联合多任务学习的对话情感分类和行为识别

刘思进 朱小飞 彭展望
(重庆理工大学计算机科学与工程学院 重庆 400054)

摘 要 对话情感分类和对话行为识别是对话系统中的两个子任务,旨在预测对话中每个语句的情感标签和行为标签.这两个任务受多种因素的影响而密切相关,而现有的模型没有合理利用对话中包含的显式和隐式信息,如说话者信息,时间信息,标签信息等,并且两个相关任务之间缺乏有效的交互.为了解决上述问题,本文提出了一个新的多任务学习模型,即说话者感知跨任务协同交互图网络(Speaker-aware Cross-task Co-interactive Graph Network, SA-CCGN).该模型首先捕捉了说话者随时间变化的情感和行为线索,以生成说话者感知的语句表示,然后通过跨任务协同交互图网络来充分建模对话内的信息传播和任务间的信息交互,其中,通过构建一个有向无环图来模拟一个对话的信息传播,每一次图传播后,使用协同交互层对两个任务进行适当的交互.最后,在解码时引入标签信息,即标签之间的区分度和关联性,对模型训练进行约束.在两个公开数据集上的实验结果表明,该模型相较于目前最先进的联合模型,在对话情感分类任务上的F1值分别提高了4.57%和3.33%,在对话行为识别任务上的F1值分别提高了2.15%和0.63%,而参数量和内存使用降低了约1/2.

关键词 多任务学习;对话系统;情感分类;行为识别
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.01947

Dialogue Sentiment Classification and Act Recognition Based on Multi-Task Learning

LIU Si-Jin ZHU Xiao-Fei PENG Zhan-Wang
(*College of Computer Science and Engineering, Chongqing University of Technology, Chongqing* 400054)

Abstract Currently, social media platforms allow users to universally express their opinions and sentiments due to their convenience and openness. As one of the most common ways of communication, dialogue contains rich information and sentiment expression of participants. Dialogue sentiment classification and dialogue act recognition are two sub-tasks in dialogue systems that aim to predict the sentiment and act label of each utterance in a dialogue. In the past few years, these two tasks gained attention from the NLP community due to the increase of public availability of dialogue data. They can be used to analyze dialogues that take place on social media or other scenes and provide support for downstream tasks, such as dialogue response generation. They can also aid in analyzing dialogues in real times, which can be public opinion monitoring, interviews, psychological consulting and more. These two tasks are influenced by multiple factors and closely related. However, existing models do not make reasonable use of the explicit and implicit information contained in a dialogue, such as speaker information, temporal information, and label information,and simply or coarse-grained modeling the interaction of two tasks. To solve the above problems, this paper proposes a new multi-task learning model, namely Speaker-aware

收稿日期:2022-07-28;在线发布日期:2023-01-18.本课题得到国家自然科学基金项目(No.62141201)、重庆市自然科学基金面上项目(CSTB2022NSCQ-MSX1672)、重庆市教育委员会科学技术研究计划重大项目(No.KJZD-M202201102)资助.**刘思进**,硕士研究生,中国计算机学会(CCF)学生会员,主要研究领域为自然语言处理、情感分析.E-mail:liusijin@2020.cqut.edu.cn.**朱小飞**(**通信作者**),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为自然语言处理、数据挖掘、信息检索.E-mail:zxf@cqut.edu.cn.**彭展望**,硕士研究生,中国计算机学会(CCF)学生会员,主要研究领域为自然语言处理、数据挖掘.

Cross-task Co-interactive Graph Network (SA-CCGN). The model first captures speaker-aware sentiment and act cues along with the time to generate speaker-aware utterance representations, and then adequately models information propagation within a conversation and information interaction between tasks through a cross-task co-interactive graph network, where information propagation of a conversation is modeled by constructing a directed acyclic graph, and after each graph propagation, appropriate interaction between two tasks is performed using the co-interactive layer. Finally, the label information is introduced, i.e., differentiation and correlation between labels, which can constrain the model training when decoding. Specifically, in the multi-loss decoder, the supervised contrastive learning loss is used to make the learned representation of different labels more differentiated and the conditional random field loss is used to constrain the generation of adjacent label sequences, then the final sentiment and act label of each utterance are obtained. In order to prove the effectiveness of the model in this paper, experiments were conducted on the two public two-way dialogue datasets: DailyDialog dataset and Mastodon dataset, and we compare our proposed method with a variety of state-of-the-art methods, including dialogue sentiment classification methods, dialogue act recognition methods and joint-train methods. Experimental results on two public datasets show that our model outperforms the current state-of-the-art joint model Co-GAT, with an improvement of 4.57% and 3.33% in F1 scores for the dialogue sentiment classification task and 2.15% and 0.63% in F1 scores for the dialogue act recognition task on the two datasets, respectively, while reducing the number of parameters and memory usage by about 1/2. The performance of SA-CCGN on two public datasets exceeds the best results in the known literature. Experiments show that this method can effectively utilize dialogue information, and has obvious advantages in dialogue sentiment classification task and dialogue act recognition task compared to previous methods.

Keywords multi-task learning; dialogue system; sentiment classification; act recognition

1 引 言

作为最常见的交流方式之一,对话包含了参与者丰富的信息和情感表达.对话情感分类(Dialogue Sentiment Classification, DSC)和对话行为识别(Dialogue Act Recognition, DAR)是对话系统中的两个具有挑战性的任务[1-2].DSC 旨在预测对话中每个语句的情感标签(如积极、消极、中性等),而 DAR 旨在预测每个语句的行为标签(如同意、询问、陈述等),这有助于对话系统生成适当的回复.最近研究人员发现,这两项任务密切相关,即它们可以通过共同执行而相互促进[3-4].

表 1 展示了一个选自数据集的对话样本片段,每个语句都有相应的情感和行为标签.直观地说,这两个任务是密切相关的,一个任务的信息可以被另一个任务所利用.当预测语句 u_2 的情感时,除其自身语义信息外,它的行为标签“不同意”和前一个语句 u_1 的情感标签“消极”都可以提供有用的参考.这与人类在做出推理的思维类似,即“不同意”的行为标签表明此时说话者 Y 对于上一句话 u_1 持有相反的观点,因此倾向于正面的情感.同样地,两句话之间的相反情感有助于来推断语句 u_2 和 u_3 的行为标签为“不同意”.此外,捕捉同一说话者的情感和行为线索也有助于最后的预测.例如在这个例子中,说话者 X 一直是偏向于消极的情感倾向.

在早期的工作中,Cerisara 等人[3]首先提出了一个多任务学习框架来联合建模这两个任务,其中两个任务共享一个编码器,以此来隐式建模两个任务的相关性.然而简单的多任务学习框架只是通过共享潜在表示来隐式地建模任务之间的联系,无法取得理想的结果,甚至低于一些独立建模两个任务的工作.而 Kim 等人[4]将对话行为、谓词和情感的识别整合到统一的模型中,明确建模两个任务之间的相互作用,但他们的框架仅考虑当前的语句.Qin

等人[5]提出了基于流水线的方法,即利用层次编码器得到情感和行为表示之后,使用一个堆叠的交互关系层将这两种类型的信息结合起来.上述模型的缺点是未能在对话中充分整合上下文信息,而 Li 等人[6]提出了一种上下文感知的动态卷积网络来捕获关键的局部上下文.最近,Qin 等人[7]又提出了一个交互图框架,其中相同任务内语句连接和不同任务间语句连接的全连通图被构造并迭代更新,实现了在一个统一的体系结构中同时建模上下文信息和交互信息.

表 1　数据集的对话样本片段

序号	说话者	语句	情感标签	行为标签
u_1	X	So instead of the nice predictable wlan0 and eth0 we have some completely random device names. Thanks systemd.	Negative	Statement
u_2	Y	… But you aren't supposed to have to know about such details, since everything is being automated for your convenience by the system (d).	Positive	Disagreement
u_3	X	I can't decide if your message is first degree …	Negative	Disagreement

如图 1 所示,以预测语句 u_3 的情感标签为例,子图(a)(b)均为以前的方法简化示意图,子图(a)是基于上下文的方法,此类方法根据上下文的文本语义学习语句表示 e_*,在建模时不区分情感表示和行为表示,也不考虑两个任务之间的交互,最后直接将语句的表示分别映射到情感和行为标签上;子图(b)是基于图交互的方法,s_* 和 a_* 分别代表情感和行为表示,此类方法建模时区分了情感表示和行为表示,但是交互时不区分说话者身份,也不考虑对话发生的时间顺序.而子图(c)是本文方法的示意图,在建模时不仅考虑了对话发生的时间顺序和说话者身份,也对情感表示和行为表示进行了适当交互.

