基于上下文无关和上下文相关的情感识别研究

A Study on Context-independent and Context-dependent Emotion Recognition

专业类别：工程

研究方向（领域）：计算机技术

作者姓名：傅雅慧

指导教师：王龙标　教授

企业导师：王林　高级工程师

答辩日期	2021 年 5 月 10 日		
答辩委员会	姓名	职称	工作单位
主席	张加万	教授	天津大学智能与计算学部
委员	路文焕	教授	天津大学智能与计算学部
	孙提	高级工程师	浪潮通用软件有限公司

天津大学智能与计算学部

二〇二一年五月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **天津大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：傅雅慧　签字日期：　2021 年 5 月 5 日

学位论文版权使用授权书

本学位论文作者完全了解 **天津大学** 有关保留、使用学位论文的规定。特授权 **天津大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：傅雅慧　　导师签名：王龙标

签字日期：　2021 年 5 月 5 日　　签字日期：　2021 年 5 月 5 日

摘　要

情感是人类固有的，因此，情感识别是机器在自然的人机交互中理解和产生情感反应的关键部分。对话中的情感识别近年来受到了广泛的关注，并且由于其在社会媒体，医疗保健，教育和人工智能交互等多个领域的广泛应用，已成为自然语言处理研究的新领域。因此，有效的情感识别算法具有重要意义，但是这仍然是一个具有挑战性的任务。基于上下文无关和上下文相关的情感识别是情感计算中的两个重要任务。对于第一个任务，近年的研究主要集中于从单句的文本，音频和对应的面部表情等提取情感特征，并不考虑上下文的影响。但考虑到多轮对话中语义和情感变化的复杂性，仅对孤立的单句建模不足以准确的预测文本的情感，因此基于上下文的情感识别（任务二）也是目前研究的热点之一。与第一个任务中的上下文无关的文本情感识别不同，为第二个任务建模有效的上下文关系尤其具有挑战性。对第一个任务，预训练的词嵌入模型在上下文无关的情感分析系统中起着重要作用，如可以提取单词多维语义特征的Word2vec, Glove(Global Vectors)等。除了词义之外,也有很多研究基于端到端的方法，通过在经过良好训练的神经网络模型如ELMO(Embeddings from Language Models)和BERT(Bidirectional Encoder Representations from Transformers)上进行微调来获得更好的性能。但是，这些特征表示是基于句法和语义信息的，它们不包含特定的情感信息。为了在训练过程中提取有效的情感特征，许多工作引入了词典信息，因为每个情感值往往代表对应单词的情感极性和情感强度，比如有的研究将文本分词后所有的单词和情感词典中的每个词进行语义相似度计算，并取最大值作为该文本在此单词维度下的情感极性。情感词典的使用可以一定程度上增强训练过程中的情感特征，但是其作用往往受限于情感词典的大小，此外在表达一个句子的情感极性时，将句子中的每个单词的情感极性相加或是取最大值的做法是不太恰当的，比如当两个正反极性的单词组合在一起，其情感极性不一定是两者相加或取最大值。注意力机制模型，即给情感特征更大的权重，这种方法在文本情感识别中也有较多应用。但是这些方法主要是基于语义和句法信息对上下文进行建模从而给不同的特征分配权重，并没有包含特定的情感信息。第一个任务存在的问题是先前的研究利用语义信息和情感词典进行情感建模来提取的情感特征。但是忽略了可能由情感标签本身传达的信息。本文认为，情感标签的语义特征可以引导网络从输入文本中提取与情感相关的特征。为了解

决这个问题，本文提出了一种基于情感相似度的注意力（SSOA）机制，该机制在编码文本特征时使用情感标签的语义信息来指导模型的注意力，从而从句子中提取与情感有关的特征。本文的方法主要包括三个部分：1）句向量编码；2）以情感相似度为导向的注意力机制计算；3）多任务学习。1）句向量编码。通过基于Transformer结构的Universal Sentence Encoder预训练模型将训练集和四类情感转换为句向量矩阵。本文并不是采取随机初始化的方法进行标签特征的学习，而是通过Transformer, 因为从预训练模型中可以获得更丰富的语义特征。此外本文使用句特征向量而不是传统的词特征向量，因为当基于词级别计算一个句子的情感时可能会导致一句话中前后情感不一致。例如，在一个句子样本中，“你不愚蠢。”单词“不”和“愚蠢”都代表否定情绪，如果只是将它们串联起来代表这句话的情绪，那就是负面情绪的，但实际上这句话表达的是积极情绪。2）以情感相似度为导向的注意力机制计算。通过所提出的以情感相似度为导向的注意力机制，计算文本和情感标签之间的情感相似度，并且将情感词典中的效价值（V）作为情感极性，来表示不同单词的情感从不开心到愉悦的情感强度。训练模型后，可以获得SSOA，其值表示每种情感的权重。接着使用convolutional neural network(CNN)提取复杂的语言特征，因为它已被广泛用于文本情感识别并显示了不错的效果。3）多任务学习。情感一般的有两种常见的表示方式，一种是离散的情感如开心，伤心，沮丧等，另一种是三维区间的情绪表示，效价（V，表示刺激的愉悦性），唤醒度（A，表示情感强度）和优势度（D，刺激控制的程度）。V，A，D可以更加细腻度的描述情感。Marsella等人提出每一种离散的情感都可以由V,A,D三个维度线性表示。并且在心理学和生理学中，情绪的表达也与V，A，D高度相关。多任务学习已被广泛使用，在一项任务中学习的知识通常可以提高另一项相关任务的性能并丰富不同类型任务的鲁棒性。综上原因，在这项工作中，通过构建情感分类和VAD情绪回归的多任务学习来提取更加鲁棒的特征。此次实验采用IEMOCAP数据库。这是个多模态数据库，内容为10名不同的说话人在自发和剧本两种不同的语境下的对话视频，包括转录文本，音频和面部动作等特征。此次实验中本文只使用文本模态。为了与前沿的实验进行对比，本文使用离散的5531句，共有四类情感，分别为开心（29.6%），中性（30.9%），生气（19.9%）和伤心（19.6%）；有三类情绪维度标签，分别为效价、唤醒度、优势度，每类标签的情绪维度为1到5。此外，本文使用ANEW情感词典，该词典在效价，唤醒度和优势度三个维度的取值范围分别为1至9，对应每个维度情绪从弱到强的程度。结果表明，在IEMOCAP数据库上，所提出的方法比以前的基于文本的情感识别性能提高了6.57%，从63.39%增至69.96%，并且具有更好的鲁棒性。对于VAD回归任务，本文使用均方误差（MSE）和皮尔逊相关系数（r）来评估性能，其中MSE值越低且相关性越高，则性能越好。实

验结果表明所提出的模型在VAD回归任务上也优于前人的工作取得了更好的效果。与第一个任务中的上下文无关的情感识别不同，在第二个上下文相关的任务中构建有效的上下文关系对更准确的检测说话人的情感是至关重要的。基于recurrent neural network(RNN)的方法在上下文建模中应用广泛，它是应用双向长短期记忆模型按照序列顺序编码对话的上下文特征。但是，这种方法面临远程传播信息丢失的问题，因此对长距离上下文信编码效果不佳。为了缓解此问题，一些变体将双向长短期记忆模型和注意力机制融合在一起，可以动态地关注最相关的上下文特征。但是，注意力机制并未考虑目标话语和上下文话语的相对位置，这对于建模过去的对话如何影响未来对话非常重要，反之亦然。因此许多工作如dialogue graph convolutional network (DialogueGCN) 和graph-based convolutional neural network towards conversations (ConGCN)使用图卷积神经网络（GCN）来建模上下文相关性，并且都取得了很好的效果，证明了GCN在上下文结构上的有效性。由于目标话语的情绪通常会受到附近话语的强烈影响，因此通过对GCN模型中的边的构建可以有效表示说话者之间的相互影响关系和自我影响关系。但是，DialogueGCN和ConGCN都仅考虑对话之间的语义信息，对于不包含明显的情感术语的隐含情感文本，由于文本中的单词相对客观和中立，如果仅考虑对话的语义信息，很难正确地识别情感。在第二个任务中，对话者的内部依赖和相互依赖对于建模动态交互并理解每轮对话的情感变化非常重要。图神经网络由于其丰富的关系结构已在多种任务上显示出有效的性能，并且可以在图编码中保留图的全局结构信息。GCN基于邻域的结构是一种合适的体系结构，可提取对话的局部和全局的上下文信息。但是上下文语义所传达的信息不足以进行情感检测，尤其是对于小型数据库和隐式情感文本而言。知识库的应用已经在诸如开放域对话系统和抑郁症检测等多个研究领域中引起了广泛的关注。知识库提供了与常识相关的背景概念等丰富的资源，可通过提供上下文特定的概念来增强对话的语义特征和情感极性。所以，基于上下文的语义建模和引入外部知识库来增强语义的理解对于上下文的情感识别任务是至关重要的。为了解决这个问题，本文提出了一种新的基于语义和知识引导的多模态图卷积神经网络（ConSK-GCN）方法，以有效地构造每段对话中的语义相关和知识相关的上下文特征。本文的方法主要包括三个部分：1）多模态特征的提取和初始化；2）基于语义图的知识提取；3）上下文语义和知识引导的图卷积神经网络(ConSK-GCN)的构建和训练。1）多模态特征的提取和初始化。在对话中，人与人交流时内容和韵律都会传达情感，因此本文同时使用声音和文字表征两种模态的情感识别。为了初始化每种模态，本文训练了不同的网络在情感标签的监督下分别提取单句级别的语言和声学特征。由于IEMOCAP和MELD数据库的差异性，针对这两个数据库，本文采取了不同的方法。对于IEMOCAP数据库，为

了与最先进的方法进行比较，本文采用了传统且使用最广泛的卷积神经网络来提取文本特征。首先，本文使用公开可用的预训练模型word2vec来初始化单词向量。然后，使用一个卷积层，一个池化层和两个全连接层来获得句级别的深层特征表示。其中100个尺寸分别为3，4，5的滤波器做卷积操作。池化层的窗口大小设置为2，激活函数是Relu。最后输入两个分别具有500和100个隐藏节点的全连接层。对于声学特征的提取，研究人员发现，大于250ms的段语音信号包含足够的情绪信息，因此本文设置每段的时长为265ms，滑动窗口设置为25ms，则频谱图尺寸为32 × 129。在MELD数据库中，每段对话的平均句数和每句话的平均单词数分别为9.6和8.0，其中在IEMOCAP数据库中分别为49.2和15.8。相比于IEMOCAP，MELD中的话语较短，上下文相关性不强。因此，卷积神经网络不足以提取MELD中话语的有效特征。考虑到BERT_BASE在许多NLP任务（例如阅读理解，抽象性摘要，文本蕴含和学习与任务无关的句子表示）中显示了最先进的性能，因此本文应用BERT_BASE，其模型架构是多层双向Transformer编码器，用于初始化MELD的文本表示。首先，本文对预训练的BERT_BASE模型进行微调，其中包含12个Transformer块，768个隐藏大小，12个自注意头以及110M总参数。然后，本文将倒数第二个全连接层的特征作为上下文无关的句级别的特征向量。声学特征的提取方面，由于MELD的平均句长为3.6s, 因此本文设置段长为2s,滑动窗长为1s,每段的频谱图尺寸为1874 × 129。最后本文使用两层BLSTM对文本特征和语音特征进行融合。模态对齐是多模态情感识别任务中具有挑战性但重要的过程。但是，跨模态的异质性增加了它的难度。在此次论文的体系结构中，本文只是将声学和语言特征拼接在一起，没有模态对齐，这也将在以后的工作中做进一步研究。2）基于语义图的知识提取。在本文中，主要使用了常识知识库ConceptNet和情感词典NRC_VAD。ConceptNet是一种大规模的多语言语义图，通过带有标记的加权边将自然语言的单词和短语连接起来，旨在帮助理解语句中所涉及的常识，从而改善自然语言的应用，协助自然语言应用程序更好地理解人们使用的词语背后的含义。ConceptNet中，节点代表概念，边代表关系，每组<concept1，relation，concept2>都有对应的置信度得分。例如："奖学金具有同义词助学金，置信度分数为0.741"。对于英语，ConceptNet包括590万组连接，310万个概念和38种关系。然后，本文根据每个语义图中的语义依赖在ConceptNet中选择相应的概念。NRC_VAD词典中包含超过20,000个的英语单词，对应其效价（V），唤醒度（A）和优势度（D）分数。每个维度的VAD实值得分分别在0-1的范围内，对应于从低到高的程度。其中本文计算V和A的值作为每一个知识概念的情感极性。3）上下文语义和知识引导的图卷积神经网络(ConSK-GCN)的构建和训练。本文分别构建了三个图网络模型，分别是基于上下文语义的图卷积神经网络（S-GCN)，基于知识的卷积神经

