



研究与开发

基于语篇解析和图注意力网络的对话情绪识别

郝秀兰¹, 魏少华¹, 曹乾², 张雄涛¹

(1. 湖州师范学院浙江省现代农业资源智慧管理与应用研究重点实验室, 浙江 湖州 313002;

2. 中国建设银行股份有限公司湖州分行, 浙江 湖州 313001)

摘要: 对话情绪识别研究主要聚焦于融合对话上下文和说话者建模的相互关系。当前研究通常忽略对话内部存在的依存关系, 导致对话的上下文联系不够紧密, 说话者之间的关系也缺乏逻辑。因此, 提出了一种基于语篇解析和图注意力网络 (discourse parsing and graph attention network, DPGAT) 的对话情绪识别模型, 将对话内部的依存关系融入语境建模过程中, 使语境信息更具有依赖性和全局性。首先, 通过语篇解析获取对话内部的话语依存关系, 构建语篇依存关系图和说话者关系图。随后, 通过多头注意力机制将不同类型的说话者关系图进行内部融合。此外, 在图注意力网络的基础上, 结合依存关系进行循环学习, 以达到上下文信息和说话人信息的有效融合, 实现对话语境信息的外部融合。最终, 通过分析内、外部融合的结果还原完整对话语境, 并对说话者的情绪进行分析。通过在英文数据集 MELD、EmoryNLP、DailyDialog 和中文数据集 M3ED 上进行评估验证, F1 分数分别为 66.23%、40.03%、59.28%、52.77%, 与主流模型相比, 所提模型具有较好的适用性, 可在不同的语言场景中使用。

关键词: 对话情绪识别; 语篇解析; 图注意力网络

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2024149

Emotion recognition in conversations based on discourse parsing and graph attention network

HAO Xiulan¹, WEI Shaohua¹, CAO Qian², ZHANG Xiongtao¹

1. Zhejiang Province Key Laboratory of Smart Management & Application of Modern Agricultural Resources,
Huzhou University, Huzhou 313002, China

2. China Construction Bank Co., Ltd. Huzhou Branch, Huzhou 313001, China

Abstract: The research on emotion recognition in conversations (ERC) focuses on the interrelationship between conversational context and speaker modeling. The current research usually ignores the dependency within the conversation, which leads to the weak connection between the context of the conversation and the lack of logic between the

收稿日期: 2024-01-12; 修回日期: 2024-04-25

通信作者: 郝秀兰, hx12221_cn@zjhu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.62376094); 湖州师范学院研究生科研创新项目 (No.2024KYCX46, No.YJK24005)

Foundation Items: The National Natural Science Foundation of China (No.62376094), Postgraduate Research and Innovation Project of Huzhou University (No.2024KYCX46, No.YJK24005)

speakers. Therefore, an emotion recognition in conversations model based on discourse parsing and graph attention network (DPGAT) was proposed to integrate the inter-dependency of conversation into the context modeling to make contextual information more dependent and global. Firstly, the discourse dependency relationships within the conversation were obtained through discourse parsing, and the discourse dependency graph and the speaker relationship graph were constructed. Subsequently, different types of speaker diagrams were internally integrated by multi-head attention mechanisms. Based on the graph attention network, cyclic learning was combined with dependency relationships to achieve the effective integration of contextual information and speaker information, realizing the external integration of context-related information in conversations. Finally, by analyzing the results of internal and external integration, the complete conversation context was restored, and the speaker's emotions were analyzed. By evaluating and verifying on English dataset MELD, EmoryNLP, DailyDialog and Chinese dataset M3ED, F1 scores were 66.23%, 40.03%, 59.28% and 52.77%, respectively. Compared with mainstream models, the proposed model at least reaches state-of-the-art, and can be used in different language scenarios.

Key words: emotion recognition in conversations, discourse parsing, graph attention network

0 引言

情感分析是自然语言处理领域的研究热点之一。作为情感分析任务的一个分支,对话中的情感识别旨在对话过程中识别每个语句的情感类别^[1]。该任务面临的挑战和难点主要涵盖对话语境和说话人参与这两个方面^[2-3]:首先,对话语境具有口语化严重、富含隐式情感信息等特点,在语句内容分析时仅考虑单一的语句进行情感分类会导致缺乏上下文关联、无法处理多说话者情境,所以需要对当前语句进行上下文依赖关系分析,以进行全面的语境建模;其次,说话人的参与表示对话的主体是说话人,相同的话语如果由不同说话者表达可能会产生不同的影响。因此,在语境还原时需要关注说话者的状态以及说话者之间的依赖关系。

已有研究中,很多工作采用预训练模型作为下游模型输入的特征提取器或直接使用下游数据集对其进行微调,但依旧停留在语句级分类阶段,如BERT、BART、RoBERTa等,这些方法将每个对话都看作独立的语句,以至于未考虑其上下文相关性和说话者信息^[4]。此外,部分研究采用多层递归神经网络和图神经网络,将深度神经网络应用到图神经网络成为当前热点^[5],如

DisGCN、DAG-ERC等,对语境进行上下文建模,但仍存在多次循环训练之后趋向平稳的问题^[6]。除此之外,还有研究采用外部推理知识注入的方式,如COSMIC、CauAIN等,这些方法通过引入情感词典、主题信息和心理知识等外部资源丰富语境表征,从而增强语境理解^[6]。文献[7]基于语篇解析,提出通过预测依存关系和构建篇章依存树用于解析多方对话,辅助还原语境,以便提高对语境的依赖性和全局性。

本文提出一种基于语篇解析和图注意力网络的对话情绪识别模型:通过语篇解析的方式注入语句之间的依存关系,进而利用图注意力网络^[8](graph attention network, GAT)融合上下文语境和说话人关系有效整合对话语境。本文的主要贡献如下。

(1) 为了使对话语境具有依赖性和全局性,对完整对话通过语篇解析构建依存关系,并融入语境建模中,更好地还原对话语境。

(2) 为了使模型更好地融合上下文信息和说话人信息,使用内部融合、外部融合两个模块,分别利用多头注意力机制和图注意力网络进行循环学习,以达到按照权重分析信息的效果。