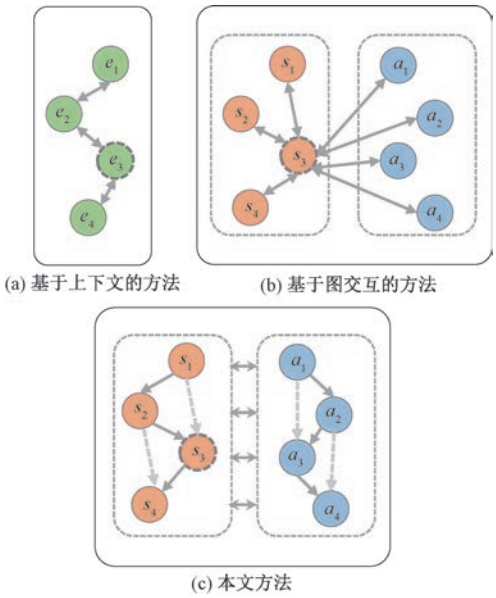

图 1　联合多任务学习的 DSC 和 DAR 方法简化示意图

因此,尽管以往的工作产生了显著的效果,但是他们还存在一些缺陷:(1)没有按照对话时间顺序建模说话者的信息.例如在 Co-GAT[7]中,说话者感知的编码器只是简单地构建了一个连接相同说话者的无向图再进行图传播,而没有考虑时间顺序.其他的一些研究工作[3-6]并没有单独建模说话者的情感和行为线索.(2)对整个对话的上下文建模不充分.Co-GAT 将图注意力网络(GAT)[8]应用于一个无向的图,该图是一个全连通图,其无法区分相同还是不同说话者之间的互动关系.(3)现有研究工作主要关注建模标签之间的关联性,而忽略了标签之间的区分度,对于语义相似但不同的情感类别,如“愤怒”和“厌恶”等,模型很难对其进行区分.

针对上述问题,本文提出了一种用于对话情感分类和对话行为识别的说话者感知跨任务协同交互图网络(SA-CCGN).它由四个主要模块组成:(1)语句编码器;(2)说话者感知交互层;(3)跨任务协同交互图网络;(4)多损失解码器.首先,语句编码器利用 Bi-LSTM 获取对话中语句的文本特征.然后在说话者感知交互层,通过对同一个说话者表述的语句之间应用一个单独的 Bi-LSTM 来捕捉说话者随时间变化的情感和行为线索.在跨任务协同交互图网络中,将语句作为图的节点,通过构建一个有向无环图来模拟一个对话的信息传播,并且使用关系感知特征转换来区分不同说话者之间的互动,充分建模了整个对话的上下文信息.每一次图传播后,使用协同交互层对两个任务进行适当的交互.最后,利用多损失解码器的监督对比学习(Supervised Contrastive Learning, SCL)损失来使得学习到的不同标签的表示更具有区分度,条件随机场(Conditional Random Field, CRF)损失来约束相邻标签序列的产生,得到每个语句最后的情感和行为标签.

综上,本文主要做出了以下贡献:

(1)使用说话者感知交互层捕捉了带有时间顺序的说话人情感和行为线索,以此学习到说话者感

知的语句表示.

(2)提出了跨任务协同交互图网络,充分建模了整个对话的上下文信息,利用有向无环图模拟对话信息传播,并使用协同交互层对两个任务进行适当交互.

(3)考虑了标签之间的区分度和关联性,在解码时引入标签信息对模型训练进行约束,使用 SCL 损失来让学习到的不同标签的表示更具有区分度,CRF 损失来约束相邻标签序列的产生.

(4)两个公开数据集上的实验表明,本文提出的模型优于现有基线模型,取得了先进的性能.

2 相关工作

2.1 对话情感分类

随着对话数据集的发展,对话情感分类在 NLP 界获得了越来越多的关注. 与普通的句子/话语情感分类[9-10]不同,对话情感分类在理想情况下需要对单个话语进行上下文建模. 早期的研究将对话中的语料视作一个句子级的序列,通过联合句子级别的上下文来帮助判定当前句子的情感,用基于递归的网络对整个对话进行建模. Poria[11]提出了一个基于 LSTM 的网络来捕捉对话中的语境信息. DialogueRNN[12]通过跟踪说话者状态和对话中的情感动态,用多个 GRU 来建模语句情感. COSMIC[13]采用了 DialogueRNN 相似的架构,并且结合常识知识来增强对话中每个语句的表示. 另一系列的工作采用基于图的模型来捕捉更复杂的对话结构信息. DialogueGCN[14]使用图卷积网络来处理由对话中参与者的自我和说话人之间的依赖关系构建的图. RGAT[15]是一个关系感知的图注意力网络,包含了关系位置编码以建模说话者的依赖关系和语句顺序. DAG-ERC[16]提出用有向无环图(DAG)建模对话语境的新思路,它应用一个新颖的有向无环图网络来为每个语句收集远程和邻近的语句,并进行关系感知的特征转换.

2.2 对话行为识别

对话行为分类任务是探究对话动因中关键的组成部分之一,它通过作为对话系统的自然语言理解舵手来满足对话系统的要求. 早期的研究如 Reithinger 和 Klesen[17]以及 Grau 等人[18]把重点放在词汇、句法和韵律特征的分类上. 在另一项工作中,Ortega 等人[19]采用了 CNNs[20]和 CRFs[21]来建模对话. Lee 等人[22]提出了一种基于 CNNs[20]和 RNNs[23]的方法,利用以前的语句来预测当前语句的对话行为. Raheja 等人[24]提出的模型在 RNN 上使用自注意力机制. Chen 等人[25]提出了一个 CRF 注意结构网络,以利用结构化注意机制捕捉长期的语境依赖. 张志昌等人[26]提出了基于独立循环神经网络(Independently Recurrent Neural Network,IndRNN)和词级别注意力融合的用户意图分类方法. 周俊佐等人[27]提出一种混合神经网络模型,综合利用多个深度网络模型的多样性输出.

2.3 联合模型

考虑到对话情感分类和行为识别两个任务之间的相关性,许多联合模型被提出来考虑这两个任务之间的互动. Cerisara 等人[3]首次探索了多任务框架来模拟这两项任务之间的相关性,其中两个任务共享一个编码器,以此来隐式建模两个任务的相关性. Kim 等人[4]提出了一个集成的神经网络,用于识别对话行为、谓词和对话的情感. 近年来,Qin 等人[5]提出了一个深度关系网络,采用一个堆叠的交互层来明确地整合对话情感与行为. Li 等人[6]基于动态卷积网络提出了一种上下文感知的动态卷积网络,以便在生成卷积核时更好地捕捉关键的上下文,并将框架扩展为用于对话情感分类和行为识别的双通道版本. Qin 等人[7]提出了一个新颖的 Co-GAT 模型,该模型将图注意力网络应用于一个完全连通的无向图上,该无向图由两种类别的连接组成,分别是任务内的语句连接和任务间语句连接.

综上,以上大多数基于深度学习的方法都缺乏对说话者信息和标签信息的有效利用,而本文的方法可以更好地捕捉说话者随时间变化的线索,并且使用标签信息更好地约束模型的训练. 另外,当联合模型开始探索两个任务的交互时,采用的多是隐式建模或者粗粒度的建模,而本文提出的模型构建了更为复杂的有向无环图来建模丰富的对话信息,并采用协同交互层对两个任务进行合理的交互.

3 模 型

3.1 问题定义

给定一个由 N 个语句组成的对话 U,$U=\{u_1,u_2,\cdots,u_N\}$,其中 N 是语句的数量. 对于每个语句,原始输入是一个单词序列,即 $u_i=\{w_{i,1},w_{i,2},\cdots,w_{i,n}\}$,其中 n 表示语句的长度. $Y^s=\{y_1^s,y_2^s,\cdots,y_N^s\}$ 和 $Y^a=\{y_1^a,y_2^a,\cdots,y_N^a\}$ 分别是对话 U 对应的情感和行为标签序列,其中 $y_i^s\in\gamma^s$,$y_i^a\in\gamma^a$ 是语句

u_i 的情感和行为标签，γ^s 和 γ^a 表示情感和行为标签集. 在一个对话中，有 2 个独立的说话者 $P=\{p_X,p_Y\}$，其中 X、Y 代表不同说话者. $p_{\phi(u_i)}$ 表示 u_i 的对应说话者，其中 $\phi(u_i)\in\{X,Y\}$ 是一个映射函数，它将 u_i 映射到其对应说话者的索引.

模型的目标是在根据对话的上下文信息及说话者信息从预先定义的情绪标签集合 γ^s 和行为标签集合 γ^a 中预测每个语句 u_i 的情绪标签 y_i^s 和行为标签 y_i^a. 本文提出的模型 SA-CCGN 如图 2 所示.

