计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(9):213−224 [doi: 10.15888/j.cnki.csa.009947] [CSTR: 32024.14.csa.009947]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于模态去异质性和自适应融合的多模态会话情感识别①

程佳玮[1], 朱小飞[1], 李曜辰[1], 曹均皓[1], 陈　旭[2]

[1](重庆理工大学 计算机科学与工程学院, 重庆 400054)
[2](重庆理工大学 会计学院, 重庆 400054)
通信作者: 朱小飞, E-mail: zxf@cqut.edu.cn

摘　要: 多模态会话情感识别任务旨在通过分析会话中产生的多种数据类型, 例如文本、音频和视觉, 从而理解会话中话语所要表达的情感. 因此许多基于多模态信息融合的方法被提出并且取得了不错的性能. 然而这些工作往往忽略了在不同情况下, 模态所展现的重要性是不同的. 此外这些工作也并没有考虑到多模态数据的异质性, 该问题会导致模态特征之间差距过大从而无法有效地进行模态信息的融合. 因此本文提出一种基于模态去异质性和自适应融合的会话情感识别模型, 以解决上述所提到的问题.首先利用共享的编码器将不同模态的特征映射到共享的语义空间中初步减少模态特征的差距, 然后使用共享卷积网络来最大化模态之间的共有语义信息从而消除模态特征之间的差距, 同时使用私有卷积网络来保持模态特征的多样性. 之后通过自注意力机制来学习出每个模态自身的重要性从而实现模态信息的自适应融合. 最后在两个公开的数据集上的实验结果表明, 本文所提出的模型优于以往的基线模型.

关键词: 情感识别; 特征融合; 注意力机制; 异质性

引用格式: 程佳玮,朱小飞,李曜辰,曹均皓,陈旭.基于模态去异质性和自适应融合的多模态会话情感识别.计算机系统应用,2025,34(9):213–224. http://www.c-s-a.org.cn/1003-3254/9947.html

Multimodal Emotion Recognition in Conversation Based on Modality De-heterogenization and Adaptive Fusion

CHENG Jia-Wei[1], ZHU Xiao-Fei[1], LI Yao-Chen[1], CAO Jun-Hao[1], CHEN Xu[2]

[1](College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)
[2](Accounting School, Chongqing University of Technology, Chongqing 400054, China)

Abstract: Multimodal emotion recognition in conversation aims to understand the emotions behind utterances in conversation by analyzing various types of data generated during the conversation, such as text, audio, and visual data. Therefore, numerous multimodal information fusion-based methods have been proposed and achieved notable performance. However, these methods often neglect that modality importance varies across different contexts, and they overlook the heterogeneity of multimodal data, which can lead to a significant gap between modal features, thereby hindering effective multimodal fusion. Therefore, this study proposes a modality de-heterogenization and adaptive fusion model for emotion recognition in conversation to address the aforementioned issues. First, a shared encoder is used to map features from different modalities into a shared semantic space to preliminarily reduce the gap between modal features. Then, shared convolutional networks are employed to maximize mutual semantic information across modalities,

① 基金项目: 国家自然科学基金 (62141201); 重庆市自然科学基金面上项目 (CSTB2022NSCQ-MSX1672); 重庆英才计划 (CSTC2024YCJH-BGZXM0022); 重庆市教育委员会科学技术研究计划重大项目 (KJZD-M202201102)
收稿时间: 2025-01-21; 修改时间: 2025-02-24; 采用时间: 2025-03-14; csa 在线出版时间: 2025-07-23
CNKI 网络首发时间: 2025-07-23

eliminating the gap between modal features, and private convolutional networks are used to maintain the diversity of modal features. Subsequently, the self-attention mechanism is employed to learn the importance of each modality, thereby achieving adaptive fusion of modal information. Finally, experimental results on two public datasets demonstrate that the proposed model outperforms existing baseline models.