网络（K-GCN），和基于上下文语义和知识的图卷积神经网络（SK-GCN）。在上下文语义引导的图卷积神经网络（ConS-GCN）中，本文通过构建图模型描述了对话者之间的上下文交互信息和说话人自己的语义连贯性。在这个基于上下文建模的语义图中，每句话都可以看作是单个节点，一对节点/对话之间的关系边则表示这些对话的说话者之间的依赖关系。语义图中，每个节点代表每句话的多模态特征。边表示每段对话中的上文语义相似度。本文首先计算两句话的余弦相似度，然后利用arccos将余弦相似度转换为角距离，从而来计算两句话之间的语义相似度。在知识图网络（ConK-GCN）中,本文引入了一个外部知识库，该知识库可以帮助理解对话内容和生成适当的回答，并通过构建知识指导的图卷积神经网络来丰富上下文中每个概念的语义含义。此外，本文将情感词典引入知识图的构建中，以丰富每个知识的情感极性。知识图中，每个节点/概念特征可以通过有效的语义空间ConceptNet Numberbatch获取,该语义特征是从分布式语义如word2vec和ConceptNet中学习而来。不包含在ConceptNet中的概念通过“fastText”方法进行初始化，该方法是用于有效学习单词表示的库。对于不在NRC_VAD中的概念，本文将V和A的值设置为中性值0.5。知识图中的边表示不同概念之间的知识关联性。最后，本文利用语义权重矩阵和情感增强的知识权重矩阵来构建ConSK-GCN的新邻接矩阵，以在上下文情感识别任务中获得更好的性能。语义和知识引导的图中，与语义图相同，每个节点代表每句话的多模态特征。边矩阵是对知识图和语义图的边矩阵加权求和，并同过模型参数w_k，用于平衡知识和语义对每段对话中上下文相关性的影响。然后本文将多模态特征和边矩阵输入到R-GCN中得到既具有上下文语义又具有知识的局部上下文信息。为了在ConSK-GCN的训练中找到知识权重和语义权重之间的最佳平衡值w_k，本文分别在IEMOCAP和MELD中测试w_k从0，0.1，···,1不同取值的效果。结果显示知识感知和语义感知的语境构建对于会话中的情感识别非常重要，但是不同的权重（0.1到0.9）对情绪检测的影响并不明显。本文在两个多模态对话数据库上评估提出的ConSK-GCN模型，分别是IEMOCAP和MELD。本文只使用了语音和文本模态用于情感识别。然而，人类交互中，除了语调和说话内容，面部表情也可表示情感的变化，因此视觉特征也是情感检测中的重要因素之一，对于融合这三种模态的多模态情感识别将作为未来的工作之一。MELD数据库共有1433段对话，共约13000句，其情感分布为46.95％中立，16.84％欢乐，11.72％愤怒，11.94％惊喜，7.31％悲伤，2.63％厌恶，和2.61％害怕。在多模态语料库IEMOCAP和MELD上的实验表明，本文的方法可以有效地构建对话中的上下文相关性。特别是对于包含隐性情感的文本，可以有效的提高情感识别的准确率。具体而言，在IEMOCAP上进行的实验表明，本文的方法在单模态和多模态情感识别方面均优于目前最新的方法。单模态下，在平均精度和F1两个衡量指

标方面都至少提高了1.3%，而在多模态情感识别中提高了4%以上。在MELD数据库上进行的实验表明，所提出的ConSK-GCN在单模态和多模态情感识别方面，在F1的指标上，皆具有优于最新方法至少5.7%的性能。此论文主要针对两种不同的任务中存在的问题提出了两种解决方案。在上下文无关的情感识别任务中，为了解决语义和情感词典不足以提取有效的情感特征进行情感建模的问题，本文提出了一种面向情感相似度的注意力机制，该机制可用于指导网络从输入文本中提取与情感有关的信息，以提高上下文无关的情感分类任务的准确度和减小了情感回归任务的误差。在上下文相关的任务中，为了解决上下文语义所传达的信息不足以对小型数据库和隐式情感文本进行情感检测的问题，本文提出了一种新的基于语义和知识的上下文图卷积网络（ConSK-GCN）用于上下文相关的情感识别，并且有效的运用了文本和音频两种模态。在这种方法中，本文通过基于对话图的图卷积网络来构造说话者之间和说话者自己的上下文交互。然后将语义图和常识知识图结合起来，对语义相关和知识相关的上下文动态进行建模。本文所提出的以情感相似度为导向的注意力模型在IEMOCAP数据库上，相比以前的工作情感识别准确率从63.39%增至69.96%，提高了6.57%；在VAD回归任务上也优于前人的工作取得了更低的均方误差和更高的皮尔逊相关性。验证了所提出的模型在情感识别任务中能够提取更有效和更具鲁棒性的情感特征。本文所提出的语义和知识引导的图卷积神经网络（ConSK-GCN），在IEMOCAP和MELD数据库上，在单模态和多模态情感识别任务中皆优于前人的工作取得了更好的结果。其中在IEMOCAP上进行的实验表明，ConSK-GCN优于目前最新的方法,单模态下，平均精度和F1值都至少提高了1.3%，而在多模态情感识别中也提高了4%以上。在MELD数据库上，ConSK-GCN在单模态和多模态情感识别方面，优于最新方法F1值分别提高5.7%和7.3%。这验证了上下文语义信息和外部知识的引入对正确检测对话中的情感的必要性以及ConSK-GCN的有效性。总结来说，与现有研究相比，本文的贡献是：1）本文提出了一种面向情感相似度的注意力机制，通过结合标签语义来对情感信息表示进行编码。2）本文结合训练两项相关任务（即离散的情感分类和维度的VAD回归）之间的相互学习来改善彼此的表现。3）本文提出了一种新的面向上下文语义和知识的图卷积网络（ConSK-GCN）方法，该方法同时利用了语义信息，常识知识和多模态（文本和语音）特征。知识库的引入丰富了每段对话的语义，而情感词典则增强了对话中每个概念的情感极性。此外，这两项技术可以作为人机交互系统的重要组成部分，应用到增强情感互动并改善用户体验等相关任务中。

关键词： 情感特征增强，多任务学习，图卷积神经网络，知识，多模态情感识别

ABSTRACT

Emotion recognition 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 emotion recognition algorithms are of great significance.

Emotion recognition based on context-dependence and context-independence are two major tasks in the community. For the first task, previous studies utilize contextual semantics and emotion lexicon for affect modeling. However, they ignore information that may be conveyed by the emotion labels themselves. Different from the non-conversational text in the first task, it is particularly challenging to model the effective context-aware dependence for the second task. It is difficult 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.

To address the problem in the first task, we propose the sentiment similarity-oriented attention (SSOA) mechanism, which uses the semantics of emotion labels to guide the model's attention when encoding the input conversations. Thus to extract emotion-related information from sentences. Then we use the convolutional neural network (CNN) to extract complex informative features. In addition, as discrete emotions are highly related with the Valence, Arousal, and Dominance (VAD) in psychophysiology, we train the VAD regression and emotion classification tasks together by using multi-task learning to extract more robust features. The proposed method outperforms the benchmarks by an absolute increase of over 3.65% in terms of the average F1 for the emotion classification task, and also outperforms previous strategies for the VAD regression task on the IEMOCAP database.

To further the problems in the second task, 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. 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. On the one hand, we construct 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. Furthermore, we leverage the semantic edge weights and affect enriched knowledge edge weights to construct a new adjacency matrix of our ConSK-GCN for better performance in the ERC task. 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. Experiments on IEMOCAP illustrate that our proposed model performs better than all of the baseline approaches, with an improvement of at least 1.3% in terms of average accuracy and F1 with unimodality and more than 4% with multimodality. Experiments on MELD show that the proposed ConSK-GCN has a better performance with more than 5.7% than the state-of-the-art approaches in terms of F1 in both unimodal and multimodal emotion recognition, illustrating that our methodology could effectively construct the contextual dependence of the utterances in a conversation.

KEY WORDS: Emotion Feature Enhancement, Multi-task Learning, Graph Convolutional Neural Network, Knowledge, Multimodal Emotion Recognition

目　　录

摘　　要 I

ABSTRACT VII

CHAPTER 1　Introduction 1

1.1　Research background 1

1.2　Problem statement 2

1.3　Research motivation 3

1.4　Research contents and contributions 4

1.5　Thesis organization 5

CHAPTER 2　Literature Review 7

2.1　Multimodal emotion recognition 7

2.1.1　Acoustic modality 7

2.1.2　Text modality 8

2.1.3　Multi modality 9

2.2　Emotion recognition in conversations 9

2.2.1　Variables in conversations 10

2.2.2　Conversational context modeling 11

2.3　Graph convolutional neural network 13

2.4　Knowledge base in emotion recognition 16

CHAPTER 3　Sentiment Similarity-oriented Attention Model for Context-independent Emotion Recognition 19

3.1　SSOA mechanism with multi-task learning 19

3.1.1　Sentence encoder 19

3.1.2　Sentiment similarity-oriented attention 19

3.1.3　Multi-task learning 21

3.2　Experiments and analysis 22

3.2.1　Database and lexicon 22

3.2.2　Experimental setup 22

3.2.3　Experimental results and analysis 23

3.3　Conclusion 27

CHAPTER 4 Semantic- and Knowledge-guided Graph Convolutional Network for Context-dependent Emotion Recognition 29
4.1 ConSK-GCN model 30
4.1.1 Database preparation 30
4.1.2 Multimodal features extraction 30
4.1.3 Knowledge retrieval 33
4.1.4 ConSK-GCN construction 34
4.2 Experiments and analysis 37
4.2.1 Databses 37
4.2.2 Experimental setup 38
4.2.3 Comparison methods 38
4.2.4 Experimental results and analysis 39
4.2.5 Effect of Context Window 43
4.2.6 Case study 44
4.2.7 Effect of w_k 44
4.3 Conclusion 46
CHAPTER 5 Conclusions and Future Works 47
5.1 Summary 47
5.2 Contributions 47
5.3 Future works 48
REFERENCES 49
发表论文和参加科研情况说明 57
致　谢 59

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)[1], convolutional neural network (CNN)[2], and recurrent neural network (RNN)[3] are the most commonly used architectures for emotion recognition and usually achieve competitive results.

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[4] 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[5]. To mitigate this issue, some variants like AIM[6] and DialogueRNN[7] 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 the target and context utterances, which is important for modeling how past utterances influence future utterances and vice versa. DialogueGCN[8] and ConGCN[9] employ a graph convolutional neural network (GCN) to model the contextual dependence and all achieve a new state of the art, proving the effectiveness of the GCN on context structure. 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

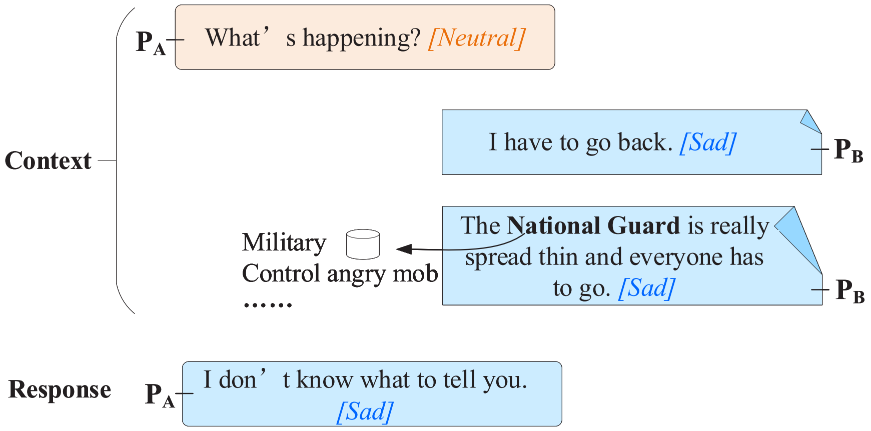

Context
PA
What' s happening? [Neutral]
I have to go back. [Sad]
PB
Military
Control angry mob
......
The National Guard is really spread thin and everyone has to go. [Sad]
PB
Response
PA
I don' t know what to tell you. [Sad]

Figure 1-1 An example conversation with annotated labels from the IEMOCAP dataset.

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

There are two major tasks in the emotion recognition community, one is the context-independent emotion recognition and another is context-dependent emotion recognition. It is significant to extract effective emotional features for emotion recognition but still a challenging task.