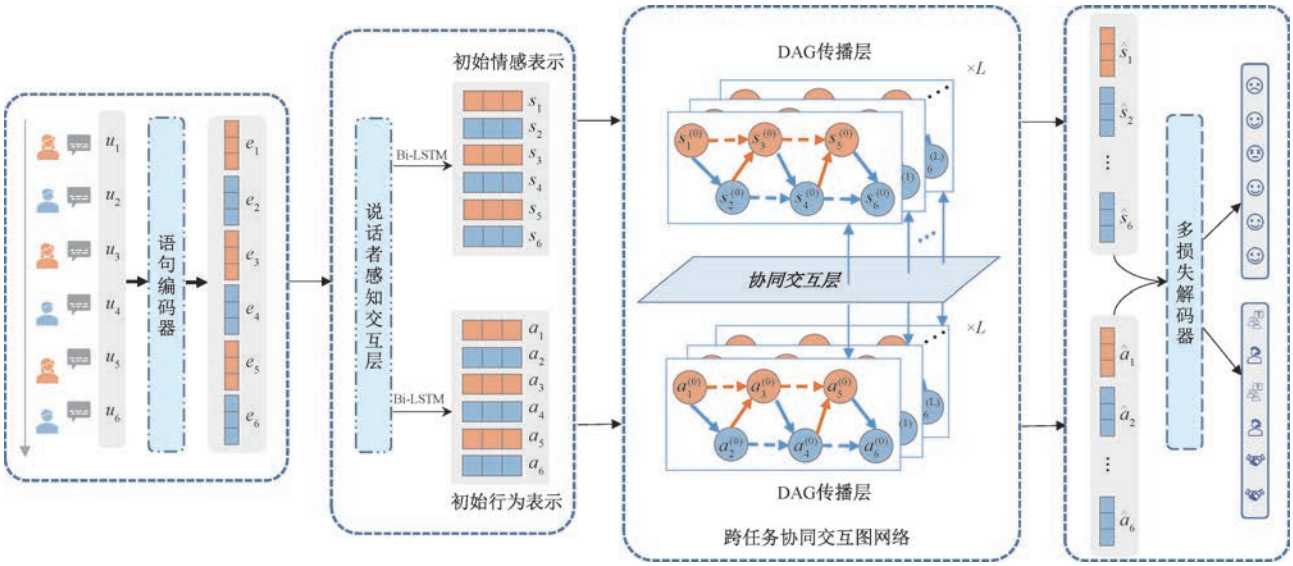

图 2　SA-CCGN 模型的总体框架图

3.2　语句编码器

与最新模型类似[7]，语句编码器采用了 Bi-LSTM 用于提取与对话上下文无关的语句级特征向量. 具体来说，对于对话 U 中的每个语句 $u_i=\{w_{i,1},w_{i,2},\cdots,w_{i,n}\}$，$u_i$ 是由 n 个单词构成的单词序列. 编码器首先采用随机初始化的方式生成每个单词的向量表示，即嵌入函数 $\phi^{emb}(\cdot)$ 将 u_i 中的每一个单词 $w_{i,j}$ 映射到向量表示，然后应用 Bi-LSTM，对于每个语句 u_i 的向量表示序列，分别在前向和后向两个方向生成每个单词的隐藏状态，公式如下：

$$\overrightarrow{\boldsymbol{h}_{i,j}}=\overrightarrow{\mathrm{LSTM}}(\phi^{emb}(w_{i,j}),\overrightarrow{\boldsymbol{h}_{i,j-1}}) \tag{1}$$

$$\overleftarrow{\boldsymbol{h}_{i,j}}=\overleftarrow{\mathrm{LSTM}}(\phi^{emb}(w_{i,j}),\overleftarrow{\boldsymbol{h}_{i,j+1}}) \tag{2}$$

$$\boldsymbol{h}_{i,j}=[\overrightarrow{\boldsymbol{h}_{i,j}}\|\overleftarrow{\boldsymbol{h}_{i,j}}] \tag{3}$$

其中，$\|$ 表示拼接操作，$\overrightarrow{\boldsymbol{h}_{i,j}}$ 和 $\overleftarrow{\boldsymbol{h}_{i,j}}$ 分别是第 i 句话中第 j 个单词的前向和后向的隐藏状态.

经过双向编码之后，我们将第 j 个单词前向和后向的隐藏状态拼接起来，作为其表示 $\boldsymbol{h}_{i,j}$，因此，对于语句 $u_i=\{w_{i,1},w_{i,2},\cdots,w_{i,n}\}$ 中每个单词拼接其双向的隐藏状态，得到编码后的 u_i 的词向量序列 $\{\boldsymbol{h}_{i,1},\cdots,\boldsymbol{h}_{i,n}\}$.

对 u_i 中的所有单词表示使用平均池化获得 u_i 的语句总体表示 $\boldsymbol{e}_i$，公式如下：

$$\boldsymbol{e}_i=avg\_pool(\boldsymbol{h}_{i,1},\cdots,\boldsymbol{h}_{i,n}) \tag{4}$$

经过语句编码器编码后，得到对话中所有语句的初始的表示，记为 $\boldsymbol{E}=\{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_N\}$.

3.3　说话者感知交互层

为了更好地捕捉说话者信息，使用说话者感知交互层来使其随着对话顺序交互，这使得模型能够更好理解同一说话者的情绪和行为线索如何随着时间变化.

具体来说，对每个说话者使用一个单独的 Bi-LSTM. 形式上，将说话者 X 的发出的所有语句序列表示为 $U^X=\{u_1^X,\cdots,u_{L_X}^X\}$，说话者 Y 的所有语句序列表示为 $U^Y=\{u_1^Y,\cdots,u_{L_Y}^Y\}$，其中 L_X 和 L_Y 分别是这两个序列包含的语句数(即 $L_X+L_Y=N$). 设 $\boldsymbol{E}^X=\{\boldsymbol{e}_{\psi(u_1^X)},\cdots,\boldsymbol{e}_{\psi(u_{L_X}^X)}\}$ 是由语句编码器学习的说话者 X 对应语句表示的序列，其中 $\psi(u_i^X)$ 是一个映射函数，它将语句 u_i^X 映射到对话 U 中对应语句的索引. 然后 $\boldsymbol{E}^X$ 被送入 Bi-LSTM 层：

$$\overrightarrow{\boldsymbol{p}_{\psi(u_j^X)}}=\overrightarrow{\mathrm{LSTM}}(\boldsymbol{e}_{\psi(u_j^X)},\overrightarrow{\boldsymbol{p}_{\psi(u_{j-1}^X)}}) \tag{5}$$

$$\overleftarrow{\boldsymbol{p}_{\psi(u_j^X)}}=\overleftarrow{\mathrm{LSTM}}(\boldsymbol{e}_{\psi(u_j^X)},\overleftarrow{\boldsymbol{p}_{\psi(u_{j+1}^X)}}) \tag{6}$$

其中，$\overrightarrow{\boldsymbol{p}_{\psi(u_j^X)}}$ 和 $\overleftarrow{\boldsymbol{p}_{\psi(u_j^X)}}$ 分别是说话者 X 第 j 句话前向和后向的语句表示.

与上面类似，将说话者 X 的第 j 句话最后的语句表示为二者的拼接，公式如下：

$$\boldsymbol{p}_j^X=[\overrightarrow{\boldsymbol{p}_{\psi(u_j^X)}}\|\overleftarrow{\boldsymbol{p}_{\psi(u_j^X)}}] \tag{7}$$

对于另一个说话者 Y，采取与上述相同的操作，得到两个说话者对应的说话者感知的语句表示 $\boldsymbol{P}^X=\{\boldsymbol{p}_1^X,\cdots,\boldsymbol{p}_{L_X}^X\}$ 及 $\boldsymbol{P}^Y=\{\boldsymbol{p}_1^Y,\cdots,\boldsymbol{p}_{L_Y}^Y\}$. 再将其按对话中原语句序列的顺序映射回去，表示为 $\boldsymbol{P}=\{\boldsymbol{p}_1,\cdots,\boldsymbol{p}_N\}$. 继 Qin 等人[7]之后，为了获得任务特定的语句表示，我们分别在 $\boldsymbol{P}$ 上应用两个单独的 Bi-LSTM 来获取情感和行为的语句表示，以使其更具体于任务. 这整个过程可以简单表述为 $\boldsymbol{S}=\text{Bi}-\text{LSTM}_s(\boldsymbol{P})$ 和 $\boldsymbol{A}=\text{Bi}-\text{LSTM}_a(\boldsymbol{P})$，其中 $\boldsymbol{S}=\{\boldsymbol{s}_1,\cdots,\boldsymbol{s}_N\}$ 和 $\boldsymbol{A}=\{\boldsymbol{a}_1,\cdots,\boldsymbol{a}_N\}$ 可以被视为对话情感和对话行为的初始表示序列.

3.4 跨任务协同交互图网络

在得到说话者感知的语句表示之后，本文设计了一个跨任务协同交互图网络，将语句作为图的节点，通过构建一个有向无环图来模拟一个对话中信息的传播. 每一次图传播后，使用协同交互层对两个任务进行交互.