Key words: emotion recognition; feature fusion; attention mechanism; heterogeneity

1 引言

情感是人类对话中的重要组成部分[1]. 会话情感识别任务 (emotion recognition in conversation, ERC) 是分析会话中的每个话语并给出相应的情感. 这项任务最近受到了学术界和工业界越来越多的关注, 因为它具有重大的潜在应用价值, 如人机交互[2]和社交媒体中的观点挖掘[3].

传统的会话情感识别范式要么基于会话中独立的句子, 要么基于单个模态的信息, 例如文本. 然而在大多数情况下, 人们的情感是复杂的并且展现的方式也是多样化的, 仅依靠单一模态的信息是无法准确地捕捉真实的情感状态[4]. 在现实生活中人们实际上是结合说话者的表情、语调以及话语内容来识别情感. 因此在会话情感识别的基础上, 研究人员引入了多模态会话情感识别 (multimodal emotion recognition in conversation, MERC) 并提出了许多的多模态信息融合的工作[5,6], 本文把这些工作主要分为两类: 基于图和基于 Transformer. 例如 SCMFN[7]构建多模态说话者感知图用于捕获不同说话者和不同模态之间的联合关系. MGLRA[8]通过构建多模态异质图来结合模态之间的信息. CTNet[9]引入跨模态 Transformer 来显式地挖掘模态之间的关联从而增强模态的表示. CMCF-SRNet[10]则是使用位置信息和说话者信息来引导模态间信息的融合从而增强模态的表示.

尽管这些先驱研究工作在情感识别方面取得了可喜的效果, 但他们往往忽略了以下两个问题: (1) 不同模态的重要性程度各不相同. 图 1 是一个多模态会话例子, 以第 1 个句子为例, 该句子的文本和视觉所传达的情感信息是模糊的, 但是响亮的声音传达了强烈的情绪, 因此音频模态作为了情感的主导模态. 同理第 4 个句子中音频和视觉传达的情感并不清晰而文本信息中的“unsettling”一词展现出了强烈的情感, 因此文本模态作为了情感的主导模态. 以往大多数工作融合多模态信息都是基于一个隐式的假设, 即不同模态之间的重要性是相同的. 实际上, 从图 1 可以发现不同模态之间的重要性其实是有差异的. 在多模态信息融合的过程中忽视了不同模态的相对重要性会导致次优的情感识别的性能. (2) 模态之间的异质性阻碍了多模态信息的融合. 由于每个模态的原始数据结构和数据分布各不相同, 这就会导致提取出的模态特征之间存在着巨大的差距. 如果不消除这个差距, 那么将会无法有效地融合模态信息, 从而实现次优的情感识别效果.

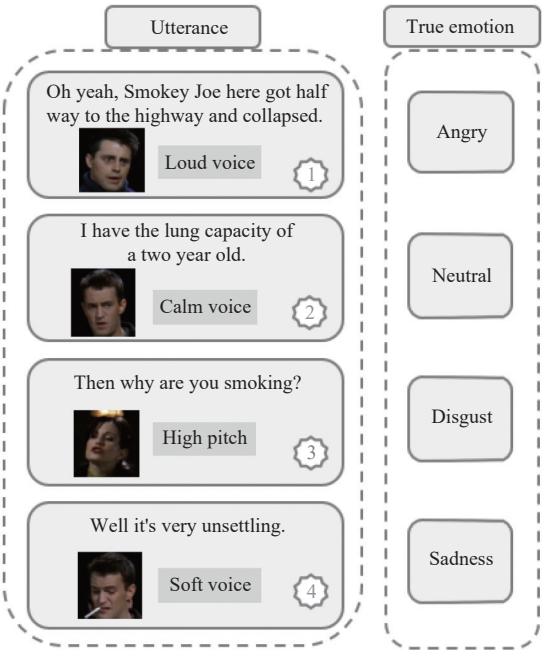

图 1 多模态会话例子

因此受到上述分析的启发, 本文提出了一个更加全面的多模态融合模型, 命名为模态去异质性和自适应融合的会话情感识别模型, 用于提高情感识别的性能. 具体而言, 为了解决模态异质性带来的问题, 本文使用共享的编码器首先将不同模态的特征映射到同一个共享语义空间中, 然后共享的卷积网络被用于学习模态之间共有的语义信息, 从而最大化多模态信息的共性以达到去除模态之间的特征差距. 考虑到在不同

