Springer, Cham, 2021: 278-289.

Patent

- Longbiao Wang, Yahui Fu, Jianwu Dang, and Lili Guo, "A Method for Textual Emotion Recognition based on Sentiment Similarity-oriented Attention," November 2020. Tianjin University, Chinese patent, Patent No. 202010665789.8, pending