(3) 为了更好地检验所提模型在中、英文对话中的有效性,在英文数据集MELD、Emo-



ryNLP、DailyDialog 和中文数据集 M3ED 上进行评估验证, 结果表明模型具有较好的适用性, 可在不同的语言场景中使用。

1 相关工作

对话情绪识别任务在其发展历程中主要分为以下几个方向。

语句级情绪识别: 将每个话语视作独立的句子, 即进行简单的句子分类。有的研究直接采用预训练模型如 BERT、BART 和 RoBERTa 等作为下游模型输入的特征提取器或直接对其进行微调, 也有研究将神经网络用于独立提取语句特征来进行分类^[4]。

上下文建模: 通过分析上下文信息, 构建语句之间的关系和逻辑^[9], 以辅助理解对话内容, 从而更准确地识别当前语境中句子所具备的情绪。例如, 文献[10]引入长短期记忆网络, 用于建模对话中的上下文信息, 证明了上下文信息对情绪预测的辅助效果; 文献[11]通过多层递归神经网络提取语句信息并处理对话中连续的句子。

说话人建模: 考虑说话者情绪状态会受到其他说话者的影响, 一些研究将说话者之间的关系进行建模。例如 DialogueGCN 引入图神经网络, 将对话建模为完全连接的图, 考虑了全局的说话者信息和上下文信息; DAG-ERC 采用有向无环图来建模说话者和话语之间的交互^[6]。

外部信息注入: 还原对话语境的前提是融合上下文信息和说话者信息, 一些研究者提出引入其他辅助信息来丰富语境。COSMIC 包含了不同的常识元素, 如心理状态、事件和因果关系, 并在此基础上学习参与对话的对话者之间的互动; CauAIN 因果感知交互网络借助情感原因检测来彻底理解会话上下文^[6]。

语篇解析: 文本语篇解析通过识别基本语篇单元之间的链接和关系, 将一段文本转换为结构化格式。对话示例 (源自 MELD 数据集^[12]) 如图

1 所示, 关系标注类别源自 STAC 语料库^[7], 包含“问答对”“结果”“解释阐述”等 16 种关系。

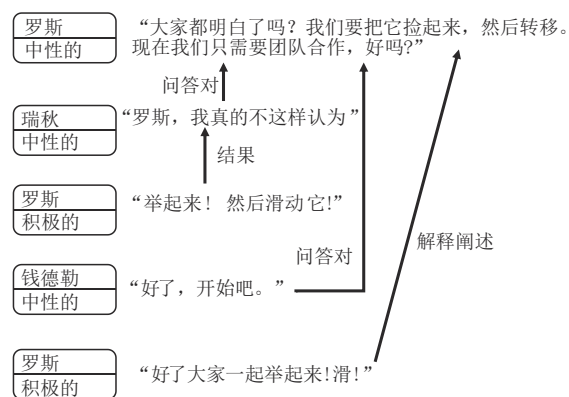


图1 对话示例(源自 MELD 数据集^[12])

文献[7]提出预测依存关系和联合交替构建篇章依存树用于深度序列模型解析多方对话。文献[13]构建了一个基于 Transformer 的解析器, 从数据和语言建模的角度进行领域集成, 以提高泛化能力。MUCDN^[14]提出相互对话分离网络, 结合语篇解析分析的话语依存关系, 通过将对话分解为独立的线程, 在独立的线程内执行上下文和说话者特定建模。DualGATs^[6]模型同时考虑话语依赖关系和讲话者感知上下文的互补方面。

尽管研究者在上下文建模和说话者建模融合方面取得了显著成果, 但忽视了语句之间依赖性的重要性以及语句依存关系对于全局性的影响。为了能够更全面地考虑语句之间的依赖关系, 将对话语境还原得更为细致和全面, 提出基于语篇解析和图注意网络的模型, 为情绪识别提供更为准确的背景信息。

2 DPGAT 对话情绪识别模型

首先给出对话情绪识别任务的定义: 数据由多个对话组成, 每段对话包含 N 个连续语句和 M 个说话人 ($M > 2$) 的会话。任务的目标是为对话中的每一句话分配一个情感标签, 例如喜悦、悲伤、生气等, 其中 Y 表示可能的情感标签集合。