模态之间追求共享信息可能会导致模态特征变得过于相似, 本文还使用私有编码器将每种模态映射到私有的语义空间中, 并且采用私有卷积网络来学习每种模态的私有上下文信息. 这种私有上下文信息被注入共享上下文信息中, 以提高共享语义空间中模态特征的多样性. 在模态信息融合阶段, 考虑到不同模态之间的重要性, 本文设计了 MATransformer (modality-aware Transformer), 它能够通过注意力机制显式地计算每个模态的重要性权重从而实现自适应融合模态的信息. 最后在两个公开的数据集的实验结果表明, 本文的模型在总体准确率和 W-*F*1 分数方面都超越了现有的基线模型. 总的来说, 这个工作的主要贡献如下.

(1) 将多模态特征通过投影到共享空间中并最大化它们之间的共有语义信息来消除模态特征之间的差距, 同时也使用私有空间来保留住模态自身的私有语义信息从而提高模态特征的多样性.

(2) 提出了 MATransformer 框架, 用于建模模态自身的重要性从而合理地融合模态的信息, 最后达到增强单一模态表示的效果.

(3) 所提出的模型在两个公开数据集上优于以往的基线模型, 为情感识别建立了一个健壮和可解释性的多模态话语特征.

2 相关工作

会话情感识别作为自然语言处理的一个重要研究领域, 近年来受到了广泛的关注. 现有的关于 ERC 的研究主要有两种类型的数据输入: 基于文本和基于多模态. 接下来, 会按照这两种类型来介绍相关工作.

2.1 基于文本模态的 ERC

文本模态中包含着丰富的语义信息, 早期大多数工作都是围绕如何更好地提取文本模态中丰富的语义上下文信息和利用说话者信息而展开. Majumder 等[11]使用 3 个不同的门控循环单元 (gate recurrent unit, GRU)[12]来分别更新说话者信息、上下文信息和情感状态, 并且所有的 GRU 都是以一种循环的方式连接在一起. Ghosal 等[13]使用图神经网络建模说话者自身和说话者之间的依赖关系以及对话序列信息. Ishiwatari 等[14]提出了拥有相对位置编码的 R-GAT 模型, 该模型不但能够获取说话者之间的依赖性, 同时还能为关系图结构提供序列信息. Shen 等[15]设计了一个有向无环图网络结构用于编码现实对话中句子结构的信息, 从而以一种显式的方式来获取对话中说话者的信息. Zhu 等[16]则是使用编码-解码结构的模型将对话中的话题信息和外部常识进行结合. Jiao 等[17]使用两个不同的 GRUS 来分别建模单词和句子之间的关系. Ma 等[18]利用多视图网络对会话中单词和话语之间的依赖关系进行建模. 尽管上述方法在精度上有所提升, 但是都只关注于单一模态从而忽略了在对话中所展现出来的各种模态数据之间的相互补充和增强.

2.2 基于多模态的 ERC

考虑到单一模态信息的限制并且有研究证明多模态信息学习比单模态信息学习[19]有效之后, 许多研究人员聚焦于多模态信息融合. Mao 等[20]使用层级 Transformer 和多粒度交互融合模块来挖掘模态内部和模态之间的情感线索. Hu 等[21]构建全连接网络来充分建模多模态和远距离信息, 并且将说话者信息作为外部信息直接加入到网络当中. Hu 等[22]设计了一个基于图的动态融合模块来减少冗余信息和增强模态之间的互补信息. 宗林林等[23]通过设计超图来建模模态之间的相关性从而实现模态间的信息融合. Hu 等[24]在句法和语义级别上执行模态融合, 并引入模态间对比学习来区分样本之间的融合表示. Tsai 等[25]使用跨模态 Transformer 来建模模态之间的远距离依赖关系. Yuan 等[26]同样也是采用 Transformer 编码器来获取模态内部和模态之间的交互关系.