In the traditional studies for context-independent emotion recognition, distributed representations or pre-trained embeddings are playing important roles in state-of-the-art sentiment analysis systems. For example, predictive methods Word2Vec[10] and Glove[11], which can capture multi-dimensional word semantics. Beyond word-semantics, there has been a big efforts toward End-to-End neural network models[12] and achieved better performance by fine-tuning the well pre-trained models such as ELMO[13] and BERT[14]. However, these representations are based on syntactic and semantic information, which do not enclose specific affective information.

In the task of context-dependent emotion recognition, 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

For the first issue, previous studies utilize semantics and emotion lexicon for affect modeling but ignore information that may be conveyed by the emotion labels themselves. The key idea centers on the fact that the label embeddings can guide the network to extract emotion-related information from input sentences.

For the second issue, 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 both inter-interaction and self-inertial of the speakers. 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 contents and contributions

The objective of this thesis is to propose a sentiment similarity-oriented attention mechanism and a new semantic- and knowledge-aware graph convolutional neural network for emotion recognition.

To address the first problem, we propose the sentiment similarity-oriented attention (SSOA) mechanism, which uses the semantics of emotion labels to guide the model's attention when encoding the input conversations. Thus to extract emotion-related information from sentences. Then we use the convolutional neural network (CNN) to extract complex informative features. In addition, as discrete emotions are highly related with the Valence, Arousal, and Dominance (VAD) in psychophysiology, we train the VAD regression and emotion classification tasks together by using multi-task learning to extract more robust features.

To further tackle the second problem, 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 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.

This thesis proposed a a sentiment similarity-oriented attention mechanism and a new semantic- and knowledge-aware graph convolutional neural network for emotion recognition. 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, both two technologies can be applied as an important part of the human-robot system to enhance emotional interaction and improve user experience.

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 extracting effective emotion features for a better performance in emotion recognition. 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 introduce our proposed method for context-independent emotion detection in detail, which is sentiment similarity-oriented attention model with multi-task learning for text-based emotion recognition.

Chapter 4:

In this chapter, we make detailed description about our proposed method for context-dependent emotion detection, which is conversational semantic- and knowledge-guided graph convolutional network for multimodal emotion recognition.

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.

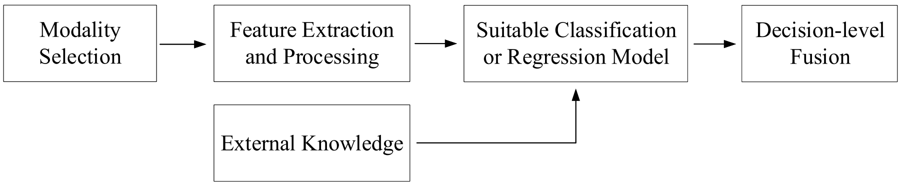

Modality Selection
Feature Extraction and Processing
Suitable Classification or Regression Model
Decision-level Fusion
External Knowledge

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.

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[15] and it has applications in many applications, such as human-computer interaction[16], robots[17], psychological assessment[18] 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[19] tool with 384 dimensions proposed in[20]. 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[21]. 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[22], and then applied the Bidirectional Long Short-Term Memory (BLSTM) to recognize emotions. The CNN-BLSTM model[21,23] 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[10] and Glove[11], which can capture multi-dimensional word semantics. Beyond word-semantics, there has been a big efforts toward End-to-End neural network models[12] and achieved better performance by fine-tuning the well pre-trained models such as ELMO[13] and BERT[14].

Table 2-1 Auditory-based feature set

LLDs (162)	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

To enrich the affective information into training,[1,24–27] 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.[26] proposed a lexicon-based supervised attention model to extract sentiment-enriched features for document-level emotion classification. Similarly,[27] 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.

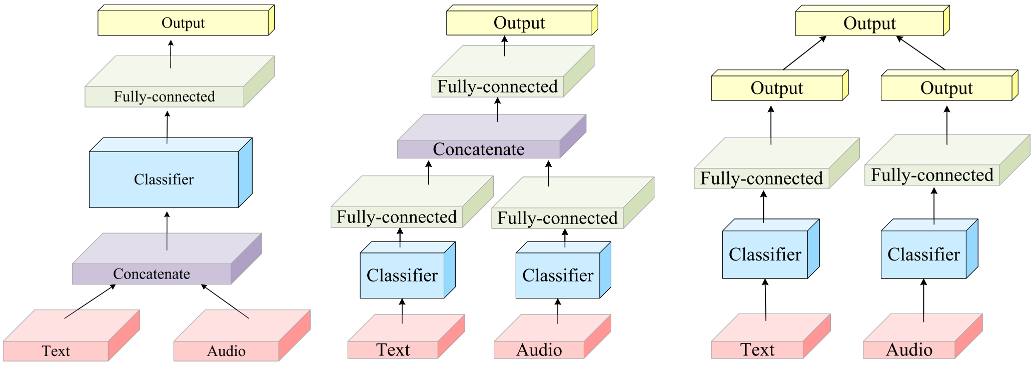
2.1.3 Multi modality

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[28]. A multimodal structure is thus necessary for using both the text and audio as input data[29]. The current research such as[1,3,30] 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[31] 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[31]. 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[32]. 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.

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[4,7–9]. ERC can be used to analyze conversations that happen on social media to mine emotion and



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

Figure 2-2 Main fusion strategies multimodaltiy.

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[33].

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[33]. 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, Li et al[34] 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[35]. 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 spectrogram to extract deep acoustic features. To capture contextual information from target utterances' surroundings in the same video, Poria et al[4] proposed a LSTM-based

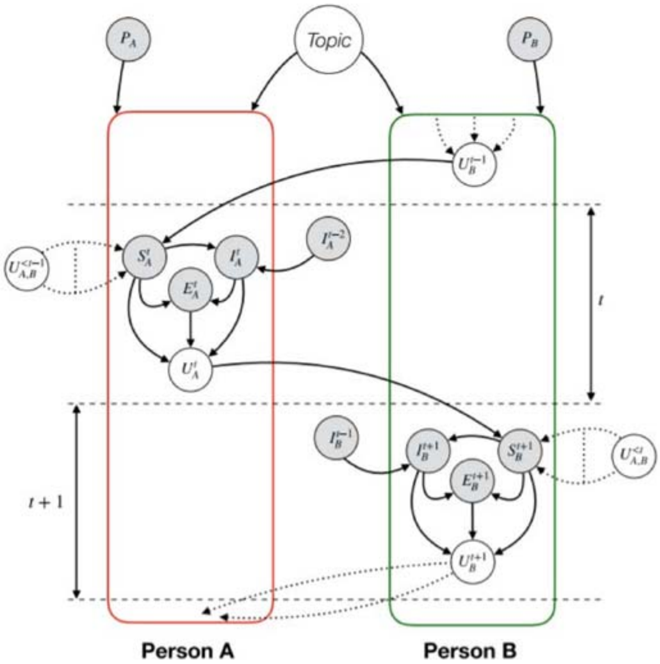


Figure 2-3 Interaction among different controlling variables during a dyadic conversation between speakers[33].

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. Majumder et al[7] applied three gated recurrent units (GRU)[36] to track the update of global context ,emotion and speaker state respectively. Yeh et al[37] proposed a new interaction-aware attention network (IAAN) that integrated contextual information in the learned acoustic representation through an attention mechanism. Hazarika et al[38] 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[4,7,35,38] 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*[39]. Self-dependency can be also understood as *emotional inertia*[33], 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.

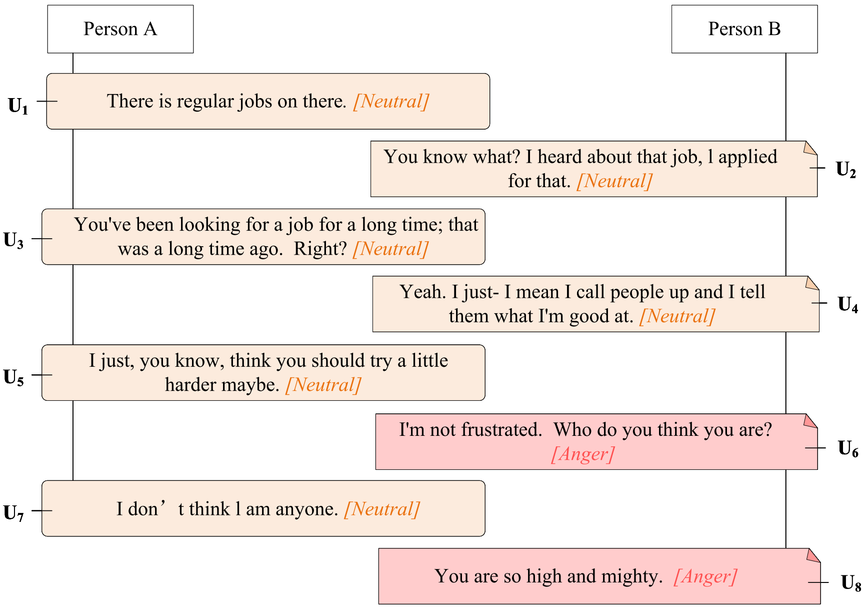

Person A
Person B
U1
There is regular jobs on there. [Neutral]
You know what? I heard about that job, I applied for that. [Neutral]
U2
U3
You've been looking for a job for a long time; that was a long time ago. Right? [Neutral]
Yeah. I just- I mean I call people up and I tell them what I'm good at. [Neutral]
U4
U5
I just, you know, think you should try a little harder maybe. [Neutral]
I'm not frustrated. Who do you think you are? [Anger]
U6
U7
I don' t think I am anyone. [Neutral]
You are so high and mighty. [Anger]
U8

Figure 2-4 An example conversation from the IEMOCAP dataset

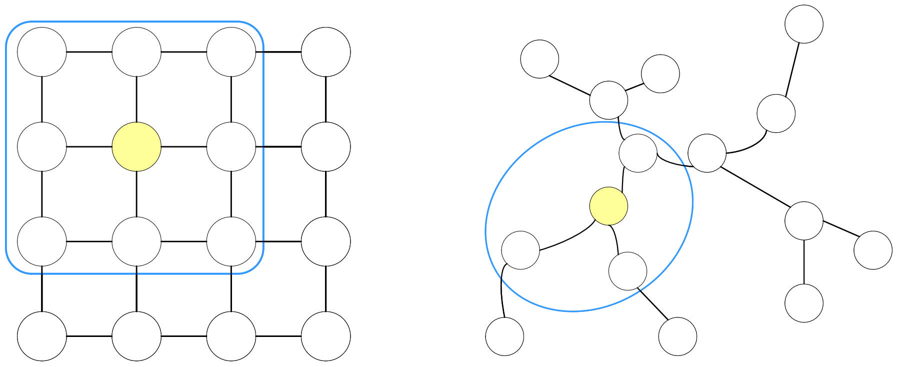
We assume that the surrounding utterances affect most for the target response, however, not only the contextual information from the local but also the distant conversational history are important for context modeling, especially in the situation that speaker refer to the topic and information from the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

RNN-based methods such as bc-LSTM[4] 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[5]. To mitigate this issue, some variants like DialogueRNN[7] 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 the target and context utterances, which is important for modeling how past utterances influence future utterances and vice versa. DialogueGCN[8] and ConGCN[9] employ a graph convolutional neural network (GCN) to model the contextual dependence and all achieve a new state of the art, proving the effectiveness of the GCN on context structure. However, both DialogueGCN and ConGCN only consider the semantic information between utterances. Thus, for implicit emotional texts that do not contain obvious emotional terms, and the words are relatively objective and neutral, it is difficult to correctly distinguish the emotions if only the semantics of the utterances are considered. Both semantic context and commonsense knowledge are essential for the machine to analyze emotion in conversations. Figure 1.1 shows an example demonstrating the importance of context and knowledge in the detection of the accurate emotion of implicit emotional texts. In the literature, only a limited number of studies have explored the incorporation of context and commonsense knowledge via GCN for the ERC task. In the next section 2.3 and 2.5, we will briefly introduce the review of graph convolutional network and knowledge base in conversational emotion recognition.

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[2] 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[40]. 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.

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



(a) 2-D convolution

(b) Graph convolution

Figure 2-5 Euclidean Structure versus Non-Euclidean Structure.

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[41].

There are several variances in graph neural networks, such as Recurrent GNNs (RecGNN)[42], Convolutional GNNs (ConvGNNs)[43], Convolutional recurrent GNNs (GCRN)[44], Graph Autoencoders (GAEs)[45], and Spatial-Temporal GNNs (STGNNs)[46]. 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[43] 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[41]. 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[43]:

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

where $\tilde{A} = A + I_N$ is the adjacency matrix of the undirected graph with added selfconnections. 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 $ReLU(\cdot) = max(0, \cdot)$. $H^{(l)} \in \mathbb{R}^{N \times D}$ 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[47], aspect-level sentiment classification[48], emotion recognition in conversations[9], and have achieved competitive performance, where GCN is used to encode the syntactic structure of sentences.