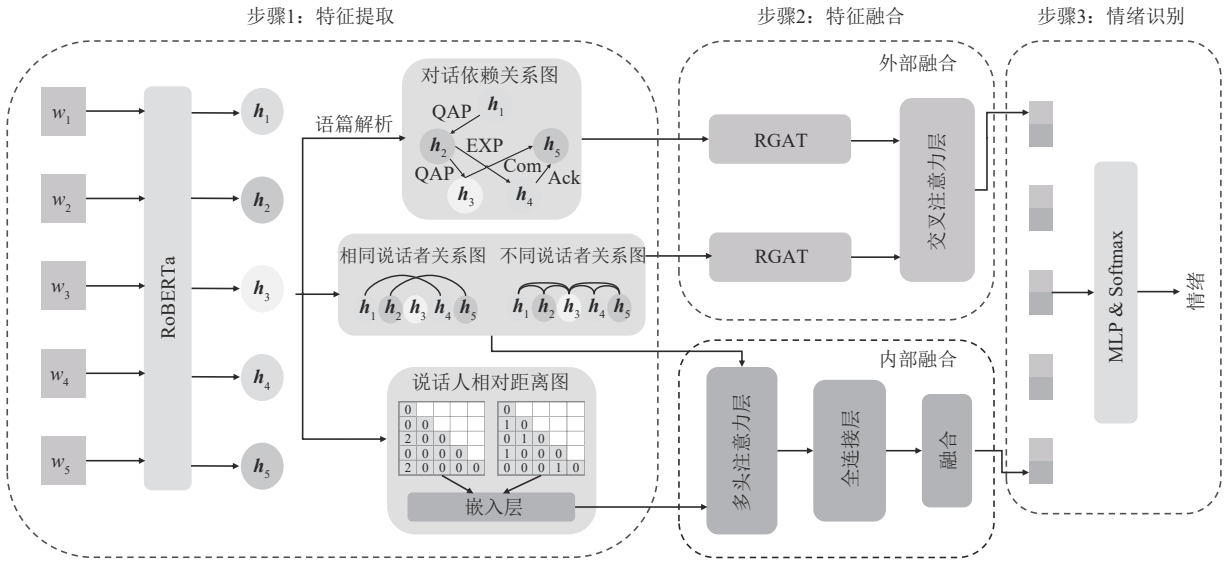


图2 模型架构

模型架构如图2所示，本文模型的主要架构分为4个部分：特征提取、内部融合、外部融合和情绪识别。在特征提取模块中，对话文本通过预训练语言模型RoBERTa进行特征提取，随后传入基于Transformer的语篇解析器^[13]以提取话语依存关系，并构建语篇依存关系图和说话人关系图。内部融合模块考虑不同类型的说话者和相互依赖关系，在多头注意力机制的作用下完成建模。外部融合模块将上下文信息和说话者信息在图注意力网络的基础上经过多层迭代训练，实现上下文建模和说话者建模的交叉融合，更好地捕捉语境信息。最终，情绪识别模块对整合的特征进行分类，完成对话情绪的识别任务。

2.1 特征提取

2.1.1 语句级特征提取

使用预训练语言模型RoBERTa来执行语句级特征提取。对于每个语句 $u_i = \{w_1, w_2, \dots, w_L\}$ ，将一个特殊的令牌[CLS]连接到语句的开头，得到输入序列 $\{[CLS], w_1, w_2, \dots, w_L\}$ 。通过RoBERTa模型进行微调，提取最后一个隐藏层的[CLS]令牌输出激活，得到特征结果。在微调之后，为了获得由[CLS]令牌表示的每个语句级特征向量 h_i ，将

每个语句以与 $\{[CLS], w_1, w_2, \dots, w_L\}$ 相同的输入格式进行输入，并对每个特征进行归一化，使得它们服从均值为0、标准差为1的分布：

$$h_i = \text{LayerNorm}(\text{RoBERTa}([CLS], w_1, w_2, \dots, w_L)) \quad (1)$$

其中， $h_i \in \mathbf{R}^{d_m}$ ， d_m 为RoBERTa中隐藏状态的维数。经过COSMIC模型处理，最后4层的[CLS]令牌计算平均以获得每个语句的语句级特征向量。将每个话语向量 h_i 变换到 d_h 的维度，并进行线性投影，对话 H 的向量化表示为 $\{h_1, h_2, \dots, h_N\}$ 。

2.1.2 语篇解析

基于Transformer的语篇解析器在STAC和Molweni语料上表现良好^[13]。每一段对话都被视为一个基本篇章单元（elementary discourse unit, EDU），解析器用于预测EDU之间的链接和对应关系类型。给定每个对话 $\{u_1, u_2, \dots, u_N\}$ ，使用这个预训练的解析器来预测对话中的话语依赖关系：

$$\{(j, i, r_{ij}), \dots\} = \text{Parser}(u_1, u_2, \dots, u_N) \quad (2)$$

其中， r_{ij} 是有向边，首为 u_j ，尾为 u_i ， i 和 j 是对话 $(i < j)$ 中话语的索引。预测的结果构成一棵依赖



树，它是有向无环图（directed acyclic graph, DAG）的一种特殊类型，其中每个节点与单个父节点相连。该话语依赖图中， $[j, i]$ 记为边， r_i 记为其对应的关系类型 r （ $r \in R$ ）。其中 R 的分类利用了STAC中的16种话语依存类型，包括注释、澄清疑问、详述、确认、延续、解释、条件、问答对、替代、问答、结果、背景、叙述、更正、平行和对比。

2.2 特征内部融合

相同的话语，如果发出的说话者不同也会对结果产生很大的影响。受到MUCDN^[14]模型中提到的对于说话者建模分为是否来自同一说话人的启发，构造带有话语依赖关系信息注入的说话者关系图（speaker relationship graph, SRG）。根据两句话是否来自同一个说话人，以邻接矩阵的形式得到 SRG^{intra} 和 SRG^{inter} ：

$$SRG_{i,j}^{intra} = \begin{cases} 0, & j \leq i \text{ \& } \varphi(u_i) = \varphi(u_j) \\ -\infty, & \text{其他} \end{cases} \quad (3)$$

$$SRG_{i,j}^{inter} = \begin{cases} 0, & j < i \text{ \& } \varphi(u_i) \neq \varphi(u_j) \\ -\infty, & \text{其他} \end{cases} \quad (4)$$

在还原对话语境时，考虑到时间上更近的语句可能会对目标语句产生较为直接的影响，因此在一定范围内的语句之间，根据话语依赖关系来计算语句之间的相对距离。按照Dijkstra算法为不同说话人类型计算最短距离；若两个节点彼此不可达，则将位置值设置为零，从而得到两个位置矩阵 L^{intra} 和 L^{inter} 。之后经过位置编码，即采用位置嵌入层将 L^{intra} 或 L^{inter} 中的每个元素编码为一个标量值：

$$Loc^t = \text{Embedding}(L^t) \quad (5)$$

根据多头注意力机制^[15]，首先将对话初始特征 H 经过线性投影的方式得到 Q 、 K 、 V 3个向量；其次，将SRG和 Loc^t 加入 Q 与 K 的点积结果中，实现说话人信息的传播：

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)M \quad (6)$$

其中，Concat表示拼接操作， H 表示注意力头数， M 表示最终输出的线性变化矩阵， $M \in \mathbb{R}^{N \times dh}$ 。

$$\text{head}_i = \text{Attention}(Q, K, V, SRG^t, Loc^t) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + SRG^t + Loc^t\right)V \quad (7)$$

其中，Attention表示图中的缩放点积注意力， $t \in \{\text{intra}, \text{inter}\}$ 表示SRG和Loc的类型。 Loc^t 可以被看作一个额外的权重，影响说话者之间关系的融合过程。

$$H_temp = H \odot \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) \quad (8)$$