3.4.1 DAG 传播层

首先，应用有向无环图(DAG)来模拟对话中的信息传播. 形式上，将 DAG 表示为 $\mathcal{G}=(V,\varepsilon,\mathcal{R})$. DAG 中的节点是对话中的语句，即 $V=\{u_1,\cdots,u_N\}$. 边代表语句之间的信息传播，例如，$(i,j,r_{ij}\in\varepsilon)$ 表示信息从 u_i 传播到 u_j，边关系类型为 $r_{ij}\in\mathcal{R}$，其中 $\mathcal{R}=\{0,1\}$ 是边的关系类型集. 如果两个相连的语句 u_i 和 u_j 由同一说话者说出，则 $r_{ij}=1$；如果两个语句由不同说话者说出，则 $r_{ij}=0$.

在对话中，信息按时间顺序在说话者的互动中流动，DAG 的构造应该模拟对话中的信息传播. 特别地，本文考虑了三个约束来决定在 DAG 中何时连接两个语句.

约束 1(有向性). 信息只能从先前的语句传播到未来的语句，即 $\forall j>i,j,i,r_{ji}\notin\varepsilon$. 此约束确保对话是有向无环图；

约束 2(远程信息). 对于每个语句 u_i(第一个除外)，其远程信息被定义为语句 u_τ，其中 u_τ 表示与 u_i 相同的说话者所说的前一个语句，即 $\exists\tau<i$，$p_{\phi(u_\tau)}=p_{\phi(u_i)},\tau,i,r_{\tau i}\in\varepsilon$ and $\forall j<\tau,j,i,r_{ji}\notin\varepsilon$. 它假设 u_τ 包含应该传播到 u_i 的远程信息. 此远程约束表示 u_τ 是远程信息的截止点；

约束 3(局部信息). 考虑在第二个约束中定义的 u_τ 和 u_i，u_τ 和 u_i 之间的所有语句包含局部信息，该信息应该传播到 u_i，即 $\forall l,\ \tau<l<i,\ l,i,r_{li}\in\varepsilon$. 局部约束给出局部信息分界点.

在每个 DAG 传播层，由于信息随时间流动，需要依次计算从第一个到最后一个语句的所有语句的隐藏状态. 对于每个语句 u_i，以情感分类任务为例，使用 u_i 在 $(l-1)$ 层的隐藏状态和 u_i 在 l 层的前驱 u_j 的隐藏状态来计算 u_i 与其前驱 u_j 之间第 l 层的注意力权重 $\alpha_{ij}^{(l)}$：

$$\alpha_{ij}^{(l)}=\text{Softmax}_{j\in\mathcal{N}_i}(\boldsymbol{W}_a^{(l)}[\boldsymbol{s}_j^{(l)}\|\boldsymbol{s}_i^{(l-1)}]) \tag{8}$$

其中，$\boldsymbol{W}_a^{(l)}$ 是可训练参数，$\|$ 代表拼接操作，$\mathcal{N}_i$ 表示 u_i 的前驱集合. 使用 $\boldsymbol{s}_i$ 来初始化第 0 层语句节点的情感表示 $\boldsymbol{s}_i^{(0)}$.

此外，本文还引入关系感知特征转换来对不同关系类型的边进行建模，并在第 l 层获得 u_i 的聚合表示 $\boldsymbol{m}_i^{(l)}$，如下所示：

$$\boldsymbol{m}_i^{(l)}=\sum\nolimits_{j\in\mathcal{N}_i}\alpha_{ij}\boldsymbol{W}_{r_{ij}}^{(l)}\boldsymbol{s}_j^{(l)} \tag{9}$$

其中，$\boldsymbol{W}_{r_{ij}}^{(l)}\in\{\boldsymbol{W}_0^{(l)},\boldsymbol{W}_1^{(l)}\}$ 是关系转换的可训练参数，去学习不同边类型的特征.

得到每个句子 u_i 所需要的聚合表示 $\boldsymbol{m}_i^{(l)}$ 之后，应用门控递归单元(GRU)将其与 u_i 在 $(l-1)$ 层的隐藏状态 $\boldsymbol{s}_i^{(l-1)}$ 合并，获得 u_i 在第 l 层的情感节点特征表示 $\tilde{\boldsymbol{s}}_i^{(l)}$.

$$\tilde{\boldsymbol{s}}_i^{(l)}=\text{GRU}_s^{(l)}(\boldsymbol{s}_i^{(l-1)},\boldsymbol{m}_i^{(l)}) \tag{10}$$

上面的 GRU 层利用 $\boldsymbol{m}_i^{(l)}$ 来控制 u_i 在隐藏状态 $\boldsymbol{s}_i^{(l-1)}$ 的传播. 类似地，为了让 $\boldsymbol{s}_i^{(l-1)}$ 控制 $\boldsymbol{m}_i^{(l)}$ 的传播，使用另一个 GRU 层，其公式如下：

$$\boldsymbol{t}_i^{(l)}=\text{GRU}_m^{(l)}(\boldsymbol{m}_i^{(l)},\boldsymbol{s}_i^{(l-1)}) \tag{11}$$

将每个节点 u_i 每一层的两种信息通过加和和拼接进行融合，得到每个节点在第 l 层的最后情感表示 $\boldsymbol{s}_i^{(l)}$.

$$\boldsymbol{s}_i^{(l)}=\tilde{\boldsymbol{s}}_i^{(l)}+\boldsymbol{t}_i^{(l)} \tag{12}$$

类似地，对于对话行为识别任务，也得到每个语句节点在第 l 层的最后行为表示 $\boldsymbol{a}_i^{(l)}$.

3.4.2 协同交互层

为了使得情感分类和行为识别两个任务进行充分交互，互相促进. 使用协同交互层在每个 DAG 传播层传播之后对两个任务的表示进行交互. 具体来说，使用门控机制[28]来确定两种表示的融合比例. 在上一小节，已经得到在 l 层的情感表示分别为 $\boldsymbol{s}_i^{(l)}$ 和 $\boldsymbol{a}_i^{(l)}$.

$$gate_s=\sigma(\boldsymbol{W}_s[\boldsymbol{s}_i^{(l)}\|\boldsymbol{a}_i^{(l)}]) \tag{13}$$

$$gate_a=\sigma(\boldsymbol{W}_a[\boldsymbol{s}_i^{(l)}\|\boldsymbol{a}_i^{(l)}]) \tag{14}$$

$$\bar{\boldsymbol{s}}_i^{(l+1)}=gate_s\odot\boldsymbol{s}_i^{(l)}+(1-gate_s)\odot\boldsymbol{a}_i^{(l)} \tag{15}$$

$$\bar{\boldsymbol{a}}_i^{(l+1)}=gate_a\odot\boldsymbol{a}_i^{(l)}+(1-gate_a)\odot\boldsymbol{s}_i^{(l)} \tag{16}$$

其中 $gate_s$ 和 $gate_a$ 是分别控制情感表示和行为表示两边的融合比例的参数，$\boldsymbol{W}_s$ 和 $\boldsymbol{W}_a$ 是可训练参数，σ 是 sigmoid 函数，$\odot$ 是哈达玛积操作.

此时得到交互后的每个节点的表示. 而每次 DAG 传播层的上一层节点表示，使用的是经过协同交互后的新的节点表示，即将之前的公式(8)，(9)和(11)更新为

$$\alpha_{ij}^{(l)}=\mathrm{Softmax}_{j\in\mathcal{N}_i}(\boldsymbol{W}_a^{(l)}[\boldsymbol{s}_j^{(l)}\|\bar{\boldsymbol{s}}_i^{(l-1)}]) \tag{17}$$

$$\tilde{\boldsymbol{s}}_i^{(l)}=\mathrm{GRU}_s^{(l)}(\bar{\boldsymbol{s}}_i^{(l-1)},\boldsymbol{m}_i^{(l)}) \tag{18}$$

$$\boldsymbol{t}_i^{(l)}=\mathrm{GRU}_m^{(l)}(\boldsymbol{m}_i^{(l)},\bar{\boldsymbol{s}}_i^{(l-1)}) \tag{19}$$

在经过 (l) 层的 DAG 传播层和协同交互层之后，对于每个任务，将所有层的表示拼接在一起，作为最后的节点特征表示，公式如下：

$$\hat{\boldsymbol{s}}_i=\|_{l=0}^{L}\boldsymbol{s}_i^{(l)} \tag{20}$$

$$\hat{\boldsymbol{a}}_i=\|_{l=0}^{L}\boldsymbol{a}_i^{(l)} \tag{21}$$

其中，L 是 DAG 传播层的总层数，$\|$ 是拼接操作.