这些工作在进行多模态信息融合时并未考虑多模态数据的异质性和模态自身的相对重要性. 与这些工作不同的是, 本文提出多模态融合框架, 它不仅能够通过共享的学习方式解决异质性问题, 还能够学习模态自身的重要性从而合理地融合模态之间的信息.

3 预定义

3.1 任务定义

在会话情感识别任务中, 每个对话由一系列话语组成$U=\{u_1,u_2,\cdots,u_n\}$, 其中n代表话语的数量. 这些话语由k个说话者说出, 用$P(\cdot)$表示映射函数, 用于获取话语的说话者索引. 每个话语都可以表示为$u_i=\{\boldsymbol{u}_{t,i},\boldsymbol{u}_{a,i},\boldsymbol{u}_{v,i}\}$. 其中$\boldsymbol{u}_t\in\mathbb{R}^{d_t}$, $\boldsymbol{u}_a\in\mathbb{R}^{d_a}$和$\boldsymbol{u}_v\in\mathbb{R}^{d_v}$代表$u_i$的文本、音频和视觉特征. 该任务的目的是基于多模态数据识别出每个话语u_i的情感标签y_i.

3.2 模态特征抽取

对于文本特征抽取来说, 大多数工作都基于预训

练语言模型. 先前实验证明了使用不同的预训练语言模型会展现不同的性能. 早期工作使用 GloVe[27]进行特征提取, 还有一些工作使用 BERT[28]或者 RoBERTa[29]来进行特征提取. 本文和先前工作一样使用 RoBERTa 来进行文本模态的特征提取. 对于音频和视觉特征抽取来说, 本文使用 OpenSMILE[30]和 DenseNet[31]来抽取对应的模态特征. 同时为了捕获上下文信息和处理多模态特征中的维度不一致问题, 本文对文本模态使用双向门控循环单元 (bidirection gate recurrent unit, BiGRU) 并且对音频模态和视觉模态使用全连接层将每个模态$m \in \{t,a,v\}$的话语特征$\boldsymbol{u}_{m,i} \in \mathbb{R}^{d_m}$映射成固定大小的话语特征$\boldsymbol{e}_{m,i} \in \mathbb{R}^{d}$. 具体操作如下:

$$\boldsymbol{e}_{t,i}, \boldsymbol{o}_{t,i} = BiGRU_t\left(\boldsymbol{u}_{t,i}, \boldsymbol{o}_{t,i-1}\right) \tag{1}$$

$$\boldsymbol{e}_{\{a,v\},i} = Linear_{\{a,v\}}\left(\boldsymbol{u}_{\{a,v\},i}\right) \tag{2}$$

其中, $\boldsymbol{o}_{t,i-1}$和$\boldsymbol{o}_{t,i}$表示隐藏状态. 考虑到情感识别中说话者的重要性[32], 另一个 BiGRU 用于提取说话者特征.

$$\boldsymbol{s}_{m,i}, \boldsymbol{o}_{m,i} = BiGRU_p\left(\boldsymbol{u}_{m,i}, \boldsymbol{o}_{m,j}\right),\ 1 \leqslant j < i \tag{3}$$

$$\boldsymbol{h}_{m,i} = \boldsymbol{e}_{m,i} + \boldsymbol{s}_{m,i} \tag{4}$$

其中, $\boldsymbol{o}_{m,j}$是u_j的隐藏状态并且$P(u_j) = P(u_i)$, 即u_j和u_i由同一个说话者说出. $\boldsymbol{s}_{m,i}$代表说话者特征.