其中， H 表示注意力头数，用注意力分数的和与数据特征相乘得到 H_temp ，使具有较大注意力权重的特征得到更多关注，而注意力分数较小的特征则受到较少关注。

将经过组合的特征通过全连接层（fully connected layer, FC），之后加入ReLU激活函数得到 C^t ， $t \in \{\text{intra}, \text{inter}\}$ ，如式（9）所示；再通过Sigmoid激活函数将全连接层的输出转换为在（0,1）范围内的概率值，即权重值 g ，如式（10）所示。

$$C^t = \text{ReLU}\left(\text{FC}\left([H, M^t, H - M^t, H \odot M^t]\right)\right) \quad (9)$$

$$g = \text{Sigmoid}\left(\text{FC}\left[C^{intra}, C^{inter}\right]\right) \quad (10)$$

最后，将两个说话人关系图结果表征以门控的方式调节输入的权重进行融合，如式（11）所示：

$$I = g \odot C^{intra} + (1 - g) \odot C^{inter} \quad (11)$$

2.3 特征外部融合

考虑上下文和说话者的影响，除了关注语句附近的信息，也考虑对话的全局性特征依赖。同时，鉴于相同话语在不同的语境和说话者的前提下会产生不同的情绪状态，而GAT具有处理输入图中多关系类型能力，每个节点只聚合其邻边的信息^[16]。为了融合上下文信息和说话者信息，在GAT的基础上考虑节点（语句）之间依存关系，构造模块（relationship-aware graph attention network, RGAT）完成图节点间的信息传播和聚合

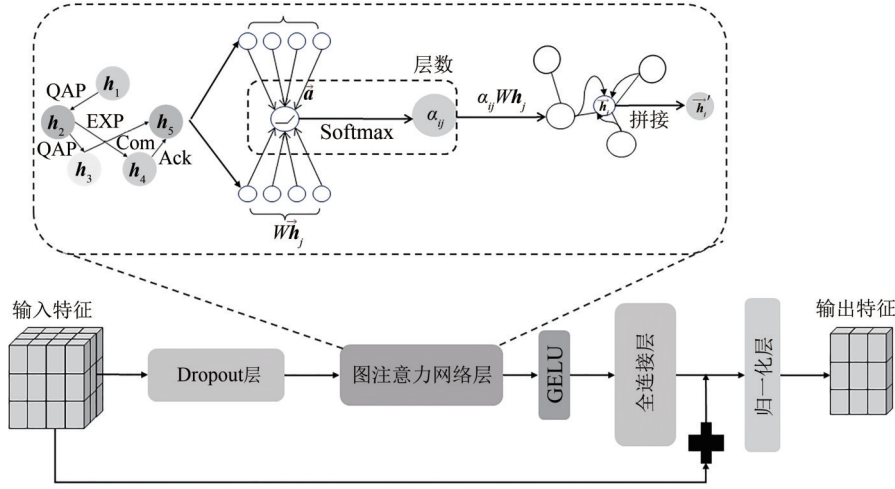


图3 RGAT 架构

篇章结构信息，RGAT 架构如图 3 所示。将语句级特征和关系矩阵图输入 RGAT 模块中，使得每个节点在信息传播的过程中能够更加关注与其相关的邻居节点以及连接边所蕴含的信息。

根据解析器获得的话语依存关系，构造 $L \times L$ 的结构邻接矩阵，记为语篇依存关系图 (dialogue relation graph, DRG)。矩阵的元素表示节点之间的关系，对于边 (i, j) ，其关系类型为 r ，其中 17 可以代表某种默认关系或者占位符：

$$\text{DRG}_{i,j} = \begin{cases} r+1, & (i,j)=r \\ 0, & (i,j)=\emptyset \\ 17, & i=j \end{cases} \quad (12)$$

对于给定节点 V_i^{dia} ，RGAT 模块将其邻居节点的信息按式 (13)、式 (14) 进行聚合：

$$\alpha_{i,j} = \text{Softmax}_i \left(\text{LeakyReLU} \left(a^T [Wh_i \| Wh_j \| V_{ij}^{\text{dia}}] \right) \right) \quad (13)$$

其中， α_{ij} 表示节点 V_i^{dia} 到其邻居 V_j^{dia} 的边权重即注意力系数， W 和 a 表示训练参数， $V_{ij}^{\text{dia}} \in R^{|\text{dia}|}$ 表示节点 V_i^{dia} 和 V_j^{dia} 之间的话语依赖类型对应的独热编码， $\|$ 表示拼接操作。

$$h_i^{\text{dia}} = \sum_{j \in N_i^{\text{dia}}} \alpha_{ij} Wh_j \quad (14)$$

其中， N_i^{dia} 表示节点 V_i^{dia} 在 DRG 中的邻居节点，

$h_i^{\text{dia}} \in R^{\text{dh}}$ 表示节点 V_i^{dia} 经过 RGAT 更新后的隐藏表示。所有节点更新后的隐藏表示记为 $D \in R^{N \times \text{dh}}$ ，上下文信息的计算过程如式 (15) 所示：

$$D = \text{RGAT}(H_{\text{temp}}, E^{\text{DRG}}) \quad (15)$$

探究说话者信息时，将前述构造的相同说话者关系图和不同说话者关系图合并为说话者关系图 SRG。同理，使用 RGAT 模块将说话者信息的计算过程定义为式 (16)：

$$S = \text{RGAT}(H_{\text{temp}}, E^{\text{SRG}}) \quad (16)$$

然后，采用相互交叉注意力作为桥梁交互集成这两个模块，有效地交换相关信息。整个交互过程迭代地提炼和交换篇章结构信息和说话人感知的上下文信息，对初始层的计算过程进行了泛化，如式 (17)~式 (19) 所示。

$$D^{[k]} = \text{Softmax} \left(D^{[k-1]} \odot W_1 \odot S^{[k-1]T} \right) S^{[k]} \quad (17)$$

$$S^{[k]} = \text{Softmax} \left(S^{[k-1]} \odot W_2 \odot D^{[k-1]T} \right) D^{[k]} \quad (18)$$