得到最后的情感表示序列和行为表示序列 $\hat{\boldsymbol{S}}=\{\hat{\boldsymbol{s}}_1,\cdots,\hat{\boldsymbol{s}}_N\}$ 和 $\hat{\boldsymbol{A}}=\{\hat{\boldsymbol{a}}_1,\cdots,\hat{\boldsymbol{a}}_N\}$.

3.5　多损失解码器

对于最后的标签解码，传统的交叉熵损失函数仅考虑了分类模型预测标签的准确性，而忽略了样本的标签之间的区分度以及标签之前存在的关联. 因此，为了充分挖掘样本标签信息，本文采用交叉熵损失、SCL 损失和 CRF 损失这 3 种损失函数联合训练的方式，对模型的训练过程进行约束. 在减小分类误差的同时，约束样本的类间距离和类内距离，并通过标签之间的关联性对整个输出标签序列进行约束. 接下来，以情感分类任务为例，对多损失解码器进行阐述.

首先，本文采用监督对比学习(SCL)[29]以缓解相似标签分类的困难，充分利用标签信息，使相同情绪的样本具有内聚性，不同情绪的样本相互排斥. SCL 将同批次(batch)中所有具有相同标签的样本视为正样本，不同标签的样本视作负样本. 如果批次中某一类别只存在一个样本，它不能直接应用于计算损失，因此对情感表示复制一份副本，其梯度被分离，此时参数优化保持稳定. 对于一批 N 个训练样本的批次，通过上述操作以获得 $2N$ 个样本，SCL 损失可表示如下：

$$X=[\hat{S},\hat{S}_{copy}] \tag{22}$$

$$\mathcal{L}_{scl}^{s}=\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}SIM(p,i) \tag{23}$$

$$SIM(p,i)=\log\frac{exp(\mathrm{sim}(X_i,X_p)/\tau)}{\sum_{a\in A(i)}exp(sim(X_i,X_a)/\tau)} \tag{24}$$

其中，$X\in\mathbb{R}^{2N\times d}$，表示一个双视图批次里所有的样本表示，$i\in I=\{1,2,\cdots,2N\}$ 表示在一个双视图批次里的样本索引. $\tau\in\mathcal{R}^{+}$，表示控制样本之间距离的温度系数，$P(i)=I_{j=i}-\{i\}$ 表示除其自身以外与 i 类别相同的样本，$A(i)=I-\{i,N+i\}$ 表示双视图批次中除其自身以外的样本. $sim()$ 表示余弦相似度函数.

其次，随着对话信息的流动，可以将 DSC 视为序列标记任务，即顺序解码对话中语句的情感标签. 由于语句标签之间有很强相关性，因而在最终表示 $\hat{\boldsymbol{S}}=\{\hat{\boldsymbol{s}}_1,\cdots,\hat{\boldsymbol{s}}_N\}$ 之上使用 CRF 层[21]来预测语句情感标签.

给定句子序列 $U=\{u_1,u_2,\cdots,u_N\}$ 及其相应的真实情感标签序列 $Y^s=\{y_1^s,\cdots,y_N^s\}$ 和所有有效的情感标签序列 γ_s. 对于对话 U，利用情感特征表示 $\hat{\boldsymbol{S}}$，可以获得 $\hat{\boldsymbol{s}}_i$ 对应标签 y_i^s 的分数 $\boldsymbol{F}_{i,y_i^s}$. 标签序列 Y^s 的分数计算如下：

$$score(\hat{\boldsymbol{S}},Y^s)=\sum_{i=1}^{N}(\boldsymbol{A}_{y_{i-1}^s,y_i^s}+\boldsymbol{F}_{i,y_i^s}) \tag{25}$$

其中，$\boldsymbol{A}_{y_{i-1}^s,y_i^s}$ 指的是标签 y_{i-1}^s 到 y_i^s 的转移分数，$\boldsymbol{F}_{i,y_i^s}$ 是标签 y_i^s 在情感表示 $\hat{\boldsymbol{S}}$ 下的发射分数.

给定句子 U，标签序列 Y^s 的条件概率如下所示：

$$p(Y^s|\hat{\boldsymbol{S}})=\frac{e^{score(\hat{\boldsymbol{s}},Y^s)}}{\sum_{y'\in\gamma_s}e^{score(\hat{\boldsymbol{s}},y')}} \tag{26}$$

对于 CRF，最小化对数似然损失来优化模型，损失函数定义为

$$\mathcal{L}_{crf}^{s}=-\ln(p(Y^s|\hat{\boldsymbol{S}})) \tag{27}$$

而对于交叉熵损失，根据最终的表示 $\hat{S}$，可以得到 N 个节点的情感标签预测 $\hat{y}s=\{\hat{y}s_1,\cdots,\hat{y}_N^s\}$，公式如下：

$$p_i^s=softmax(\boldsymbol{W}_f\hat{\boldsymbol{s}}_i+\boldsymbol{b}_f) \tag{28}$$

$$\hat{y}_i^s=argmax(p_i^s) \tag{29}$$

其中 $\boldsymbol{W}_f$ 和 $\boldsymbol{b}_f$ 是可训练参数，p_i^s 是标签的概率分布.

交叉熵损失可以定义为

$$\mathcal{L}_{ce}^{s}=-\sum_{i=1}^{N}\sum_{c=1}^{C}y_{i,c}^s\cdot\ln p_{i,c}^s \tag{30}$$

同理，可以得到对话行为识别任务的三种损失，

即 $\mathcal{L}_{ce}^{a}$,$\mathcal{L}_{scl}^{a}$,$\mathcal{L}_{crf}^{a}$. 最后的损失定义为对话情感分类和对话行为识别两个任务的三种损失的加权和:

$$\mathcal{L}=\mathcal{L}_{ce}^{s}+\mathcal{L}_{ce}^{a}+\alpha\,\mathcal{L}_{scl}^{s}+\beta\,\mathcal{L}_{scl}^{a}+\gamma\,\mathcal{L}_{crf}^{s}+\delta\,\mathcal{L}_{crf}^{a} \tag{31}$$

其中 α、β、γ、δ 四个超参数,分别控制两个任务对应的 SCL 损失和 CRF 损失的权重.

4 实 验

4.1 实施细节

本文使用验证集来调整超参数,并在训练期间使用 AdamW 优化器. 可调超参数包括学习率、批量大小、随机失活率,DAG 传播层的数量和损失权重. 对于其他超参数,语句的隐藏表示设置为 300 维. 在 Mastodon 数据集上,学习率为 3e-5,批量大小为 8,随机失活率为 0.5,DAG 传播层层数为 1,损失权重 α,β,γ,δ 为别为 0.5,0.2,0.3,0.3,迭代次数为 200. 在 DailyDialog 数据集上,学习率为 2e-5,批量大小为 64,随机失活率为 0.5,DAG 传播层层数为 3,损失权重 α,β,γ,δ 为别为 0.7,0.2,0.2,0.1.

所有呈现的结果均为 5 次运行的平均值,实验是在 Intel 核心 CPU I7-9700K 3.6 GHz 和 NVIDIA GeForce GTX 2080TI 的硬件上进行的. 为了使本文的实验可复现,我们的代码和数据将公布在 https://github.com/Sydneylsj/SA-CCGN 上.

4.2 数据集

本文在两个公开数据集上进行实验:Mastodon[3] 和 DailyDialog[30].

Mastodon 是一个对源自于 Mastodon 社交网络的英语对话进行注释的数据集. 从 Mastodon 实例中抓取了大约 80w 篇帖子,并自动筛选出非英语帖子,然后按照回复链接将所有帖子组织成对话树,两名拥有语言学硕士学位且英语流利的学生分别为对话中的每个帖子分配了两个标签. 一个是情感标签,分为积极、消极和中性 3 个类别,另一个是行为标签,共有 15 类,如声明、同意、请求等.

DailyDialog 是从英语学习者的日常交流中收集的双向对话数据集. 它包含 7 种情绪:中性(none)、愤怒(anger)、厌恶(disgust)、恐惧(fear)、快乐(happiness)、悲伤(sadness)或惊讶(surprise). 在 DailyDilaog 中,那些表现出模棱两可情绪的语句被标注为中性,因此在这个数据中超过 83% 的语句被标记为中性. 对于行为标签,分为通知(inform)、疑问(questions)、建议(directives)、接受/拒绝(commissive)4 个类别.

本文仅利用上述数据集的文本形式进行实验,并且采用原始数据集的训练集/验证集/测试集的划分比例,数据集的统计如表 2 所示. 对于评估指标,遵循 Cerisara 等人[1] 和 Qin 等人[5,7],对 Dailydialog 数据集采用宏平均(Macro-average)准确率(Precision,P)和召回率(Recall,R)和 $F1$ 值,在 Mastodon 数据集上,忽略了 DSC 任务中的中性标签,而在 DAR 任务中,采用了行为特定均值 $F1$ 分数,由每个对话行为的流行度加权.