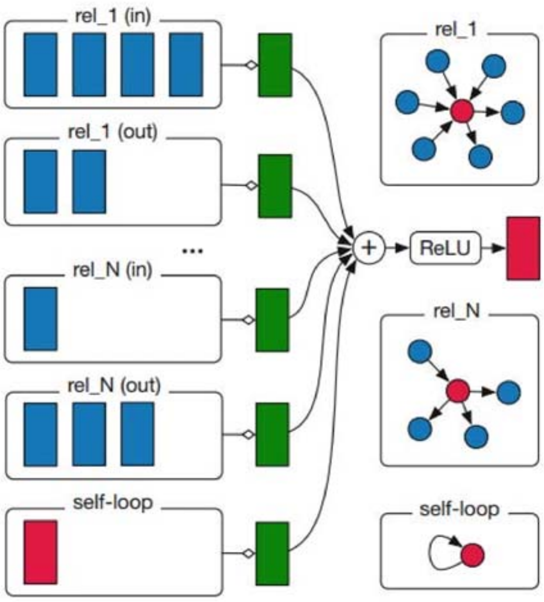


Figure 2-6 Diagram for computing the update of a single graph node/entity (red) in the R-GCN model proposed in[49].

Inspired by GCN which operates on local graph neighborhoods, Schlichtkrull et al[49] 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, 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:

$$h_i^{(1+1)} = \sigma(\sum_{r \in \Re} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)}) \tag{2-2}$$

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 $V_b^{(l)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ is the basis transformation, coefficient $a_{rb}^{(l)}$ depends on r. And $Q_{br}^{(l)} \in \mathbb{R}^{(d^{(l+1)}/B) \times (d^{(l)}/B)}$. 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[49]. 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[50], question answering systems[51], cross-domain sentiment analysis[52], aspect-based sentiment analysis[53], and emotion detection in conversations[54]. 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. Zhang et al[53] 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. Zhong et al[54] 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 Sentiment Similarity-oriented Attention Model for Context-independent Emotion Recognition

3.1 SSOA mechanism with multi-task learning

Figure 3.1 gives the overall framework. First, the sentence encoder approach is used to generate representations for all the input texts and emotional labels. Then we adopt the proposed sentiment similarity-oriented attention mechanism to obtain the sentiment-enriched text representations, followed by a CNN to extract deep informative features. In addition, we introduce multi-task learning for both emotion classification and VAD regression to extract more robust representations.

3.1.1 Sentence encoder

Cer et al[55] has published two kinds of universal sentence encoder for sentence embedding, one is trained with Transformer encoder[56], while the other is based on deep averaging network (DAN) architecture[57], and all of them can be obtained from the TF Hub website. We use the first one (USE_T) for our sentence encoder part to encode texts and emotion labels into sentence embeddings. Rather than learning label embeddings from radome, we also explore using contextual embeddings from transformer-based models. This allow us to use richer semantics derived from pre-training. The reason that we use sentence embeddings not conventional pre-trained word embeddings as when computing emotion of one sentence based on word level may cause sentiment inconsistency. For example, in a sentence sample *'You are not stupid.'* word *not* and *stupid* are both represent negative emotion, if just concatenate them to represent the emotion of this sentence, it is negative, which should be positive.

3.1.2 Sentiment similarity-oriented attention

In this section, we introduce our proposed SSOA mechanism more explicitly. The main idea behind the SSOA mechanism is to compute affective attention scores between the labels and the representations of the input sentences that is to be classified. Formally, let $S = \{s_1...s_i...s_N\}$ be the set of the sentences in the database, where N is the total number of training data set. $E = \{e_1, e_2, e_3, e_4\}$ be the set of four emotion labels (Happy,

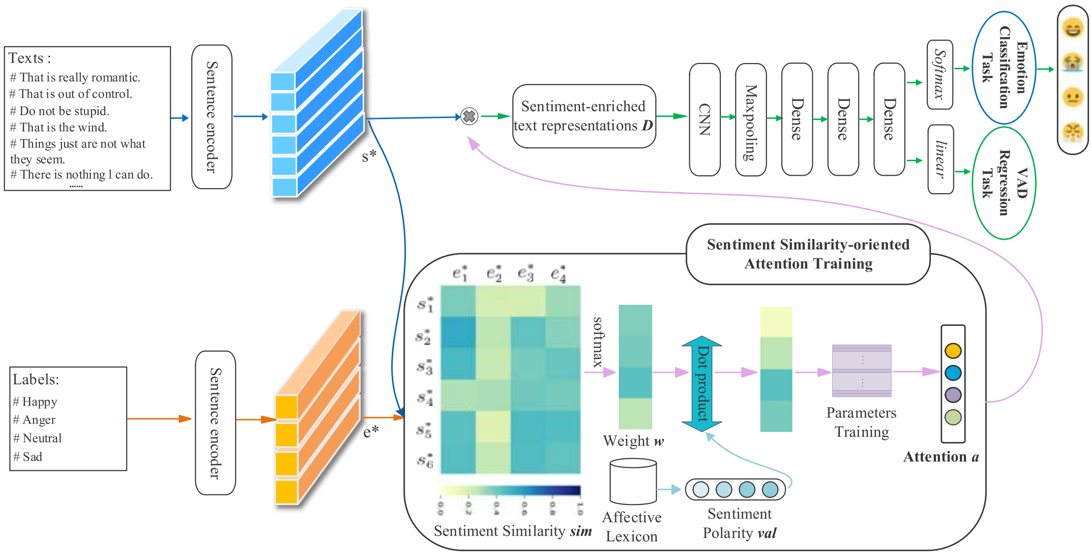

Texts :
# That is really romantic.
# That is out of control.
# Do not be stupid.
# That is the wind.
# Things just are not what they seem.
# There is nothing I can do.
......
Sentence encoder
s*
Sentiment-enriched text representations *D*
CNN
Maxpooling
Dense
Dense
Dense
Softmax
linear
Emotion Classification Task
VAD Regression Task
Labels:
# Happy
# Anger
# Neutral
# Sad
Sentence encoder
e*
Sentiment Similarity-oriented Attention Training
e_1^* e_2^* e_3^* e_4^*
s_1^* s_2^* s_3^* s_4^* s_5^* s_6^*
0.0 0.2 0.4 0.6 0.8 1.0
Sentiment Similarity *sim*
softmax
Weight *w*
Dot product
Affective Lexicon
Sentiment Polarity *val*
Parameters Training
Attention *a*

Figure 3-1 Overall framework of proposed methodology: Sentiment similarity-oriented attention model with multi-task learning.

Angry, Neutral, Sad) same as in[58], $Val = \{val_1, val_2, val_3, val_4\}$ be the set of valence scores of the emotions, which selected from ANEW lexicon[59]. We define val_i as the sentiment polarity of each emotion e_j, which is a real number and indicates the strength of each emotion.

For each s_i in S, $1 \leq i \leq l$, where l is batch size. And each e_j in E, $1 \leq j \leq 4$, we directly assess their sentence embedding s_i^* and e_j^* respectively, produced by the sentence encoder. For the pairwise sentiment similarity $sim\left(s_i^*, e_j^*\right)$, we compute it based on the method proposed in[55], that first compute the cosine similarity of the sentence embedding and emotion embedding, then use arccos to convert the cosine similarity into an angular distance, which had experimented to have better performance on sentiment similarity computing, that is,

$$sim\left(s_i^*, e_j^*\right) = \left(1 - \arccos\left(\frac{s_i^{*\top} e_j^*}{\| s_i^* \|\| e_j^* \|} / \pi\right)\right) \tag{3-1}$$

where $s_i^{*\top}$ represents the transpose of s_i^*. For each $sim\left(s_i^*, e_j^*\right)$, we use the softmax function to compute the weight probability $w_{i,j}$ as:

$$w_{i,j} = \frac{\exp\left(sim\left(s_i^*, e_j^*\right)\right)}{\sum_{j=1}^{4} \exp\left(sim\left(s_i^*, e_j^*\right)\right)} \tag{3-2}$$

Then the affective attention $a_{i,j}$ that sentence s_i oriented on each emotion is computed as below:

$$a_{i,j} = \alpha * \left(val_j w_{i,j}\right) \tag{3-3}$$

We add a scaling hyper-parameter α to increase the range of possible probability values for each conditional probability term. The sentiment-enriched text representations D can be induced as follows:

$$D = \sum_{i=1}^{l} \sum_{j=1}^{4} W_s s_i^* a_{i,j} \tag{3-4}$$

where W_s denotes sentence-level weight matrices, $D \in R^{l\times 4d^s}$, and d^s is the size of sentence embedding.

3.1.3 Multi-task learning

In this subsection, we introduce multi-task learning for both emotion classification and VAD regression task, as the knowledge learned in one task can usually improve the performance of another related task and enrich robustness of different type tasks[60,61]. Each sentence s_i in the training corpus has the following feature and label set $[s_i^*, (y_{emo,i}, y_{val,i}, y_{aro,i}, y_{dom,i})]$, where s_i^* represents the sentence embedding of s_i, and $(y_{emo,i}, y_{val,i}, y_{aro,i}, y_{dom,i})$ represent the associated categorical emotion, dimensional valence, arousal and dominance label separately. We apply CNN and three dense layers as informative feature extractor, then H^* is the final document vector. The probability of emotion classification task is computed by a *softmax* function:

$$P(y_{emo}) = softmax(W_e H^* + b_e) \tag{3-5}$$

where W_e and b_e are the parameters of the *softmax* layer. We use categorical cross entropy loss function for the first task, the objective function of this system is as follows:

$$J_e = -\frac{1}{l} \sum_{i=1}^{l} logP(y_{emo,i}) [y_{emo,i}] \tag{3-6}$$

where $y_{emo,i}$ is the expected class label of sentence s_i and $P(y_{emo,i})$ is the probability distribution of emotion labels for s_i. However, for the continuous labels, the *softmax* layer is not applicable, we use the *linear* function to predict the values for the VAD regression task. Then the predict value $y^p_{val|aro|dom,i}$ for sentence s_i is calculated using the following formula:

$$y^p_{val|aro|dom,i} = linear(W_s h_i^* + b_s) \tag{3-7}$$

where h_i^* represents the final vector of sentence s_i, W_e and b_e represent weights and bias respectively. Given l training sentences, we use the mean squared error loss function for VAD analysis, the loss between predicted dimensional values $y^p_{val|aro|dom,i}$ and original continuous labels $y^o_{val|aro|dom,i}$ is calculated as below:

$$L_{s,val|aro|dom} = \frac{1}{3l} \sum_{i=1}^{l} \left(y^p_{val|aro|dom,i} - y^o_{val|aro|dom,i}\right)^2 \tag{3-8}$$

Then the objective function for the whole system is:

$$J = J_e + \beta * (L_{s,act} + L_{s,aro} + L_{s,dom}) \tag{3-9}$$

where β is the hyper-parameter to control the influence of the loss of the regression function to balance the preference between classification and regression disagreements.

3.2 Experiments and analysis

3.2.1 Database and lexicon

3.2.1.1 The IEMOCAP emotion database

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database[62] contains videos of ten unique speakers acting in two different scenarios: scripted and improvised dialog with dyadic interactions. We only use the transcript data. To compared with state-of-the-art approaches, we use four emotion categories and three sentiment dimensions with 5531 utterances in this study. The four-class emotion distribution is: 29.6% happy, 30.9% neutral, 19.9% anger and 19.6% sad. Note that happy and excited category in the original annotation are included into happy class to balance data distribution between classes. For valence, arousal and dominance labels, self-assessment are used for annotation, in which the scale is from 1 to 5. In this paper, we focus on speaker-independent emotion recognition. We use the first eight speakers from session one to four as the training set, and session five as the test set.

3.2.1.2 The ANEW affective lexicon

The emotional values of the English words in Affective Norms for English Words (ANEW)[59] were calculated by means of measuring the psychological reaction of a person to the specific word. It contains real-valued scores for valence, arousal and dominance (VAD) on a scale of 1-9 each, corresponding to the degree from low to high for each dimension respectively. We select the *Valence* rating as the sentiment polarity which can distinguish different emotions of distinct words with the scale ranging from unpleasant to pleasant.

3.2.2 Experimental setup

Following[55], we set the dimension of the sentence embedding to 512. We use a convolutinoal layer with 16 filters each for kernel size of (4,4) and a max-pooling layer with the size of (2,2). As for dense layers, we use three hidden dense layers with 1024, 512 and 256 units and ReLU activation[63] separately. For regularization, we employ Dropout operation[64] with dropout rate of 0.5 for each layer. We set the mini-batch size as 50 and epoch number as 120, Adam[65] optimizer with a learning rate 0.0002, clipnorm as 5. And we set the parameter β to 1.0 to control the strength of the cost function for the VAD regression task.