其中， $W_1, W_2 \in R^{\text{dh} \times \text{dh}}$ 为可学习参数，当 $D^{[0]} = S^{[0]} = H^u$ 以及 $k \in [0, k-1]$ 时，有：

$$E = \text{FC}(H_{\text{temp}}, D^{[k]}, S^{[k]}) \quad (19)$$

2.4 情绪识别

完成特征处理的操作之后，需要进行最终的分类步骤，构造了一个多层感知器 (multilayer



perceptron, MLP) 分类模型, 以 H 、 I 和 E 的串联作为输入来预测语句的情感:

$$\hat{y} = \text{Softmax}(W_e[H, I, E] + b_e) \quad (20)$$

其中, W_e 和 b_e 为可训练参数。

最后, 利用交叉熵损失对模型进行训练:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{\text{Emo}} \hat{y}_i^j \cdot \log(y_i^j) \quad (21)$$

其中, Emo 是情感类别的个数, y_i^j 表示语句 i 的真实情感标签。

3 实验与分析

3.1 数据集

所提的模型主要在英文数据集 (MELD^[12]、EmoryNLP (ENRP)^[17]、DailyDialog (DD)^[18]) 和中文数据集 (M3ED^[19]) 上进行实验, 数据集的统计数据见表 1。

表 1 数据集的统计数据

数据集		MELD	ENLP	DD	M3ED
对话数量	Train	1 038	659	11 118	685
	Dev	114	89	1 000	126
	Test	280	79	1 000	179
语句数量	Train	9 989	7 551	87 170	17 427
	Dev	1 109	1 344	954	2 821
	Test	2 610	1 328	984	4 201

MELD 是源自美国电视剧《老友记》的一个多模态数据集, 在实验中仅使用了文本数据。对话情感标记包括愤怒、厌恶、悲伤、高兴、惊讶、恐惧以及中性。

EmoryNLP 也是来自电视节目《老友记》的一个数据集, 但标注了不同的情感标签类别。预先定义的情感是中性、悲伤、疯狂、害怕、强大、平和、快乐。

DailyDialog 是手动编译的有关日常交流的数据集, 仅涉及两方且没有发言者信息。该数据集中使用的标注方法是 Ekman 的情绪类型: 中性、快乐、惊讶、悲伤、愤怒、厌恶、恐惧。

M3ED 是一个大规模、高质量、多模态、多场景、多标签情感对话数据集, 实验中仅处理文本模态。该数据集精选 56 部电视剧中的 900 多个对话片段, 每句话都使用六大主流基本情感标签高兴、惊讶、悲伤、愤怒、厌恶、恐惧另加中性情绪标签。

评价指标: 对于 MELD、EmoryNLP 和 M3ED 这 3 个数据集上均采用 weighted-F1 作为评价指标。由于 DailyDialog 中“中性”占据多数, 因此采用 micro-F1 作为评价指标, 并且文献[4-6]在计算结果时忽略“中性”标签。

3.2 实验设置

在语句级特征提取部分, 使用 RoBERTa-Large 预训练模型进行微调, 批处理大小为 32, 训练采用 Adam 优化器, 学习速率 (learning rate, LR) 为 1×10^{-5} 。

DPGAT 在训练过程中, 维度来自 RoBERTa 的特征表示 d_u 为 1 024, 隐藏表示的维度 d_h 为 300, 正则化系数 λ 为 0.3, 注意力头数设置为 6。

3 个数据集的训练参数: 学习速率、Dropout 率、批次大小 (Bs)、图注意力网络层数 (GAT)、多层感知机层数 (MLP), 通过实验得出最优参数记录表见表 2。

表 2 最优参数记录表

数据集	LR	Dropout 率	Bs	GAT	MLP
MELD	1×10^{-4}	0.4	16	2	2
EmoryNLP	1×10^{-4}	0.1	32	1	1
DailyDialog	5×10^{-5}	0.4	32	3	1
M3ED	1×10^{-4}	0.4	16	2	3

每个训练和测试过程都在内存为 32 GB 的 NVIDIA Tesla V100-SXM2 GPU 上运行, 模型实现采用 pytorch 框架, 包含 100 个 epochs, 最终的实验结果均为随机 5 次实验在测试集上的平均表现。

为了研究图注意力网络层数对情感分类的影

响,在M3ED数据集上进行了1~6层的实验。图注意力网络层数的影响如图4所示,结果显示在具有两个层数时效果最佳。

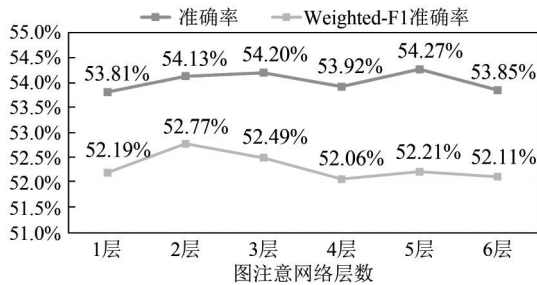


图4 图注意力网络层数的影响

实验结果呈现在某一固定范围内的震荡现象。观察准确率和Weighted-F1分数的趋势:当层数较少时,上下文信息和说话者信息无法得到足够的融合和交换;而当层数较多时,生成冗余或特征重复,导致忽略重要信息,性能反而下降。这一实验结果表明,在图注意网络的层数选择上需要权衡上下文融合和信息重复之间的关系,以实现情感分类性能的最佳提升。

3.3 对比实验

将本文模型和其他已有模型进行如下比较。

基于图和预训练模型:对比实验结果见表3,DisGCN基于对话话语结构构建图,探索上下文和说话者的特定特征;DialogXL设计了4种类型的对话感知自我注意力,使模型能够感知多方对话中的交互;DAG-ERC使用有向无环图来建模说话者和话语之间的交互;CoG-BART基于BART-Large,利用监督对比学习和响应生成辅助任务来区分具有相似情感的话语的语义;SPCL-CL^[20]用监督原型对比学习损失解决类不平衡的问题,并引入课程学习来减轻极端样本的影响;SACL^[20]基于预训练模型提取的特征,建立一个基于序列的监督对抗对比学习框架,用于以监督方式学习类传播结构化表征;denoiseGNN^[21]用一种图结构的形式表示对话,通过上下文过滤器和特征校正提高其判别能力并减少情绪预测错误。