表 2 数据集统计

数据集		Mastodon	DailyDialog
对话数量	训练集	243	11118
	验证集	26	1000
	测试集	266	1000
语句数量	训练集	967	87170
	验证集	108	8069
	测试集	1142	7740
平均对话轮数		4	10
平均句子长度		10	13
情绪类别数		3	7
行为类别数		15	4

4.3 基 线

本文将提出的模型与一些最先进的基线进行比较,并将所有对比的基线分为三类,包括:

(1)单独的对话情感分类方法:DialogueRNN[12], DialogueGCN[14],DAG-ERC[16];

(2)单独的对话行为识别方法:HEC[31],CRF-ASN[25],CASA[24];

(3)对话情感分类和行为识别联合模型:Joint-DAS[3], IIIM[4],DCR-Net[5],Co-GAT[7].

值得注意的是,近年来,对话领域也涌现了一些相关的文献,如 Wei 等人提出的 E2E DAC 模型[32]、Sunder 等人提出的 HIER 模型[33],由于这些文献使用的数据集及应用场景与本文不一致,因此我们不与其进行比较.

4.4 结果对比和分析

4.4.1 总体性能实验

本文在两个公开数据集上进行了实验,所有被比较的基线的总体性能都列在表 3 中. 表现最优的和次优的结果分别用粗体和下划线表示. 另外,为了验证统计显著性,本文对提出的模型 SA-CCGN 与之前的最优模型 Co-GAT 使用配对 t 检验进行显著性检验. 当 p 值低于 0.05 或 0.01 时,差异被认为具有统计学显著性. 其中"$^{\#}$"和"$*$"标记分别表示本

文模型在配对 t 检验(p 值<0.01 和 p 值<0.05 时)结果显著优于基线模型 Co-GAT.

如表 3 所示,本文提出的模型 SA-CCGN 与所有基线方法相比,在两个数据集上都表现出更好的性能. 与最优的基线 Co-GAT 相比,SA-CCGN 在 Mastodon 和 DailyDialog 上 DSC 任务 $F1$ 值分别提高了 4.57%和 3.33%,DAR 任务 $F1$ 值分别提高了 2.15%和 0.63%. 通过配对 t 检验的显著性验证,可以发现,SA-CCGN 在 Mastodon 数据集所有指标上显著优于 Co-GAT,在 DailyDialog 数据集上绝大多数指标显著优于 Co-GAT.

另外,可以发现多任务模型表现有时略差于单任务模型,这可能是因为任务间不恰当或不充分的交互反而会降低模型的性能. 而本文的模型利用协同交互层,使得两个任务的信息之间进行合理交互,达到了最佳的模型性能.

表 3　模型在两个数据集上的总体表现

模型	Mastodon						DailyDialog					
	DSC			DAR			DSC			DAR		
	$F1$/%	R/%	P/%	$F1$/%	R/%	P/%	$F1$/%	R/%	P/%	$F1$/%	R/%	P/%
DialogueRNN	41.5	42.8	40.5	—	—	—	40.3	37.7	44.5	—	—	—
DialogueGCN	42.4	43.4	41.4	—	—	—	43.1	44.5	41.8	—	—	—
DAG-ERC	47.2	51.0	43.8	—	—	—	49.4	45.8	63.2	—	—	—
HEC	—	—	—	56.1	55.7	56.5	—	—	—	77.8	76.5	77.8
CRF-ASN	—	—	—	55.1	53.9	56.5	—	—	—	76.0	75.6	78.2
CASA	—	—	—	56.4	57.1	55.7	—	—	—	78.0	76.5	77.9
JiontDAS	37.6	41.6	36.1	53.2	51.9	55.6	31.2	28.8	35.4	75.1	74.5	76.2
IIIM	39.4	40.1	38.7	54.3	52.2	56.3	33.0	28.5	38.9	75.7	74.9	76.5
DCR-Net	45.1	47.3	43.2	58.6	56.9	60.3	45.4	40.1	56.0	79.1	79.0	79.1
Co-GAT	48.1	53.2	44.0	60.5	60.6	60.4	51.0	45.3	65.9	79.4	78.1	81.0
SA-CCGN	**50.3**$^{\#}$	**54.6**$^{\#}$	**46.6**$^{\#}$	**61.8**$^{\#}$	**62.0**$^{\#}$	**61.6**$^{\#}$	**52.7**$^{\#}$	**46.1***	**66.0**	**79.9***	**79.1**$^{\#}$	**81.6***

4.4.2　消融实验

为了验证模型 SA-CCGN 的每个组成部分的有效性,从 SA-CCGN 中移除每个部分进行比较,所有变体列举如下:

- w/o 说话者感知交互层:去掉说话者感知交互层,模型不再单独捕捉说话者随时间变化的信息.
- w/o 协同交互层:去掉协同交互层,两个任务进行单独的图传播,不再进行交互.
- w/o SCL 损失:只使用交叉熵及 CRF 损失.
- w/o CRF 损失:只使用交叉熵及 SCL 损失.
- w/o SCL&CRF 损失:只使用交叉熵损失,去除 SCL 及 CRF 损失.

表 4 为报告的消融实验的结果,最好的结果使用粗体表示,根据实验结果得到的结论为:

表 4　消融实验

模型	Mastodon		DailyDialog	
	DSC	DAR	DSC	DAR
	$F1$/%	$F1$/%	$F1$/%	$F1$/%
SA-CCGN	**50.26**	**61.76**	**52.69**	**79.93**
w/o 说话者感知交互层	49.85	61.02	52.10	79.33
w/o 协同交互层	49.77	60.55	52.17	79.51
w/o SCL 损失	49.42	60.61	51.66	79.35
w/o CRF 损失	49.86	61.25	52.68	79.73
w/o SCL& CRF 损失	49.30	60.59	51.42	79.28

(1)与完整的模型相比,移除任一组件的变体的性能都明显下降,这表明每个组件都起到了积极的作用.

(2)去除协同交互层将导致相当大的性能下降,例如在 Mastodon 两个任务性能下降分别为 0.97%,1.96%,1.25%,4.08%. 证明了本文的模型对于两个任务的交互是合理且充分的.

(3)分别去除 SCL 损失和 CRF 损失,模型表现的下降比例各有不同,这可能是因为不同数据集不同任务之间存在差异性.

(4)同时去除 SCL 损失和 CRF 损失,将导致与单独去除这两个组件的变体相比,性能更差. 它进一步证明了多损失解码器在探索标签信息的重要性.

4.4.3　参数敏感性实验

DAG 传播层的数量 L 控制了节点信息传播的范围. 在这个实验中,把层数 L 从 1 到 7 依次改变,研究参数 L 的敏感性,结果显示在图 3 中.

可以观察到,当增加层数时,性能持续上升,进一步增加 L 时,性能相对下降. 主要原因是当 L 较小时,DAG 的结构信息没有被很好地挖掘,如当 $L=1$ 时,模型聚合来自一跳邻居的信息. 然而过大的 L 可能会导致模型从图中的所有节点收集信息,使得图节点过平滑.

此外，最佳层数随不同的数据集和不同任务而变化. 在 DailyDialog 上，更多的层数是首选，即 $L=3$ 或 $L=4$. 而在 Mastodon 上，较少的层数性能更优，如 $L=1$. 原因可能是 Mastodon 的平均语句数量更少，L 的层数过多反而会损害模型性能.

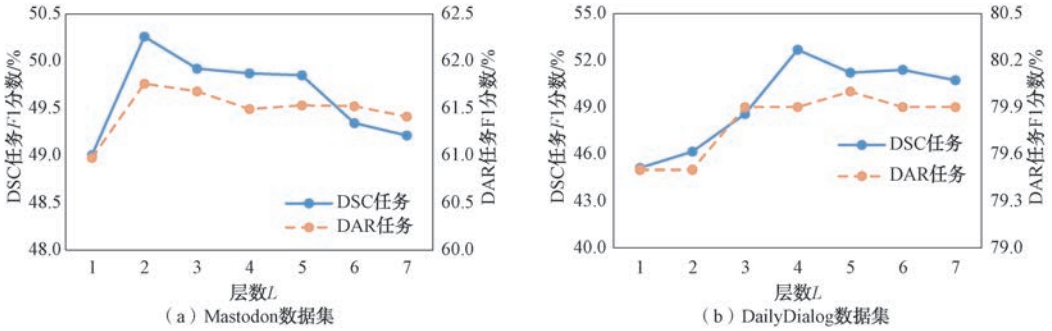

图 3 DAG 传播层取不同层数的模型性能

4.4.4 学习曲线实验

本节进一步在两个数据集上进行实验，以研究所提出的模型 SA-CCGN 与 Co-GAT 模型的收敛速度. 图 4 显示了两个模型的学习曲线，报告了在每个迭代次数(epoch)时两个任务的测试集的 F1 得分和训练集的总 loss 值. 从图 4 可以看到，SA-CCGN 模型收敛得较快，特别是在 DailyDialog 数据集上的 DSC 和 DAR 任务，SA-CCGN 分别在 5 和 12 个 epoch 左右就获得了最佳性能. 而在 Mastodon 数据集，由于此数据集数据量较小，收敛速度略慢于 DailyDialog 数据集，但是也能够在大约 50 个 epoch 时模型得到收敛. 与 SA-CCGN 相比，Co-GAT 的收敛速度略慢于 SA-CCGN，且模型的学习曲线较为抖动，尤其是在 Mastodon 数据集上更为明显. 总的来说，SA-CCGN 模型的学习曲线较为平滑与稳定，模型具有很好的收敛性和鲁棒性.