We evaluate the experimental results of both single-task learning (STL) and multi-task learning (MTL) architecture. In the single-task architecture, we build seperate systems for emotion classification and VAD regression, whereas in multi-task architecture a join-model is learned for both of these problems.

3.2.3 Experimental results and analysis

3.2.3.1 Comparison to state-of-the-art approaches:

To quantitatively evaluate the performance of the proposed model, we compare our method with currently advanced approaches. The following are the commonly used benchmarks:

Tf-idf+Lexicon+DNN[1]: Introducing affective *ANEW*[59] lexicon and the term frequency-inverse document frequency (*tf-idf*) to construct the text features with DNN for emotion classification on IEMOCAP.

CNN[2]: A efficient architecture which achieves excellent results on multiple benchmarks including sentiment analysis.

LSTMs[3]: Two conventional stacked LSTM layers for emotion detection using the text transcripts of IEMOCAP.

Deepmoji[66]: Using the millions of texts on social media with emojis to pre-train the model to learn representations of emotional contents.

BiGRU+ATT[67]: A BiGRU network with the classical attention (ATT) mechanism.

BiLSTM+CNN[68]: Incorporating convolution with BiLSTM layer to sample more meaningful information.

BERT$_{BASE}$[14]: Bidirectional encoder with 12-layer Transformer blocks, which obtains new state-of-the-art results on sentence-level sentiment analysis.

In order to evaluate the performance, we present accuracy and F1-score for emotion classification task. As for VAD regression work, we use the mean squared error (MSE) and pearson correlation coefficient (r) to evaluation the performance, in which the lower MSE value and higher r correlation, the better performance. Experimental results of different methods in single task framework are shown in Table 3.1 and Table 3.2.

Table 3-1 F1, Accuracy for the comparative experiments in emotion classifiation framework. Acc.=Accuracy(%), Average(w)=Weighted average(%). The best results are in bold.

ID	Model	IEMOCAP									
		Happy		Anger		Neutral		Sad		Average(W)	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
1	Tf-idf+Lexicon+DNN[1]	63.80	69.29	68.24	67.64	60.68	58.84	62.86	57.69	63.89	63.39
2	CNN[2]	64.71	69.00	72.35	64.23	60.16	59.08	62.45	62.70	64.92	63.75
3	LSTMs[3]	60.41	69.08	71.18	66.30	61.72	59.18	68.98	62.25	65.57	64.20
4	Deepmoji[66]	58.37	66.15	61.18	63.03	72.14	61.56	63.67	66.10	63.84	64.21
5	BiGRU+ATT[67]	60.18	68.73	76.47	67.01	59.64	58.79	71.02	64.33	66.83	64.72
6	BiLSTM+CNN[68]	63.57	70.60	71.76	67.59	63.80	61.17	66.53	62.21	66.42	65.40
7	BERT_{BASE}[14]	59.05	69.23	72.35	65.78	67.19	63.70	73.88	66.54	68.12	66.31
Proposed	USE_T+SSOA+CNN	**69.91**	**72.88**	71.18	**70.14**	**67.71**	**65.74**	72.24	**71.08**	**70.26**	**69.96**

Table 3-2 MSE and r for the comparative experiments in VAD regression framework

ID	Model	IEMOCAP					
		Valence		Arousal		Dominance	
		MSE	r	MSE	r	MSE	r
1	Tf-idf+Lexicon+DNN[1]	0.755	0.435	0.536	0.277	0.638	0.318
2	CNN[2]	0.731	0.471	0.544	0.345	0.619	0.359
3	LSTMs[3]	0.626	0.575	0.413	0.425	0.536	0.447
4	Deepmoji[66]	0.655	0.499	0.417	0.421	0.514	0.458
5	BiGRU+ATT[67]	0.674	0.478	0.439	0.378	0.561	0.416
6	BiLSTM+CNN[68]	0.685	0.466	0.433	0.400	0.531	0.442
7	BERT_{BASE}[14]	0.566	0.587	0.416	0.464	0.564	0.460
Proposed	USE_T+SSOA+CNN	**0.523**	**0.603**	**0.402**	0.446	**0.511**	**0.486**

As shown in Table 3.1, our proposed model outperforms the state-of-the-art approaches with the absolute increase of more than 3.65%, 2.14% on average weighted F1, accuracy in the emotion classification task. As for VAD regression task, we can see from Table 3.2 that the proposed model *USE_T+SSOA+CNN* has better performance of consistently lower MAE and higher *r*. The results of the comparative experiments demonstrate the effectiveness of our proposed model. In order to illustrate the performance of our proposed SSOA mechanism and multi-task training, we do further researches in the following part.

3.2.3.2 Validation studies of proposed model:

We apply Universal Sentence Encoder which is trained with Transformer[55] (USE_T) to encode input texts into sentence embeddings and use CNN as the feature extractor. Therefore **USE_T+CNN** is the basic architecture and we control it as invarient.

USE_T+ATT+CNN: In order to validate our proposed SSOA mechanism, we also consider the most useful self-attention mechanism[56], which decide the importance of features for the prediction task by weighing them when constructing the representation of text.

USE_T+SSOA+CNN (STL): It is our work in single task framework, which uses SSOA mechanism to compute attention scores between the label and the representations of the sentences in the input that is to be classified. This can then be used to appropriately weight the contributions of each sentence to the final representations.

USE_T+SSOA+CNN (MTL): To demonstrates the effectiveness of incorporating VAD regression with emotion classification, we experiment this model in the multi-task framework which trained with both categorical emotion labels and dimensional valence, arousal, dominance labels.

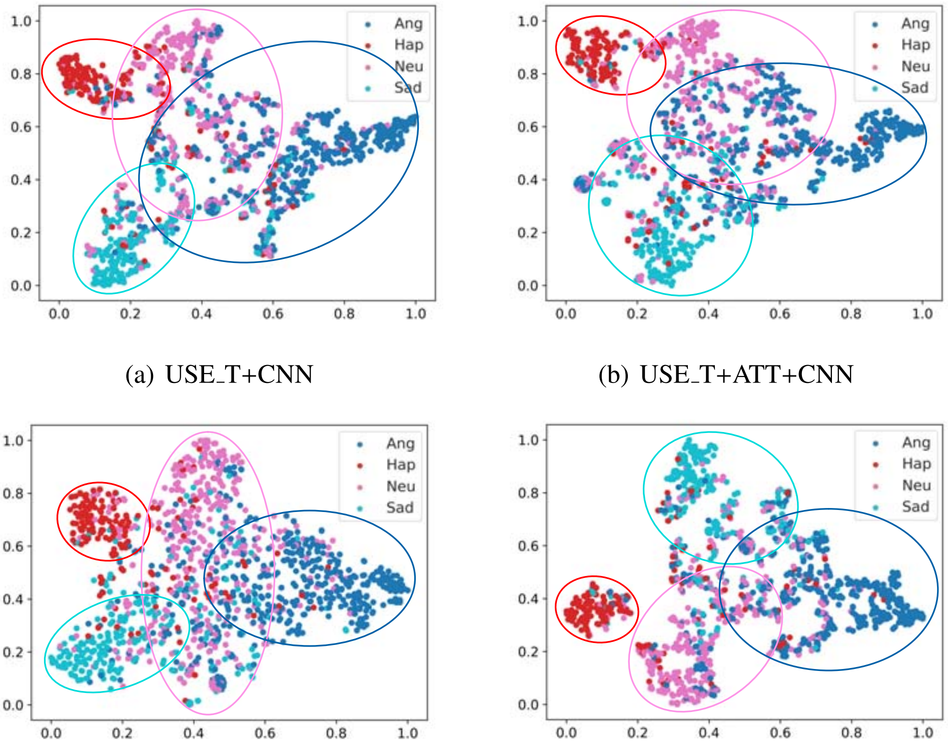
Table 3-3 Results (%) of Validation studies on emotion classification task

Model	IEMOCAP									
	Happy		Anger		Neutral		Sad		Average(W)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
USE_T+CNN	60.63	69.61	70.59	67.61	73.44	66.04	69.80	67.99	68.61	67.81
USE_T+ATT+CNN	69.00	70.77	68.82	68.82	69.01	65.11	66.12	69.53	68.24	68.56
USE_T+SSOA+CNN (STL)	66.97	73.27	70.00	71.47	71.61	64.94	69.80	68.95	69.60	69.66
USE_T+SSOA+CNN (SML)	**69.91**	72.88	**71.18**	70.14	67.71	**65.74**	**72.24**	**71.08**	**70.26**	**69.96**

Table 3-4 Results of validation studies on VAD regression task

Model	IEMOCAP					
	Valence		Arousal		Dominance	
	MSE	r	MSE	r	MSE	r
USE_T+CNN	0.595	0.570	0.431	0.418	0.563	0.464
USE_T+ATT+CNN	0.571	0.582	0.463	0.415	0.554	0.459
USE_T+SSOA+CNN(STL)	0.546	0.591	0.405	0.441	0.526	0.470
USE_T+SSOA+CNN(MTL)	**0.523**	**0.603**	**0.402**	**0.446**	**0.511**	**0.486**

From Table 3.3 and Table 3.4, some conclusions can be drawn as following: (1) Both *USE_T+ATT+CNN* with self-attention and *USE_T+SSOA+CNN* with our SSOA have a better performance than with no attention mechanism as expected. (2) Compared with *USE_T+ATT+CNN*, our *USE_T+SSOA+CNN* model achieves a relatively better result, especially achieves improvement about 2.5% in Happy, 2.65% in Anger on F1-score, and have accuracy improvement about 2.6% in Neutral, 3.68% in Sad separately. The results demonstrate that semantics of emotion labels can guide a model's attention when representing the input conversation and our proposed SSOA mechanism is able to capture sentiment-aware features, meanwhile, self-attention mechanism usually weights features based on semantic and context information which is not effective enough for emotion recognition. (3) Comparatively, as is shown in the last row, when both the problems are learned and evaluated in a multi-task learning framework, we observe performance enhancement for both tasks as well, which illustrates the effectiveness of multitask framework. And as we assume there are two reasons that VAD regression and emotion classification can assist each other task. On the one hand, emotions are high correlated with valence-arousal-dominance space. On the other hand, we take emotion labels into attention computing, which can help to capture more valence and arousal features.



(a) USE_T+CNN

(b) USE_T+ATT+CNN

(c) USE_T+SSOA+CNN(STL)

(d) USE_T+SSOA+CNN(MTL)

Figure 3-2 t-SNE visualization of validation studies on emotion classification.

Furthermore, in order to validate the effectiveness of our proposed method on different emotions, we introduce the t-Distributed Stochastic Neighbor Embedding (t-SNE)[69] for visualizing the deep representations as shown in Figure 3.2. We can see that compared with Figure 3.2 (a), the points which represent Anger in Figure 3.2 (b) can be distinguished more easily. The points which represent Happy and Sad have similar performance. Compared with Figure 3.2 (b), all the four emotion points have better discrimination in Figure 3.2(c) which means the deep representations extracted by our model are more sentiment-aware. However, we can observe from Figure 3.2(c) that most confusions are concentrated between Anger, Sad with Neutral. We assume the reason is that Anger and Sad hold the lowest percentage in IEMOCAP, which would not trained enough in our SSOA training process. Besides, the dataset we use is multimodal, a few utterances such as *"Yeah"*, *"l know"* carrying non-neutral emotions were misclassified as we do not utilize audio and visual modality in our experiments. In Figure 3.2(d), Sad can be distinguished better, we assume it's because Sad is one kind of negative valence and arousal values emotion according to Valence-Arousal representation[58], whose prediction would be more easy with the help of VAD.

Overall, the proposed *USE_T+SSOA+CNN* with multi-task learning model outperforms the other comparative and ablation studies. It is reasonable to assume that the proposed model is good at capturing both semantic and emotion features not only in emotion classification but also the VAD regression task.

3.3 Conclusion

In this section, we described our proposed sentiment similarity-oriented attention mechanism, which can be used to guide the network to extract emotion-related information from input sentences to improve classification and regression accuracy. In addition, to extract more robust features, we jointed dimensional emotion recognition using multi-task learning. The effectiveness of our proposed method has been verified under a series of comparative experiments and validation studies on IEMOCAP. The results show that the proposed method outperforms previous text-based emotion recognition by 6.57% from 63.39% to 69.96%, and show better robustness.

CHAPTER 4 Semantic- and Knowledge-guided Graph Convolutional Network for Context-dependent Emotion Recognition

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[7] [37] 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.[8] [9] [48] 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 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. Figure 4.1 is the overall architecture of our proposed model.