表3 对比实验结果

模型	MELD	EmoryNLP	DailyDialog
DisGCN	64.22	36.38	—
DialogXL	62.41	34.73	54.93
DAG-ERC	63.65	39.02	59.33
CoG-BART	64.81	39.04	56.29
COSMIC [†]	65.21	38.11	58.48
CauAIN [†]	65.46	—	58.21
MUCDN ^{†‡}	65.37	40.09	—
DualGATs ^{†‡}	65.72	39.46	59.27
SPCL+CL ^{°C}	64.93	39.45	—
SACL ^{°C}	66.45	39.65	—
denoiseGNN	66.70	39.73	59.62
DPGAT	66.23	40.03	59.28

注:与ERC的各种方法相比,†表示使用了外部知识;未标注符号的表示结果来自原始论文;°C表示结果来自文献[20]。‡标记的模型是通过使用与论文完全相同的配置和参数,成功在本地服务器上重现,结果为随机进行的5次实验的平均结果,其方差分别为0.01、0.46、0.09。

外部知识注入:在表3中,COSMIC包含不同的常识元素,如心理状态、事件和因果关系,并在此基础上学习参与对话的对话者之间的互动。CauAIN因果感知交互网络借助情感原因检测来彻底理解会话上下文。MUCDN将对话分解为独立的线程执行上下文和说话者特定建模,并通过一个相互模块进行融合。DualGATs模型同时考虑话语结构和讲话者感知上下文的互补方面。

为探究所提模型在不同语言数据上是否普遍适用,在中文数据集M3ED上实验,与其他模型进行效果对比,M3ED实验结果(文本模态)见表4。

DialogueRNN通过不同的GRU捕获全局和说话者特定的时间上下文信息以及全局情绪状态信息;DialogueGCN通过对长距离依赖和说话人交互进行建模;MDI提出了一个通用的多模态对话感知交互框架,涉及多模态融合、全局-局部上下文和说话人交互建模^[19];MMGCN^[22]提出一种最先进的基于GCN的框架,针对文本模态,对全连接图进行建模。



3.4 结果分析

由表3可以看出, DisGCN、DialogXL和DAG-ERC等基于图的方法可以融合和处理更多的节点间信息, 充分利用有效信息进行语境还原, 在对上下文进行全局性建模方面表现出优势, 尤其在对长距离上下文信息的分析和结合方面。基于监督对比学习的SACL模型也表现出明显的优势。在外部信息注入时, 模型整体性能进一步提高, 如COSMIC和CauAIN, 说明外部注入的信息有助于模型对语境的建立和理解; 通过对比MUCDN和DuaGATs的结果, 发现对外部信息不同处理方式会产生不同效果。

最新的模型denoiseGNN也是基于图神经网络, 为了减轻相关性低、信息量小的上下文线索对ERC的影响, 使用一种上下文过滤器来去除语篇间建立依赖关系过程中的噪声信息(去噪), 因而在MELD和DailyDialog数据集上呈现出略微优异的效果, 也证实在对话情感识别过程中噪声信息需要被考虑。

在中文数据集上的实验结果, 表4展现了所提模型的泛化性和适用性, 与基于RNN和GCN的模型对比, 显示了图注意网络在处理信息交互融合的角度具有优势。

表4 M3ED实验结果(文本模态)

模型	M3ED
DialogueGCN*	46.09
MMGCN	46.49
DialogueRNN*	48.80
MDI*	49.42
DPGAT	52.77

注: *表示结果均来自文献[19]; 未标注符号的表示结果来自原始论文。

综上所述, 多个模型的实验结果对比, 证实了基于图结构建立上下文依赖关系和说话者关系以及通过图注意力网络来处理外部信息均是有效的。本文的研究重点是利用解析器对语篇进行解

析作为外部信息注入, 并选择图结构来整合语篇依赖关系以及说话人信息。所提出的DPGAT不仅在说话者内部融合中研究语句之间的关系和影响权重, 还在外部融合中使用图注意力网络捕获语篇依赖关系, 在中、英文数据集上取得了较好的结果, 验证了模型的有效性和泛化性。

3.5 算法复杂度分析

用计算量每秒浮点运算次数(floating-point operations per second, FLOPS)来衡量模型的时间复杂度, 用参数量来衡量模型的空间复杂度。

选取部分模型, 分别计算它们的计算量和参数量, 算法复杂度计算见表5。对比观察表3和表5, 有以下结论。

表5 算法复杂度计算

模型	计算量/ ($\times 10^6$ FLOPS)	参数量/ 10^6 个
MuCDN	1 822.37	5.45
DualGATs	405.02	1.26
SACL	3 788.96	2.37
DPGAT	1 371.33	4.01

与MuCDN模型相比, 本文模型在时间和空间上都有所改善, 并且实验结果也有提升; 与Dual-GATs模型, 本文模型实验结果略有优势, 但在时间、空间上消耗较多; 与SACL模型相比, 虽然实验结果近似, 但是本文模型的计算量明显更低。

结合算法复杂度和实验结果, 本文模型在实验结果上有所提升, 但是在模型的轻量化方面还有待进一步完善, 也进一步给提升模型效率提供了一个新的方向。

3.6 性能分析

由于对话情绪识别任务以对话为单位进行研究, 因此对话中的类别不平衡现象会在一定程度上影响模型最终性能表现。为了研究数据集中类别不平衡对模型性能的影响, 模型分类效果统计如图5所示, 对比了4个数据集在测试集、验证集和测试集中模型针对不同类别的分