4 总体模型

在本节中会详细介绍本文所提出的模型. 模型架构如图 2 所示, 该模型主要由 4 部分组成: (1) 模态特征映射模块, 用于将每个模态的特征分别映射到其共享语义空间和私有语义空间中; (2) 模态私有和共有上下文语义信息学习模块, 用于消除模态特征差距同时保持模态特征多样性; (3) 模态感知融合模块, 用于学习模态自身的重要性并实现合理的多模态信息融合; (4) 情感分类模块, 用于输出模型的预测结果.

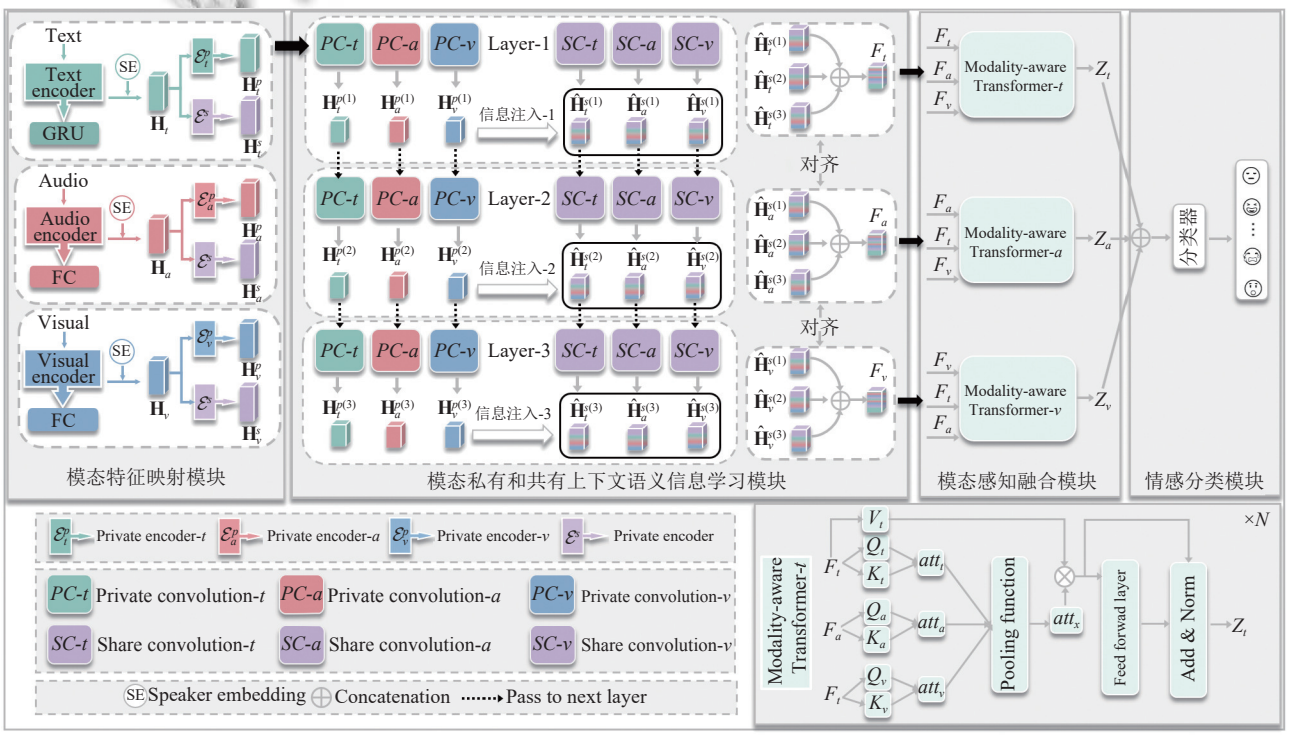

图 2 基于模态去异质性和自适应融合的多模态会话情感分析框架图

4.1 模态特征映射模块

本文首先使用共享编码器将每个模态的特征映射到相同的语义空间来初步减少模态特征之间的差距. 具体公式如下:

$$\mathbf{H}_m^s = \mathcal{E}^s(\mathbf{H}_m, \Theta),\ m \in \{t,a,v\} \tag{5}$$