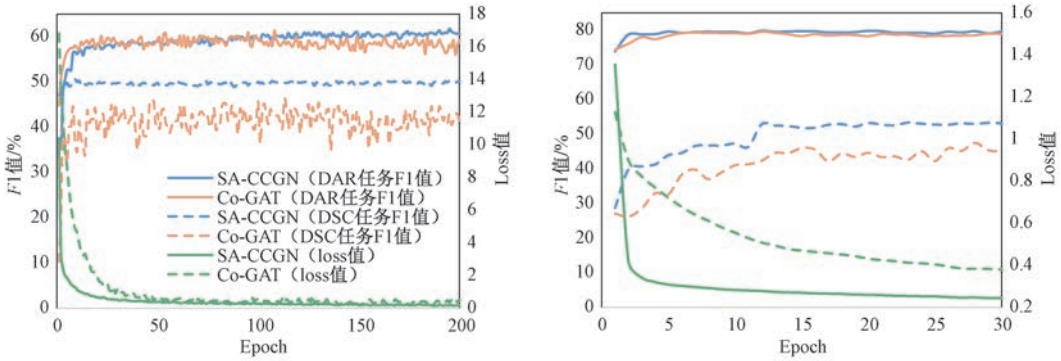

图 4 SA-CCGN 与 Co-GAT 学习曲线

4.4.5 低资源环境下的模型性能

本节将研究低资源环境下的模型性能. 由于 Mastodon 数据集本身数据量较小，不适合进行在低资源环境下的性能实验. 因此实验选取 DailyDialog 数据集，通过从原始训练集中随机选择 20%，40%，60%，80%，100%依次递增的样本数量来训练模型，并在原始测试集中进行测试. 图 5 显示了本文提出的模型 SA-CCGN 和最具竞争力的基线 Co-GAT 在不同比例的训练数据上的性能. 可以观察到，本文提出的 SA-CCGN 模型始终优于基线 Co-GAT. 训练数据的比例小时，相对于 Co-GAT 有显著提升，例如，在训练数据比例为 20%，40%时 DSC 任务的性能改进分别为 21%，25%. 实验结果验证了本文模型在低资源环境仍具有良好的性能.

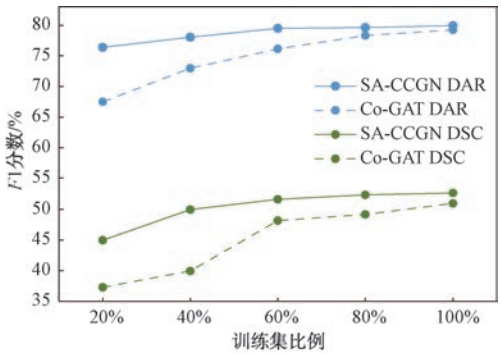

图 5 SA-CCGN 和 Co-GAT 在低资源环境的模型性能

4.4.6 t-SNE 可视化实验

为了提供更直观的评估，我们选取标签数相对

较少的 DailyDialog 数据集,分别对 SOTA 模型 Co-GAT 和本文提出的模型 SA-CCGN 在测试集上最后得到的语句表示使用 t-SNE 进行降维,结果如图 6 所示,其中不同的标签使用不同的图例表示. 由于 DailyDialog 数据集中 DSC 任务中性(none)标签占比太多(超过 83%),为了便于观察,我们删去了其中标签为中性的节点,剩余 6 类情感标签:愤怒(anger)、厌恶(disgust)、恐惧(fear)、快乐(happiness)、悲伤(sadness)或惊讶(surprise). DAR 任务使用数据集中原始的行为标签,一共 4 类:通知(inform)、疑问(questions)、建议(directives)、接受/拒绝(commissive). 通过语句嵌入可视化,我们可以看到 Co-GAT 学习到的语句表示把不同标签的节点都混到一起,不能很好地区分出各节点的类别,呈现比较模糊的边界,而 SA-CCGN 有着更清晰的边界,且相同簇的簇内距离也比 Co-GAT 模型更小. 本文提出的模型采用了监督对比损失,使得最后学习到的语句表示在相同标签之间的表示具有更高的相似性,而不同标签之间的表示相似性更低,即相同标签的样本具有内聚性,不同标签的样本相互排斥. 实验证明了我们模型的优越性,它能够较正确地分离不同标签的句子.

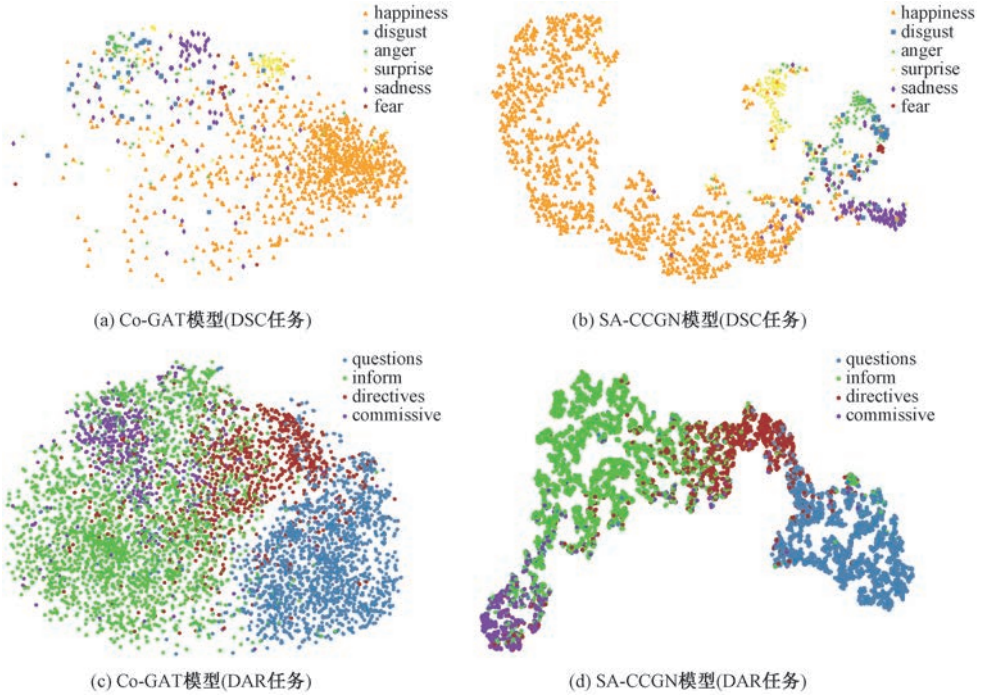

图 6　在 DailyDialog 数据集上学习到的语句表示可视化

4.4.7　计算效率对比实验

在实际应用中,除了性能外,参数数量、时间成本和所需的 GPU 内存也是重要因素. 以 Mastodon 数据集为试验基准,将本文提出的模型 SA-CCGN 与最新的 SOTA(Co-GAT)在这些因素上进行了比较,结果如表 5 所示. 值得注意的是,尽管本文模型在两个任务达到性能上都超过了 SOTA,但它将参数数量和所需的 GPU 内存减少约 1/2,因为 Co-GAT 两个任务之间通过一个全连通的无向图进行传播,而 SA-CCGN 则在两个任务分别进行传播后再可适应地协同交互,训练成本较低. 因此,在实际应用中本文提出模型是更有效的.

表 5　SA-CCGN 和 Co-GAT 的计算效率对比

模型	参数数量(↓)	每个 epoch 训练时间(↓)	GPU 内存(↓)	DSC 任务 F1 分数(↑)	DAR 任务 F1 分数(↑)
Co-GAT	6.93M	2.35s	2007MB	48.1%	60.5%
SA-CCGN	3.34M	1.78s	1007MB	50.3%	61.8%
提升	51.8%	24.26%	49.83%	4.57%	2.51%

5　总结与展望

本文提出了一个说话者感知跨任务协同交互图网络(SA-CCGN),其中的说话者感知交互层捕获了相同说话者随时间变化的情感和行为线索,而跨任务协同交互图网络能够很好联合建模两个任务之间的交互,DAG 传播层对对话信息进行传播,而协同

交互层则使两个任务能够相互迭代更新.最后的多损失解码器利用标签之间的区分度和关联性来使模型学习到更为细粒度的表示.实验验证本文所提出模型的有效性,并且超越了现有的基线,达到了先进的性能.当前,区分相似的标签会干扰我们的模型准确性,仍然是任务的重点和难点.下一步我们拟从如何更好地区分一些相似度较高的标签着手,并将探索更优的多任务学习方法,使得任务之间的交互更加合理和充分,提高模型的分类效果.