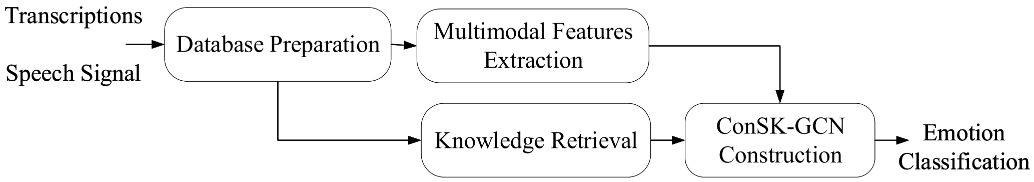
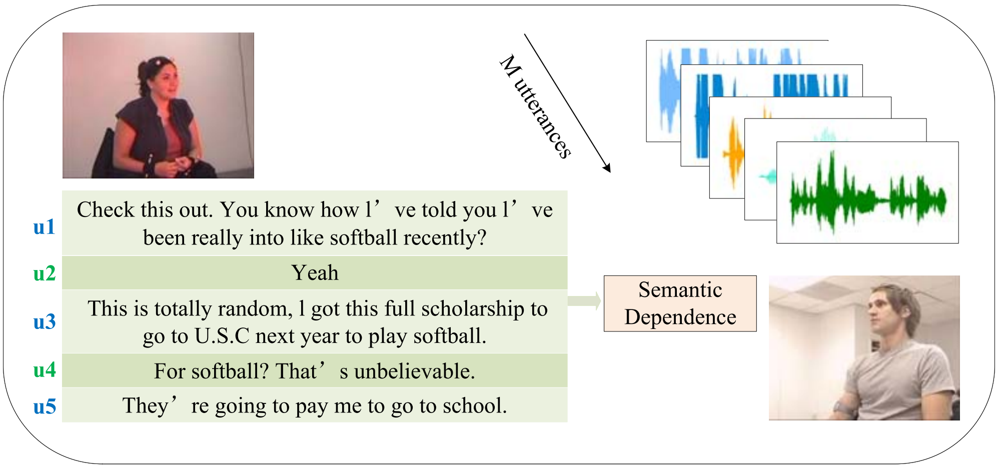


Figure 4-1 Overall architecture of our proposed ConSK-GCN approach for multimodal emotion recognition

4.1 ConSK-GCN model

4.1.1 Database preparation

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 4.2.



u1 Check this out. You know how I’ve told you I’ve been really into like softball recently?

u2 Yeah

u3 This is totally random, I got this full scholarship to go to U.S.C next year to play softball.

u4 For softball? That’s unbelievable.

u5 They’re going to pay me to go to school.

Figure 4-2 Architecture of database preparation

4.1.2 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, as shown in Figure 4.3.

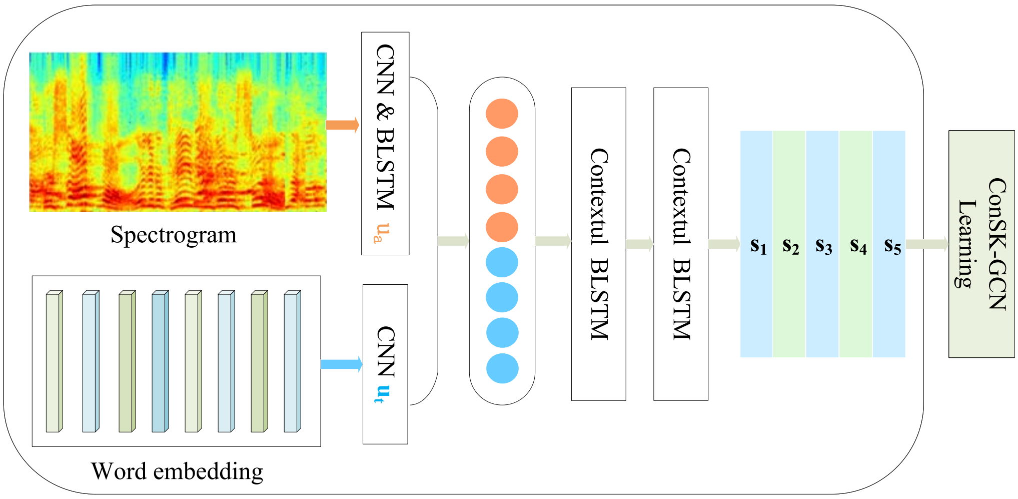


Figure 4-3 Architecture of multimodal features extraction

4.1.2.1 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[2] to extract textual embeddings of the transcripts. First, we use the publicly available pretrained word2vec[70] 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[63]. 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[39]. 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*[14] 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 for emotion label prediction from the transcript of the utterances. Then, we take the representations from the penultimate dense layer as the context independent utterance level feature vectors.

4.1.2.2 Acoustic features

In this paper, we follow the audio preprocessing method introduced in[71]. Researchers have found that a segment speech signal that is greater than 250-*ms* includes sufficient emotional information[72]. As the average utterance length of IEMOCAP dataset is around 2-*s*, and it's about 3.6-*s* in MELD dataset[39]. 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 $time \times frequency$: 32×129. 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 1874×129.

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 $\boldsymbol{\mu_a}$.

4.1.2.3 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 for sequence encoding to obtain the global utterance-level contextual information. Formally, we denote the context-aware multimodal representations as $\boldsymbol{s}$:

$$s_i = BiLSTM(s_{i(+,-)}, u_i) \tag{4-1}$$

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.

4.1.3 Knowledge retrieval

In this paper, we utilize external commonsense knowledge base ConceptNet[73] and an emotion lexicon NRC_VAD[74] 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 4.4, 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.

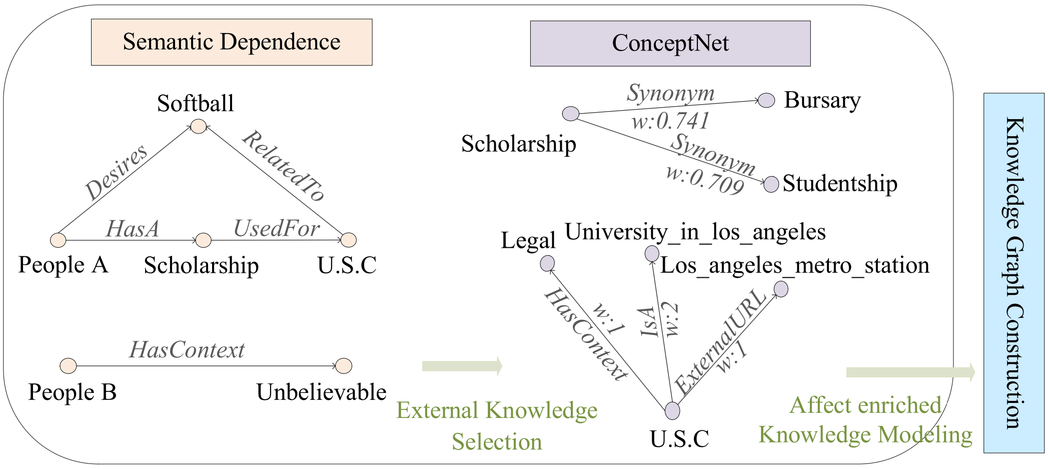


Figure 4-4 Architecture of external knowledge retrieval

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

4.1.4 ConSK-GCN construction

Figure 4.5 shows the architecture of our proposed ConSK-GCN approach for multimodal emotion recognition.

4.1.4.1 Knowledge graph construction

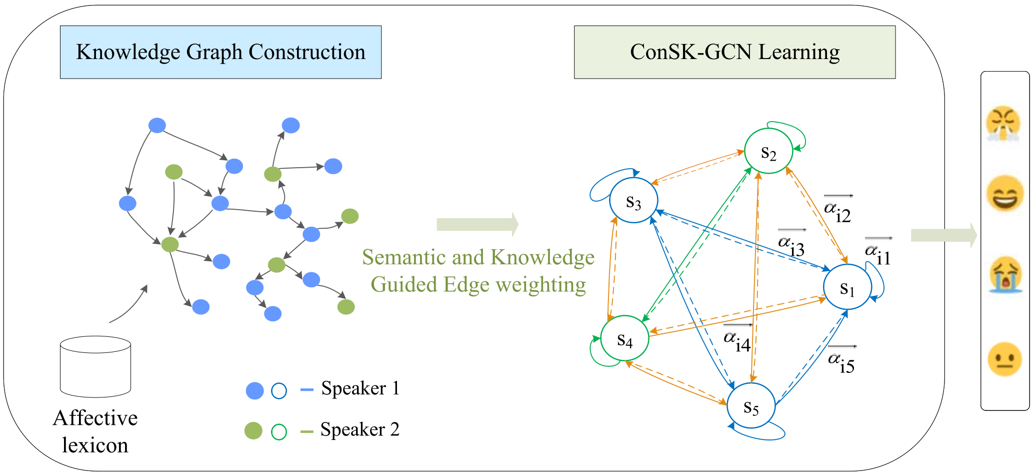

Knowledge Graph Construction
ConSK-GCN Learning
Semantic and Knowledge
Guided Edge weighting
Affective
lexicon
Speaker 1
Speaker 2
s1
s2
s3
s4
s5
$\overrightarrow{\alpha_{i1}}$
$\overrightarrow{\alpha_{i2}}$
$\overrightarrow{\alpha_{i3}}$
$\overrightarrow{\alpha_{i4}}$
$\overrightarrow{\alpha_{i5}}$

Figure 4-5 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 the concepts in V_k, we retrieve the corresponding valence (V) and arousal (A) scores from NRC_VAD, respectively.

Each node/concept in the knowledge graph is embedded into a single effective semantic space, named *ConceptNet Numberbatch*, that learns from both distributional semantics and ConceptNet. The tokens that are not included in the ConceptNet are initialized by the ”fastText” method[75], which is a library for efficient learning of word representations. For the concept not in the NRC_VAD, we set the VAD value to 0.5 as a neutral score.

The edges in the knowledge graph represent the knowledge relatedness between the concepts. First, for each concept $c_{i,m}$ in utterance i, we adopt l_2 norm to compute the emotion intensity emo_m, following[54], that is,

$$emo_m = min - max(\left\|[V(c_{i,m}) - 1/2, A(c_{i,m})/2]\right\|_2) \tag{4-2}$$

where $m = 1, ...n$, and n is the number of concepts in each utterance. $\|.\|_2$ denotes l_2 norm, $V(c_{i,m})$ and $A(c_{i,m})$ represent the corresponding valence and arousal score for each concept in utterance i. Then, following[8], we consider the past context window size of p and future context window size of f, and knowledge edge weights $a^k_{i,j}$ are defined as below:

$$k_{i,m} = emo_m c_{i,m} \tag{4-3}$$

$$a^k_{i,j} = \sum_{m=1}^{n} abs\left(\cos(k^{\top}_{i,m} W_k\left[\, k_{i-p,m}, ..., k_{i+f,m}\right])\right) \tag{4-4}$$

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

4.1.4.2 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[76] 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,

$$sim_{i,j} = 1 - \arccos\left(\frac{s_i^{\top} s_j}{\|s_i\| \left\|s_j\right\|}\right)/\pi \tag{4-5}$$

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

$$a^s_{i,j} = softmax(W_s[sim_{i-p}, ..., sim_{i+f}]) \tag{4-6}$$

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

4.1.4.3 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,

$$a_{i,j} = \omega_k a^s_{i,j} + (1 - \omega_k) a^k_{i,j} \tag{4-7}$$

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

$$h_i^{(1)} = \sigma(\sum_{r\in\Re}\sum_{j\in N_i^r}\frac{a_{i,j}}{q_{i,r}}W_r^{(1)}s_j + a_{i,i}W_0^{(1)}s_i) \tag{4-8}$$

$$h_i^{(2)} = \sigma(\sum_{j\in N_i^r}W^{(2)}h_j^{(1)} + W_0^{(2)}h_i^{(1)}) \tag{4-9}$$

where N_i^r 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 $q_{i,r} = |N_i^r|$), and $W_r^{(1)}$,$W_0^{(1)}$,$W^{(2)}$,$W_0^{(2)}$ 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 $\boldsymbol{h_i^{(2)}}$ are concatenated to obtain the final representations as following:

$$v_i = [s_i, h_i^{(2)}] \tag{4-10}$$

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

$$l_i = ReLU(W_l v_i + b_l) \tag{4-11}$$

$$P_i = softmax(W_p l_i + b_p) \tag{4-12}$$

$$\hat{y}_i = \underset{k}{argmax}(P_i[k]) \tag{4-13}$$

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:

$$L = -\frac{1}{\sum_{s=1}^{M} d(s)}\sum_{j=1}^{M}\sum_{i=1}^{d(s)} logP_{i,j}[y_{i,j}] + \lambda\,\|\theta\|_2 \tag{4-14}$$

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.