其中, $\mathbf{H}_m^s, \mathbf{H}_m \in \mathbb{R}^{n \times d}$, $\mathcal{E}^s$代表共享编码器. 该编码器由全连接层和激活函数组成. 此外, 本文还会对每个模态使用私有编码器将它们映射到私有的语义空间:

$$\mathbf{H}_m^p = \mathcal{E}_m^p(\mathbf{H}_m, \Theta_m) \tag{6}$$

其中, $\mathbf{H}_m^p \in \mathbb{R}^{n \times d}$, $\mathcal{E}_m^p$是m模态的私有编码器. 该编码器

同样由全连接层和激活函数组成.

4.2 模态私有和共有上下文语义信息学习模块

为了能够学习出会话中的上下文语义信息, 本文使用卷积网络来学习话语间的交互作用. 由于会话是顺序发生的, 因此本文限制话语信息只能从前往后传递. 具体而言, 对于每个话语u_i, 本文会使用该话语的第$l-1$层的特征$\boldsymbol{h}_{m,i}^{l-1}$和它的前继话语的第$l$层的特征$\boldsymbol{h}_{m,j}^{l}$计算出注意力权重.

$$\alpha_{ij}^{l}=Softmax_{j\in N_i}(W_{\alpha}^{l}[\boldsymbol{h}_{m,j}^{l}\|\boldsymbol{h}_{m,i}^{l-1}]) \tag{7}$$

其中, $W_{\alpha}^{l}\in\mathbb{R}^{d\times d}$是可学习参数, $\|$表示拼接操作. 此外本文对信息进行特征转化用于表示该话语信息是属于相同说话者还是不同说话者. 具体公式定义如下:

$$\boldsymbol{I}_{m,i}^{l}=\sum_{j\in N_i}\alpha_{ij}^{l}W_{r_{i,j}}\boldsymbol{h}_{m,j}^{l} \tag{8}$$

其中, $W_{r_{i,j}}^{l}\in\mathbb{R}^{d\times d}$, $W_{r_{i,j}}^{l}\in\{W_0^l,W_1^l\}$是可学习的特征转化矩阵, 其中 0/1 代表边的类型用于区分不同说话者和相同说话者. 之后本文使用两个不同的 GRU 来对这两种信息进行多方面的交互从而获得话语u_i在第l层的隐藏表示$\boldsymbol{h}_{m,i}^{l}$. 具体公式定义如下:

$$\tilde{\boldsymbol{I}}_{m,i}^{l}=GRU_1(\boldsymbol{h}_{m,i}^{l-1},\boldsymbol{I}_{m,i}^{l}) \tag{9}$$

$$\tilde{\boldsymbol{I}}_{m,i}^{l}=GRU_2(\boldsymbol{I}_{m,i}^{l},\boldsymbol{h}_{m,i}^{l-1}) \tag{10}$$

$$\boldsymbol{h}_{m,i}^{l}=\tilde{\boldsymbol{I}}_{m,i}^{l}+\tilde{\boldsymbol{I}}_{m,i}^{l} \tag{11}$$

其中, GRU_1和GRU_2分别代表节点信息到历史信息的交互以及历史信息到节点信息的交互并且$\boldsymbol{h}_{m,i}^{l}\in\mathbf{H}_m^l$.

最后本文将式 (7)–式 (11) 统一定义成如下形式:

$$\mathbf{H}_m^l=Convolution(\mathbf{H}_m^{l-1},\Theta) \tag{12}$$

其中, Θ代表卷积网络的参数. 之后本文同时在共享语义空间和私有语义空间中进行卷积操作:

$$\mathbf{H}_m^{s,l}=SC-m(\mathbf{H}_m^{s,l-1},\Theta) \tag{13}$$

$$\mathbf{H}_m^{p,l}=PC-m\left(\mathbf{H}_m^{p,l-1},\Theta_m\right) \tag{14}$$

其中, SC 和 PC 分别代表共有卷积和私有卷积.