参 考 文 献

[1] Ghosal D, Majumder N, Mihalcea R, et al. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: an empirical study//Proceedings of the Association for Computational Linguistics. Bangkok, Thailand, 2021: 1435-1449

[2] ZHAO Yang-Yang, WANG Zhen-Yu, Wang Pei, et al. A survey on task-oriented dialogue systems. Chinese Journal of Computers, 2020, 43(10): 1862-1896 (in Chinese)
(赵阳洋, 王振宇, 王佩等. 任务型对话系统研究综述. 计算机学报, 2020, 43(10): 1862-1896)

[3] Cerisara C, Jafaritazehjani S, Oluokun A, et al. Multi-task dialog act and sentiment recognition on Mastodon//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 745-754

[4] Kim M, Kim H. Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. Journal of Pattern recognition letters, 2018, 101: 1-5

[5] Qin L, Che W, Li Y, et al. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(05): 8665-8672

[6] Li J, Fei H, Ji D. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 616-626

[7] Qin L, Li Z, Che W, et al. Co-GAT: a co-interactive graph attention network for joint dialog act recognition and sentiment classification//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021, 35(15): 13709-13717

[8] Veli? kovïc P, Cucurull G, Casanova A, et al. Graph attention networks//Proceedings of the 6th International Conference on Learning Representations Vancouver, Canada, 2018

[9] ZENG YI-FU, LAN Tian, WU Zu-Feng, et al. Bi-memory based attention model for aspect level sentiment classification. Chinese Journal of Computers, 2019, 8: 1845-1857 (in Chinese)
(曾义夫, 蓝天, 吴祖峰等. 基于双记忆注意力的方面级别情感分类模型. 计算机学报, 2019, 8: 1845-1857)

[10] GUO Xian-Wei, LAI Hua, YU Zheng-Tao et al. Emotion classification of case-related microblog comments integrating emotional knowledge. Chinese Journal of Computers, 2021, 44(3): 564-578 (in Chinese)
(郭贤伟,赖华,余正涛等. 融合情绪知识的案件微博评论情绪分类. 计算机学报, 2021,44(3):564-578)

[11] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 873-883

[12] Majumder N, Poria S, Hazarika D, et al. Dialoguernn: An attentive RNN for emotion detection in conversations//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33(01): 6818-6825

[13] Ghosal D, Majumder N, Gelbukh A, et al. COSMIC: COmmonSense knowledge for eMotion identification in conversations//Proceedings of the Association for Computational Linguistics. Online, 2020: 2470-2481

[14] Ghosal D, Majumder N, Poria S, et al. Dialoguegcn: a graph convolutional neural network for emotion recognition in conversation//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 154-164

[15] Zhong P, Wang D, Miao C. Knowledge-enriched transformer for emotion detection in textual conversations//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 165-176

[16] Shen W, Wu S, Yang Y, et al. Directed acyclic graph network for conversational emotion recognition//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online, 2021: 1551-1560

[17] Reithinger N, Klesen M. Dialogue act classification using language models//Proceedings of the Fifth European Conference on Speech Communication and Technology. Rhodes, Greece, 1997: 2235-2238

[18] Grau S. Dialogue act classification using a Bayesian approach//Proceedings of the 9th Conference Speech and Computer. Saint-Petersburg, Russia, 2004: 495-499

[19] Ortega D, Li C Y, Vallejo G, et al. Context-aware neural-based dialog act classification on automatically generated transcriptions//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, 2019: 7265-7269

[20] Lecun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. Neural computation, 1989, 1(4): 541-551

[21] Lafferty J D, Mccallum A, Pereira F C N. Conditional ran-

dom fields: probabilistic models for segmenting and labeling sequence data//Proceedings of the Eighteenth International Conference on Machine Learning. Williamstown, USA, 2021: 282-289

[22] Lee J Y, Dernoncourt F. Sequential short-text classification with recurrent and convolutional neural networks//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego California, USA, 2016: 515-520

[23] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation. Parallel Distributed Processing, 1986, 1: 318-362

[24] Raheja V, Tetreault J. Dialogue act classification with context-aware self-attention//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA, 2019: 3727-3733

[25] Chen Z, Yang R, Zhao Z, et al. Dialogue act recognition via crf-attentive structured network//Proceedings of the 41st International Acm Sigir Conference on Research & Development in Information Retrieval. Ann Arbor, USA, 2018: 225-234

[26] ZHANG Zhi-Chang, ZHOU Zhen-Wen, ZHANG Zhi-Man. User intent classification based on IndRNN-attention. Journal of Computer Research and Development, 2019, 56(7): 1517-1524 (in Chinese)
(张志昌, 张珍文, 张治满. 基于 IndRNN-Attention 的用户意图分类. 计算机研究与发展, 2019, 56(7): 1517-1524)

[27] ZHOU Jun-Zuo, ZHOU Zong-Kui, HE Zheng-Qiu, et al. Hybrid neural network models for human-machine dialogue intention classification. Journal of software, 2019, 30 (11): 3313-3325 (in Chinese)
(周俊佐, 朱宗奎, 何正球等. 面向人机对话意图分类的混合神经网络模型. 软件学报, 2019, 30(11): 3313-3325)

[28] Zhao F, Wu Z, Dai X. Attention transfer network for aspect-level sentiment classification//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 811-821

[29] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673

[30] Li Y, Su H, Shen X, et al. DailyDialog: A manually labelled multi-turn dialogue dataset//Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei, China, 2017: 986-995

[31] Kumar H, Agarwal A, Dasgupta R, et al. Dialogue act sequence labeling using hierarchical encoder with crf//Proceedings of the Aaai Conference on Artificial Intelligence. New Orleans, USA, 2018: 3440-3447

[32] Wei K, Knox D, Radfar M, et al. A neural prosody encoder for end-to-end dialogue act classification//Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, 2022: 7047-7051

[33] Sunder V, Thomas S, Kuo H K J, et al. Towards end-to-end integration of dialog history for improved spoken language understanding//Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, 2022: 7497-7501

LIU Si-Jin, M. S. candidate. Her research interests include natural language processing and sentiment analysis.

ZHU Xiao-Fei, Ph. D., professor. His research interests include natural language processing, data mining and information retrieval.

PENG Zhan-Wang, M. S candidate. His research interests include natural language processing and data mining.

Background

As an important research task in the field of artificial intelligence, the dialogue system has broad application prospects and has received extensive attention from academia and industry. In recent years, since the breakthrough of deep learning in the field of natural language, the research on dialogue systems has made a rapid development. At present, most domestic and foreign research treat dialogue sentiment classification and dialogue act recognition as two independent tasks. Dialogue sentiment classification, a subtask of text sentiment, has made promising progress and results, and the dialogue act classification has shifted from exploring lexical and syntactic features of an independent text to capturing more complex long-term dialogue-level context dependencies. However, the current research related to dialogue texts is at a preliminary stage, and most scholars focus on the classification tasks of independent texts, therefore there is still great space for exploring the analysis of dialogue texts with interactive information. In addition, most deep learning-based methods lack effective utilization of speaker information and label information, and implicit modeling or coarse-grained modeling the interaction of two tasks when they explore the correlation of them.

To address the above issues, this work proposed a novel multi-task learning framework, the Speaker-aware Cross-task Co-interactive Graph Network (SA-CCGN), in which the speaker-aware interaction layer captures the sentiment and act cues of the same speakers along with time, and the cross-task co-interaction graph network can well jointly model the interaction between two tasks, and finally uses the differentiation and correlation between labels to enable the model to learn a better representation. Compared with existing methods, the approach in this paper can better capture speaker cues over time and use label information to better constrain the training of the model. In addition, we construct directed acyclic graphs to model rich conversational information, and use a co-interaction layer to reasonably interact with the two tasks. The experimental results show that the proposed model effectively improves the performance of both dialogue sentiment classification and dialogue act recognition tasks, providing new ideas and contributing to the exploration in this task.

This work was supported by the National Natural Science Foundation of China (62141201), National Natural Science Foundation of Chongqing(CSTB2022NSCQ-MSX1672), Major Project of Science and Technology Research Program of Chongqing Education Commission of China (KJZD-M202201102). Dialogue sentiment classification and act recognition can help dialogue understanding and dialogue empathy, which are crucial for some downstream tasks such as public opinion monitoring, psychological analysis and response generation. In this paper, we study the sentiment classification and act recognition methods of dialogues according to the characteristics of dialogues texts.