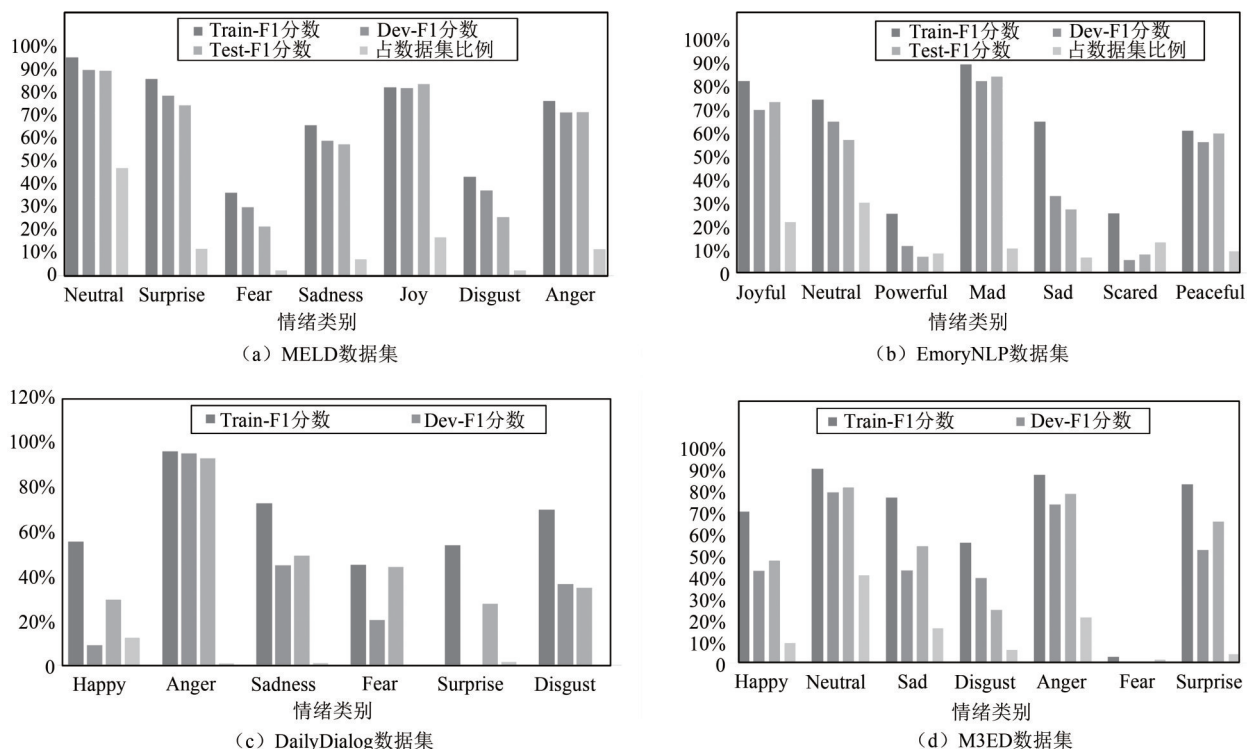


图5 模型分类效果统计

类效果，以及该类别占标签总数的比例，其中MELD、EmoryNLP、M3ED数据集依旧使用Weighted-F1分数作为评价指标，DailyDialog数据集使用Micro-F1分数。

MELD和EmoryNLP数据集的类别分布相对均衡，各个类别的样本数量较多，模型能够深入学习和训练各个类别，整体类别准确率反映了各类别在总体中的占比，展示了模型在分类任务上的良好性能，表明模型在相对均衡的数据集上能够有效地学习并取得优异的分类结果。

M3ED和DailyDialog中存在类别不平衡的情况，对模型在各个类别上的学习和泛化能力产生影响。

在M3ED中，有些类别在特征空间中更容易区分，例如，Fear和Surprise占比相近，但是模型在Surprise类别上表现很好。观察到模型在某些类别上表现优秀，表明模型能够区分特定的类别，即使类别分布相近。

在DailyDialog中，尽管“中性”类别占比较大，但是Anger在训练集、验证集或测试集中的准确率较高，可能是因为模型在学习过程中没有足够的样本来很好地理解这个少数类别，这可能导致模型在实际应用中对这个少数类别的表现不佳。

综合而言，模型在相对均衡的数据集上表现出色；在处理类别不平衡的情况下，仍然能够有效学习和泛化到少数类别，证实了模型在应对不同数据分布下的适应能力。

3.7 消融实验

通过消融研究来验证本文模型中提出的不同模块的有效性，消融实验结果见表6。

内部融合模块主要处理说话人内部特征融合，探究不同说话人对对话语境和说话人情绪状态的实际影响，实验结果表明这一模块是必要的，为后续与外部融合联合预提供了基础。

外部融合模块主要处理上下文建模和说话者建模的融合部分，仅基于初始特征和依赖关系进



表6 消融实验结果

方法	MELD		EmoryNLP		DailyDialog		M3ED	
	准确率	Weighted-F1	准确率	Weighted-F1	准确率	Micro-F1 (-中性)	准确率	Weighted-F1
内部融合	65.86	65.42	43.70	38.64	86.43	56.21	52.43	50.83
外部融合	65.90	65.78	44.11	38.06	85.83	57.19	52.11	50.67
去掉位置编码	65.86	65.79	43.72	38.23	85.87	58.08	53.09	51.58
去掉归一化层	65.82	65.27	43.70	37.55	85.9	58.97	52.22	50.98
DPGAT	66.93	66.23	44.0	40.03	86.16	59.28	54.02	52.77

行实验，结果表明融合上下文信息和说话人感知信息的重要性。

位置编码目的是将离散的说话人距离转化为密集的实数向量，模型可以从数据中学习目标之间的语义关系。因为相似含义的单词在嵌入空间中通常会更接近，所以能够更好地还原说话人之间的互相影响，实验结果显示该模块对于模型有提升效果。

归一化层的主要作用旨在加速模型的收敛、防止梯度消失或梯度爆炸，并提高模型的泛化能力，根据消融效果对比，证明了这一过程的有效性。

4 结束语

针对对话语境建模的全局性和依赖性的问题，提出了一个基于语篇解析和图注意网络的模型，通过使用语篇解析器对语句进行解析得到话语依存关系，构造语篇上下文结构图和说话人关系图，利用图注意网络对上下文建模和说话人建模进行融合，还原对话语境探究说话者情绪状态。在4个公共数据集上进行实验，结果表明这种方法不仅能够全面捕捉对话语境中的信息，而且有效整合了说话者之间的关系，为情感分类任务提供了更丰富、准确的特征表示。