值得注意的是, 在共有卷积中使用的是每个模态的共有特征$\mathbf{H}_m^s$并且都共享参数Θ. 这样的方式可以最大化模态之间的共有信息从而进一步消除模态的特征差距. 而私有卷积中使用的是模态的私有特征$\mathbf{H}_m^p$并且都有属于自己的私有参数Θ_m. 这样的方式可以单独地学习每个模态的私有信息, 然后这个私有信息会注入共有信息中用于提高共有信息的多样性:

$$\hat{\mathbf{H}}_m^{s,l}=W[\mathbf{H}_m^{s,l}\|\mathbf{H}_t^{p,l}\|\mathbf{H}_a^{p,l}\|\mathbf{H}_v^{p,l}]+b \tag{15}$$

其中, $W\in\mathbb{R}^{4d\times d}$. 之后$\hat{\mathbf{H}}_m^{s,l}$会输入到 m 模态的第$l+1$层的共享卷积网络中进行再次卷积. 这样一来, 在共享卷积的过程中还能够保持多模态特征的多样性, 最后本文通过堆叠$\hat{\mathbf{H}}_m^{s}$来获得该模块的最终输出$\mathbf{F}_m$:

$$\mathbf{F}_m=\|_{l=0}^{L}\hat{\mathbf{H}}_m^{s,l} \tag{16}$$

其中, $\mathbf{F}_m\in\mathbb{R}^{n\times d'}, d'=L\cdot d$. 和最开始的$\mathbf{H}_m$相比, $\mathbf{F}_m$不仅有效缓解了模态之间的特征差距, 同时还保留着每个模态自身的多样性. 此外, 本文使用对齐操作来进一步对齐不同模态之间的语义信息, 从而拉近模态之间的共性. 对齐操作分别由有监督对比学习和无监督对比学习组成.

有监督对比学习: 该对比学习的目的是将模态内部相同情感标签话语的特征进行对齐, 具体操作如下:

$$\begin{cases}\mathcal{L}_{\text{scl}}^{m}=-\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}1_{\hat{y}_i=\hat{y}_j}\log\frac{\exp\left(\left(\boldsymbol{f}_{m,i}\boldsymbol{f}_{m,j}^{\mathrm{T}}\right)/\tau\right)}{\displaystyle\sum_{k=1}^{n}1_{\hat{y}_k\neq\hat{y}_i}\exp\left(\left(\boldsymbol{f}_{m,i}\boldsymbol{f}_{m,k}^{\mathrm{T}}\right)/\tau\right)}\\ \mathcal{L}_{\text{scl}}=\mathcal{L}_{\text{scl}}+\mathcal{L}_{\text{scl}}^{u}+\mathcal{L}_{\text{scl}}^{v}\end{cases} \tag{17}$$

其中, $\boldsymbol{f}_{m,i}\in\mathbf{F}_m$并且$\tau$是温度参数. $1_{\hat{y}_i=\hat{y}_j}$是一个指示函数, 如果$\hat{y}_i=\hat{y}_j$那么就取 1, 否则为 0

无监督对比学习: 该对比学习的目的是将不同模态之间相同话语的特征进行对齐, 具体操作如下:

$$\begin{cases}\mathcal{L}_{\text{ucl}}^{m1,m2}=-\displaystyle\sum_{i=1}^{n}\log\frac{\exp\left(\left(\boldsymbol{f}_{m1,i}\boldsymbol{f}_{m2,i}^{\mathrm{T}}\right)/\tau\right)}{\displaystyle\sum_{j=1}^{n}\exp\left(\left(\boldsymbol{f}_{m1,i}\boldsymbol{f}_{m2,j}^{\mathrm{T}}\right)/\tau\right)}\\ \mathcal{L}_{\text{ucl}}=\mathcal{L}_{\text{ucl}}^{a}+\mathcal{L}_{\text{ucl}}^{v}+\mathcal{L}_{\text{ucl}}^{av}\end{cases} \tag{18}$$

其中, $m1,m2\in\{t,a,v\}$并且$m1\neq m2$.