4.2 Experiments and analysis

4.2.1 Databses

We evaluate our ConSK-GCN on two conversational databases, namely *Interactive Emotional Dyadic Motion Capture* (IEMOCAP)[62] and *Multimodal EmotionLines Dataset* (MELD)[39]. 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.

As described in section 3, IEMOCAP database contains videos of ten unique speakers acting in two different scenarios: scripted and improvised dialog with dyadic interactions. And 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.[77]. *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 4.1. And the statistics of all the emotions are displayed in Table 4.2.

Table 4-1 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 4-2 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

4.2.2 Experimental setup

We choose ReLU as the activation and apply the method of stochastic gradient descent based on Adam[65] 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.

4.2.3 Comparison methods

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

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

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

bc-LSTM[4]: Utilized bidirectioinal LSTM network that takes as input the sequence of utterances in a video and extracts contextual unimodal and multimodal features by modeling the dependencies among the input utterances.

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

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

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.

4.2.4 Experimental results and analysis

4.2.4.1 Experiments on IEMOCAP

Table 4.3 and 4.4 indicate 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.

Table 4-3 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 4-4 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**

According to the emotion theory introduced in[78] that the Valence-Arousal space depicts the affective meanings of linguistic concepts. We believe that both *Anger* and *Happiness* are explicit emotions in linguistic features with positive arousal, which are also contagious in the context. Therefore, the information extracted both through "ConS-GCN" and "ConK-GCN" that based on context construction affect similar for recognizing them. By contrast, *Sadness* is relatively implicit in linguistic characteristics with negative valence and negative arousal. Compared to "ConS-GCN", "ConK-GCN" have a significant improvement in *Sadness* detection, and we observe that the recognition accuracy has increased by almost 3% as shown in Table 4.3, while it shows a more significant increase by nearly 8% in Table 4.5. 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.

Table 4.5 and 4.6 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 4-5 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 4-6 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 4.3, 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.

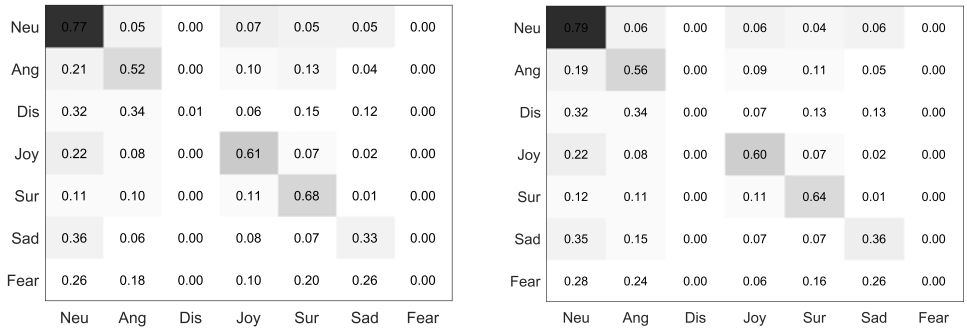
4.2.4.2 Experiments on MELD

Comparation with the state of the art: Table 4.7 and 4.8 depict the experimental comparations between our model and previous works in emotion recognition with both unimodality and multimodality. We can see from both table 4.7 and 4.8 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, 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.

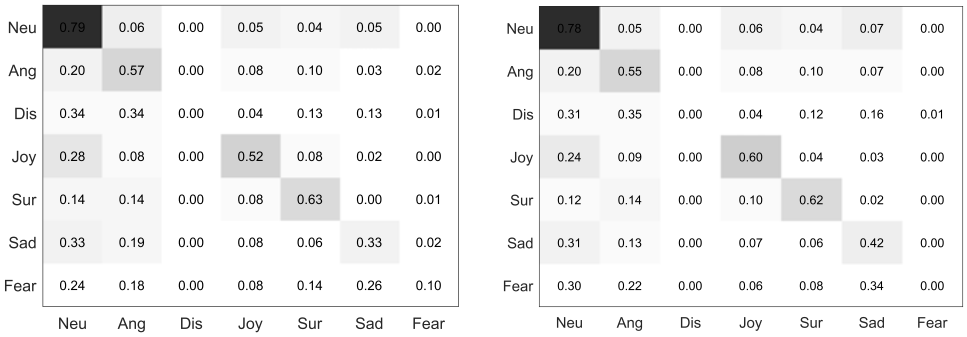
Ablation Studies: To further research and validate the performance of the proposed model, the comparative confusion matrices of classification results are shown in Figure 4.6, 4.7 and 4.8 separately.

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 4-6 Confusion matrix of the proposed ConS-GCN.



(a) Unimodality (b) Multimodality

Figure 4-7 Confusion matrix of the proposed ConK-GCN.

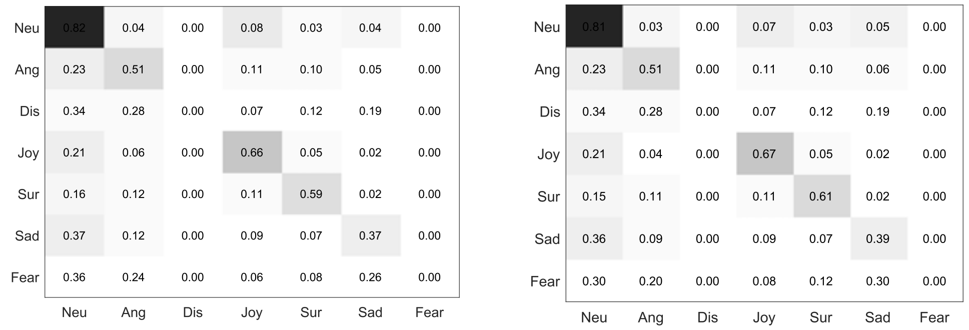
Table 4-7 Comparative experiments of different methods for unimodality (Text) emotion recognition. F1-score (%) is used as the evaluation metric. W= Weighted average.

Models	Neutral	Anger	Disgust	Joy	Surprise	Sadness	Fear	W-F1
CNN[2]	67.3	12.2	0.0	32.6	45.1	19.6	0.0	45.5
LSTMs[3]	67.6	12.3	0.0	36.0	45.7	17.2	0.0	46.0
bc-LSTM[4]	77.0	38.9	0.0	45.8	47.3	0.0	0.0	54.3
DialogueRNN[7]	73.7	41.5	0.0	47.6	44.9	23.4	5.4	55.1
DialogueGCN[8]	-	-	-	-	-	-	-	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**

We can see from Figure 4.8 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 integrating acoustic and linguistic features that are complementary in emotion recognition.

Table 4-8 Comparative experiments of different methods for multimodality (Text+ Audio) emotion recognition.

Models	Neutral	Anger	Disgust	Joy	Surprise	Sadness	Fear	W-F1
LSTMs[3]	68.1	31.4	0.0	34.5	44.9	7.24	0.0	47.6
bc-LSTM[4]	76.4	44.5	0.0	49.7	48.4	15.6	0.0	56.8
DialogueRNN[7]	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**

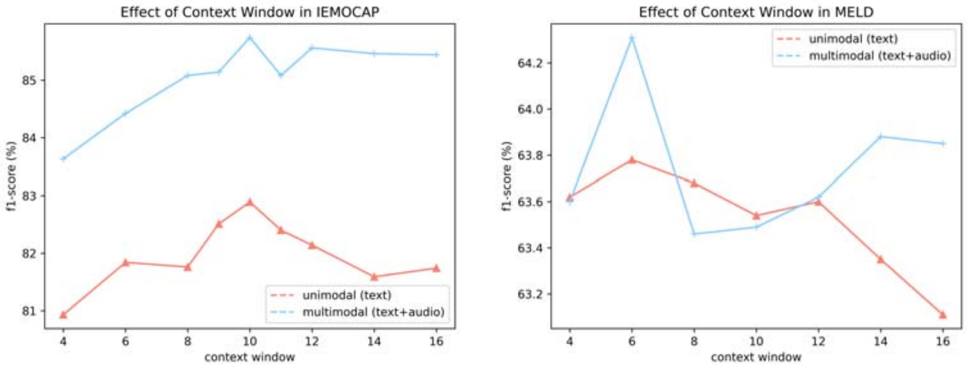


(a) Unimodality (b) Multimodality

Figure 4-8 Confusion matrix of the proposed ConSK-GCN.

4.2.5 Effect of Context Window

The accuracy of emotion detection in conversation varies with the context window. From Figure 4.9 (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.



(a) IEMOCAP (b) MELD

Figure 4-9 Effect of context window for emotion recognition in different datasets.

In the MELD dataset, we can conclude from the Figure 4.9 (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.

4.2.6 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 4.10.

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.

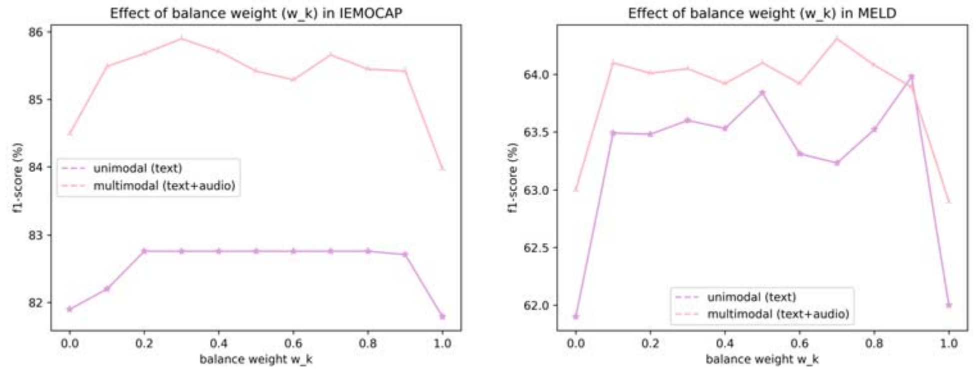
Cases in utterance 6-7 and the cases in utterance 8-9 are in the same situation with opposite results, where “ConS-GCN” weights more than “ConK-GCN” in “ConSK-GCN” learning, but information in “ConS-GCN” oriented to wrong direction in utterance 6-7, vice verse in utterance 8-9. External knowledge, sometimes 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.

4.2.7 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.

	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, l 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 4-10 Visualization of several representative examples. Blue denotes the typical concept in each utterance.



(a) IEMOCAP (b) MELD

Figure 4-11 Effect of balance weight (w_k) for emotion recognition in different dataset.

We can conclude from Figure 4.11 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 4.11 (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.

4.3 Conclusion

In this section, 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.

CHAPTER 5 Conclusions and Future Works

5.1 Summary

In this thesis, we proposed two models for extracting effective emotion features for emotion recognition. Firstly, we proposed a sentiment similarity-oriented attention mechanism, which can be used to guide the network to extract emotion-related information from input sentences to improve classification and regression accuracy in context-independent emotion recognition task. Secondly, we proposed a new conversational semantic- and knowledge-guided graph convolutional network (ConSK-GCN) for context-dependent emotion recognition, which leveraging both text and audio modaliteis. In this 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 with the state-of-the-art approaches show that our approach can significantly improve the performance of emotion detection, illustrating the effect of our proposed models.

5.2 Contributions

This thesis proposed a a sentiment similarity-oriented attention mechanism and a new semantic- and knowledge-aware graph convolutional neural network for emotion recognition. 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, both two technologies 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 have applied 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.

REFERENCES

[1] Kim E, Shin J W. Dnn-based emotion recognition based on bottleneck acoustic features and lexical features [C]. In ICASSP, 2019: 6720–6724.

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

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

[4] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos [C]. In ACL, 2017: 873–883.

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

[6] Zhao J, Chen S, Liang J, et al. Speech Emotion Recognition in Dyadic Dialogues with Attentive Interaction Modeling [C]. In INTERSPEECH, 2019: 1671–1675.

[7] Majumder N, Poria S, Hazarika D, et al. Dialoguernn: An attentive rnn for emotion detection in conversations [C]. In AAAI, 2019: 6818–6825.

[8] Ghosal D, Majumder N, Poria S, et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation [C]. In EMNLP-IJCNLP, 2019: 154–164.

[9] Zhang D, Wu L, Sun C, et al. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations [C]. In IJCAI, 2019: 5415–5421.

[10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

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

[12] Winata G I, Madotto A, Lin Z, et al. CAiRE_HKUST at SemEval-2019 Task 3: Hierarchical Attention for Dialogue Emotion Classification [J]. arXiv preprint arXiv:1906.04041, 2019.

[13] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations [J]. arXiv preprint arXiv:1802.05365, 2018.

[14] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[15] Schuller B W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends [J]. Communications of the ACM, 2018, 61 (5): 90–99.