本文提出的模型仍具有一定的不足，例如：在使用语篇解析器处理对话内容时会有一定的误差；在判定最终情绪状态时有某些相似类别难以区分；某些语言内容的情绪指向并不明显。在后续研究中，还需要从这些角度进一步改进。

参考文献：

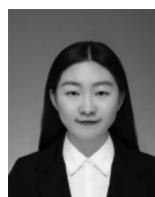
- [1] 王祥宇, 宗成庆. 基于监督对比学习的文本情绪类别表示[J]. 软件学报, 2023: 1-12.
WANG X Y, ZONG C Q. Text emotion category representation based on supervised contrast learning[J]. Journal of Software, 2023: 1-12.
- [2] 赵妍妍, 陆鑫, 赵伟翔, 等. 情感对话技术综述[J]. 软件学报, 2024, 35(3): 1377-1402.
ZHAO Y Y, LU X, ZHAO W X, et al. Survey on emotional dialogue techniques[J]. Journal of Software, 2024, 35(3): 1377-1402.
- [3] FU Y, YUAN S Y, ZHANG C, et al. Emotion recognition in conversations: a survey focusing on context, speaker dependencies, and fusion methods[J]. Electronics, 2023, 12(22): 4714.
- [4] LI S M, YAN H, QIU X P. Contrast and generation make bart a good dialogue emotion recognizer[C]// Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2022: 11002-11010.
- [5] 陶涛, 李珍, 王冀彬, 等. 基于图神经网络的权益推荐技术方案研究[J]. 电信科学, 2023, 39(8): 91-101.
TAO T, LI Z, WANG J B, et al. Research on the graphical convolution neural network based benefits recommendation system strategy[J]. Telecommunications Science, 2023, 39(8): 91-101.
- [6] ZHANG D Z, CHEN F L, CHEN X Y. DualGATs: dual graph attention networks for emotion recognition in conversations[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 7395-7408.
- [7] SHI Z, HUANG M. A deep sequential model for discourse parsing on multi-party dialogues[C]// Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019, 33(1): 7007-7014.

- [8] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB]. 2017:
- [9] 陈佳, 章坚武, 张浙亮. 基于上下文信息与注意力特征的欺骗语音检测[J]. 电信科学, 2023, 39(2): 92-102.
CHEN J, ZHANG J W, ZHANG Z L. Spoof speech detection based on context information and attention feature[J]. Telecommunications Science, 2023, 39(2): 92-102.
- [10] PORIA S, CAMBRIA E, HAZARIKA D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long Papers), Stroudsburg, PA: ACL, 2017: 873-883.
- [11] CHEN S Y, Hsu C C, KUO C C, et al. Emotionlines: An emotion corpus of multi-party conversations[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA). rue Brillat-Savarin, Paris, 2018: 1597-1601.
- [12] PORIA S, HAZARIKA D, MAJUMDER N, et al. MELD: a multimodal multi-party dataset for emotion recognition in conversations[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 527-536.
- [13] LIU Z Y, CHEN N. Improving multi-party dialogue discourse parsing via domain integration[C]//Proceedings of the 2nd Workshop on Computational Approaches to Discourse. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 122-127.
- [14] ZHAO W, ZHAO Y, QIN B. MuCDN: mutual conversational-detachment network for emotion recognition in multi-party conversations[C]//Proceedings of the 29th International Conference on Computational Linguistics. Stroudsburg, PA: ACL, 2022: 7020-7030.
- [15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, Inc., 2017: 1-15.
- [16] 王瑞琴, 黄熠旻, 纪其顺, 等. 注意力感知的边-节点交换图神经网络模型[J]. 电信科学, 2024, 40(1): 106-114.
WANG R Q, HUANG Y M, JI Q S, et al. Attention aware edge-node exchange graph neural network[J]. Telecommunications Science, 2024, 40(1): 106-114.
- [17] ZAHIRI S M, CHOI J D. Emotion detection on tv show transcripts with sequence-based convolutional neural networks[C]//Workshops at the thirty-second AAAI conference on artificial intelligence. Palo Alto, CA: AAAI Press, 2018: 44-51.
- [18] LI Y R, SU H, SHEN X Y, et al. Dailydialog: a manually labelled multi-turn dialogue dataset[C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing, Volume 1 (Long Papers) ACL-IJCNLP 2017. Stroudsburg, PA: ACL, 2017: 986-995.
- [19] ZHAO J M, ZHANG T G, HU J W, et al. M3ED: multi-modal multi-scene multi-label emotional dialogue database[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 1 (Long Papers), Stroudsburg, PA: ACL, 2022: 5699-5710.
- [20] HU D, BAO Y N, WEI L W, et al. Supervised adversarial contrastive learning for emotion recognition in conversations.[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. ACL, 2023: 10835-10852.
- [21] GAN C Q, ZHENG J H, ZHU Q Y, et al. A graph neural network with context filtering and feature correction for conversational emotion recognition[J]. Information Sciences, 2024, 658: 120017.
- [22] HU J W, LIU Y C, ZHAO J M, et al. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Volume 1 (Long Papers), Stroudsburg, PA: ACL, 2021: 5666-5675.

[作者简介]



郝秀兰 (1970-), 女, 博士, 湖州师范学院浙江省现代农业资源智慧管理与应用研究重点实验室副教授、硕士生导师, 主要研究方向为智能信息处理、数据与知识工程、自然语言理解等。



魏少华 (2000-), 女, 湖州师范学院浙江省现代农业资源智慧管理与应用研究重点实验室硕士生, 主要研究方向为自然语言处理、情感分析。



曹乾 (1997-), 男, 中国建设银行股份有限公司湖州分行研究员, 主要研究方向为自然语言处理、数据挖掘等。

张雄涛 (1984-), 男, 博士, 湖州师范学院浙江省现代农业资源智慧管理与应用研究重点实验室副教授、硕士生导师, 主要研究方向为机器学习、深度学习、模糊系统等。