[16] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction [J]. IEEE Signal processing magazine, 2001, 18 (1): 32–80.

[17] Huahu X, Jue G, Jian Y. Application of speech emotion recognition in intelligent household robot [C]. In 2010 International Conference on Artificial Intelligence and Computational Intelligence, 2010: 537–541.

[18] Low L-S A, Maddage N C, Lech M, et al. Detection of clinical depression in adolescents’ speech during family interactions [J]. IEEE Transactions on Biomedical Engineering, 2010, 58 (3): 574–586.

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

[20] Schuller B, Steidl S, Batliner A. The interspeech 2009 emotion challenge [C]. In Tenth Annual Conference of the International Speech Communication Association, 2009.

[21] Guo L, Wang L, Dang J, et al. Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine [J]. IEEE Access, 2019, 7: 75798–75809.

[22] Zhang S, Zhang S, Huang T, et al. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching [J]. IEEE Transactions on Multimedia, 2017, 20 (6): 1576–1590.

[23] Satt A, Rozenberg S, Hoory R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms [C]. In Interspeech, 2017: 1089–1093.

[24] Fu Y, Guo L, Wang L, et al. A Sentiment Similarity-Oriented Attention Model with Multi-task Learning for Text-Based Emotion Recognition [C]. In International Conference on Multimedia Modeling, 2021: 278–289.

[25] Araque O, Zhu G, Iglesias C A. A semantic similarity-based perspective of affect lexicons for sentiment analysis [J]. Knowledge-Based Systems, 2019, 165: 346–359.

[26] Zou Y, Gui T, Zhang Q, et al. A lexicon-based supervised attention model for neural sentiment analysis [C]. In Proceedings of the 27th International Conference on Computational Linguistics, 2018: 868–877.

[27] Khosla S, Chhaya N, Chawla K. Aff2Vec: Affect–Enriched Distributional Word Representations [J]. arXiv preprint arXiv:1805.07966, 2018.

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

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

[30] Gu Y, Yang K, Fu S, et al. Multimodal affective analysis using hierarchical attention strategy with word-level alignment [C]. In Proceedings of the conference. Association for Computational Linguistics. Meeting, 2018: 2225.

[31] Dasarathy B V. Decision fusion [M]. IEEE Computer Society Press Los Alamitos, 1994.

[32] Williams J, Comanescu R, Radu O, et al. Dnn multimodal fusion techniques for predicting video sentiment [C]. In Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML), 2018: 64–72.

[33] Poria S, Majumder N, Mihalcea R, et al. Emotion recognition in conversation: Research challenges, datasets, and recent advances [J]. IEEE Access, 2019, 7: 100943–100953.

[34] Li J, Zhang M, Ji D, et al. Multi-task learning with auxiliary speaker identification for conversational emotion recognition [J]. arXiv e-prints, 2020: arXiv–2003.

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

[36] Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv:1412.3555, 2014.

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

[38] Hazarika D, Poria S, Zadeh A, et al. Conversational memory network for emotion recognition in dyadic dialogue videos [C]. In Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, 2018: 2122.

[39] Poria S, Hazarika D, Majumder N, et al. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations [C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 527–536.

[40] Bronstein M M, Bruna J, LeCun Y, et al. Geometric deep learning: going beyond euclidean data [J]. IEEE Signal Processing Magazine, 2017, 34 (4): 18–42.

[41] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks [J]. IEEE transactions on neural networks and learning systems, 2020.

[42] Li Y, Tarlow D, Brockschmidt M, et al. Gated Graph Sequence Neural Networks [C]. In ICLR, 2016.

[43] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [J]. arXiv preprint arXiv:1609.02907, 2016.

[44] Seo Y, Defferrard M, Vandergheynst P, et al. Structured sequence modeling with graph convolutional recurrent networks [C]. In International Conference on Neural Information Processing, 2018: 362–373.

[45] Pan S, Hu R, Long G, et al. Adversarially regularized graph autoencoder for graph embedding [C]. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 2609–2615.

[46] Jain A, Zamir A R, Savarese S, et al. Structural-rnn: Deep learning on spatio-temporal graphs [C]. In Proceedings of the ieee conference on computer vision and pattern recognition, 2016: 5308–5317.

[47] Yao L, Mao C, Luo Y. Graph convolutional networks for text classification [C]. In AAAI, 2019: 7370–7377.

[48] Zhou J, Huang J X, Hu Q V, et al. SK-GCN: Modeling Syntax and Knowledge via Graph Convolutional Network for aspect-level sentiment classification [J]. Knowledge-Based Systems, 2020, 205: 106292.

[49] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks [C]. In European Semantic Web Conference, 2018: 593–607.

[50] Young T, Cambria E, Chaturvedi I, et al. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge [C]. In AAAI, 2018: 4970–4977.

[51] Mihaylov T, Frank A. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge [C]. In ACL, 2018: 821–832.

[52] Ghosal D, Hazarika D, Roy A, et al. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis [C]. In ACL, 2020.

[53] Zhang B, Li X, Xu X, et al. Knowledge Guided Capsule Attention Network for Aspect-Based Sentiment Analysis [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2538–2551.

[54] Zhong P, Wang D, Miao C. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations [C]. In EMNLP-IJCNLP, 2019: 165–176.

[55] Cer D, Yang Y, Kong S-y, et al. Universal sentence encoder for English [C]. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2018: 169–174.

[56] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]. In Advances in neural information processing systems, 2017: 5998–6008.

[57] Iyyer M, Manjunatha V, Boyd-Graber J, et al. Deep unordered composition rivals syntactic methods for text classification [C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015: 1681–1691.

[58] Giannakopoulos T, Pikrakis A, Theodoridis S. A dimensional approach to emotion recognition of speech from movies [C]. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009: 65–68.

[59] Warriner A B, Kuperman V, Brysbaert M. Norms of valence, arousal, and dominance for 13,915 English lemmas [J]. Behavior research methods, 2013, 45 (4): 1191–1207.

[60] Tafreshi S, Diab M. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning [C]. In Proceedings of the 27th international conference on computational linguistics, 2018: 2905–2913.

[61] Akhtar M S, Chauhan D S, Ghosal D, et al. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis [J]. arXiv preprint arXiv:1905.05812, 2019.

[62] Busso C, Bulut M, Lee C, et al. IEMOCAP: Interactive emotional dyadic motion capture database [J]. Language resources and evaluation, 2008, 42 (4): 335.

[63] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines [C]. In ICML, 2010.

[64] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv preprint arXiv:1207.0580, 2012.

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

[66] Felbo B, Mislove A, Søgaard A, et al. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm [J]. arXiv preprint arXiv:1708.00524, 2017.

[67] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification [C]. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016: 1480–1489.

[68] Zhou P, Qi Z, Zheng S, et al. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling [J]. arXiv preprint arXiv:1611.06639, 2016.

[69] Van Der Maaten L. Accelerating t-SNE using tree-based algorithms [J]. The Journal of Machine Learning Research, 2014, 15 (1): 3221–3245.

[70] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [C]. In ICLR, 2013.

[71] Guo L, Wang L, Dang J, et al. A feature fusion method based on extreme learning machine for speech emotion recognition [C]. In ICASSP, 2018: 2666–2670.

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

[73] Speer R, Chin J, Havasi C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge [C]. In AAAI, 2017: 4444–4451.

[74] Mohammad S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words [C]. In ACL, 2018: 174–184.

[75] Bojanowski P, Grave E, et al. Enriching Word Vectors with Subword Information [J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135–146.

[76] Cer D, Yang Y, Kong S, et al. Universal Sentence Encoder for English [C]. In EMNLP, 2018: 169–174.

[77] Hsu C-C, Chen S-Y, Kuo C-C, et al. EmotionLines: An Emotion Corpus of Multi-Party Conversations [C]. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[78] Osgood C E. The nature and measurement of meaning [J]. Psychological bulletin, 1952, 49 (3): 197–237.

[79] Liu B. Sentiment analysis and opinion mining [J]. Synthesis lectures on human language technologies, 2012, 5 (1): 1–167.

[80] Zhang M, Liang Y, Ma H. Context-aware affective graph reasoning for emotion recognition [C]. In ICME, 2019: 151–156.

[81] Zhao P, Hou L, Wu O. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification [J]. Knowledge-Based Systems, 2020, 193: 105443.

[82] Barrett L F. Discrete emotions or dimensions? The role of valence focus and arousal focus [J]. Cognition & Emotion, 1998, 12 (4): 579–599.

[83] Cliche M. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms [J]. arXiv preprint arXiv:1704.06125, 2017.

[84] Du J, Gui L, He Y, et al. A convolutional attentional neural network for sentiment classification [C]. In 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2017: 445–450.

[85] Li J-L, Lee C-C. Attentive to Individual: A Multimodal Emotion Recognition Network with Personalized Attention Profile [J]. Proc. Interspeech 2019, 2019: 211–215.

[86] Lian Z, Tao J, Liu B, et al. Conversational Emotion Analysis via Attention Mechanisms [J]. arXiv preprint arXiv:1910.11263, 2019.

[87] Marsella S, Gratch J. Computationally modeling human emotion [J]. Communications of the ACM, 2014, 57 (12): 56–67.

[88] Mohammad S M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text [M] // Mohammad S M. Emotion measurement. Elsevier, 2016: 2016: 201–237.

[89] Nielsen F Å. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs [C]. In Workshop on'Making Sense of Microposts: Big things come in small packages, 2011: 93–98.

[90] Osgood C E. The nature and measurement of meaning [J]. Psychological Bulletin, 1952, 49 (3): 197–237.

[91] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C]. In Proceedings of the 40th annual meeting on association for computational linguistics, 2002: 417–424.

[92] Picard R W. Affective computing: challenges [J]. International Journal of Human-Computer Studies, 2003, 59 (1-2): 55–64.

[93] Smetanin S. EmoSense at SemEval-2019 Task 3: Bidirectional LSTM Network for Contextual Emotion Detection in Textual Conversations [C]. In Proceedings of the 13th International Workshop on Semantic Evaluation, 2019: 210–214.

[94] Tammewar A, Cervone A, Messner E-M, et al. Modeling user context for valence prediction from narratives [J]. arXiv preprint arXiv:1905.05701, 2019.

[95] Tang D, Wei F, Qin B, et al. Sentiment embeddings with applications to sentiment analysis [J]. IEEE transactions on knowledge and data Engineering, 2015, 28 (2): 496–509.

[96] Xu P, Madotto A, Wu C-S, et al. Emo2vec: Learning generalized emotion representation by multi-task training [J]. arXiv preprint arXiv:1809.04505, 2018.

[97] Yang Y-H, Lin Y-C, Su Y-F, et al. A regression approach to music emotion recognition [J]. IEEE Transactions on audio, speech, and language processing, 2008, 16 (2): 448–457.

[98] Yang X, Macdonald C, Ounis I. Using word embeddings in twitter election classification [J]. Information Retrieval Journal, 2018, 21 (2-3): 183–207.

[99] Zhong P, Miao C. ntuer at semeval-2019 task 3: Emotion classification with word and sentence representations in rcnn [J]. arXiv preprint arXiv:1902.07867, 2019.

[100] Kuncheva L I, Bezdek J C, Duin R P. Decision templates for multiple classifier fusion: an experimental comparison [J]. Pattern recognition, 2001, 34 (2): 299–314.

[101] Hazarika D, Poria S, Mihalcea R, et al. ICON: interactive conversational memory network for multimodal emotion detection [C]. In Proceedings of the 2018 conference on empirical methods in natural language processing, 2018: 2594–2604.

发表论文和参加科研情况说明

（一）发表的学术论文

[1] 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, CCF-B)

[2] 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. (CCF-C)

（二）申请专利

[1] 王龙标，傅雅慧，党建武，郭丽丽，“一种基于情感相似度注意力机制的文本情感识别方法。”申请号：202010665789.8

（三）参与的科研项目

[1] “基于语言认知机理的类脑自然语音识别与交互”, 科技部国家重点研发计划“智能机器人”专项课题（No. 2018YFB1305200）, 2019.6-2022.5

[2] “面向机器人的复杂环境语音对话关键技术及系统实现”, 新一代人工智能科技重大专项（No.18ZXZNGX00330）, 2018.10-2021.9

[3] “面向混响环境的多口音语音识别研究”, 国家自然科学基金面上项目（No.61771333）, 2018.1-2021.12

致 谢

I would like to express my sincere thanks to my supervisor, professor Wang Longbiao, thank him for his guidance and support. Professor Wang has been giving me many suggestions on my research work, without his advising, this work could not be completed. 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.

At the same time, I would like to express my thanks to my supervisor in JAIST, professor Okada Shogo. 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. 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.

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 study and life.