4.3 模态感知融合模块

多模态信息融合一直都是多模态学习的一个挑战, 分析和挖掘模态之间的关联性和互补性是提高多模态信息融合效果的重要步骤. 基于此, 许多工作提出了复杂的融合机制并展现了令人可喜的效果. 然而他们似乎没有考虑到模态之间的重要性是不一致的. 为了解决这个问题, 本文提出 MATransformer 来建模模态的重要性从而对多模态信息进行自适应的融合. 具体而言, 对于第n层 MATransformer 来说本文使用注意力操作来获取每个模态自身的重要性:

$$att_m^n = \frac{\left(W_m^{n,Q}\mathbf{F}_m^n\right)\left(W_m^{n,K}\mathbf{F}_m^n\right)^{\mathrm{T}}}{\sqrt{d'}},\ m \in \{t,a,v\} \tag{19}$$

求和操作被用来结合这些模态重要性以得到模态重要性感知的注意力权重矩阵:

$$att_x^n = sum(att_t^n, att_a^n, att_v^n) \tag{20}$$

这样一来, att_x^n能够指导模型对模态信息进行合理的融合从而增强模态的特征表示. 由于每个模态的增强操作都类似, 这里以文本模态为例子来说明. 增强操作定义如下:

$$\begin{cases} \tilde{\mathbf{F}}_t^n = \dfrac{att_x^n \cdot (W_t^{n,V}\mathbf{F}_t^n)}{\sqrt{d'}} \\ \hat{\mathbf{F}}_t^n = Norm(\tilde{\mathbf{F}}_t^n + \mathbf{F}_t^n) \\ \mathbf{F}_t^{n+1} = Norm(\sigma(FFN(\hat{\mathbf{F}}_t^n)) + \hat{\mathbf{F}}_t^n) \end{cases} \tag{21}$$

其中, $Norm$表示归一化, 加法操作是为了防止在训练过程中出现梯度消失问题. σ代表激活函数, FFN表示前馈神经网络. 本文把式 (19) 和式 (21) 定义为如下形式:

$$\mathbf{Z}_t = MAT - t^N(\mathbf{F}_t, \mathbf{F}_a, \mathbf{F}_v, \Theta) \tag{22}$$

其中, MAT表示 MATransformer. 同理可以获得增强的音频模态和视觉模态特征表示:

$$\begin{cases} \mathbf{Z}_a = MAT - a^N(\mathbf{F}_a, \mathbf{F}_t, \mathbf{F}_v, \Theta) \\ \mathbf{Z}_v = MAT - v^N(\mathbf{F}_v, \mathbf{F}_t, \mathbf{F}_a, \Theta) \end{cases} \tag{23}$$

值得注意的是, MATransformer 也是以共享参数的方式运用在每个模态上面.

4.4 情感分类模块

在获得了 3 个模态的增强表示后, 本文使用拼接操作将它们进行结合以得到最终的多模态表示, 然后送入到分类器中获取最终的预测结果:

$$\begin{cases} \mathbf{Z}_x = \mathbf{Z}_t \,\|\, \mathbf{Z}_a \,\|\, \mathbf{Z}_v \\ P = classifier(\mathbf{Z}_x) \\ y_i = \mathrm{argmax}(P_i) \end{cases} \tag{24}$$

其中, $classifier$ 由全连接层、激活函数和归一化层组合而成.

4.5 训练目标

本文使用分类交叉熵损失来定义在训练过程中预测的情感类别和真实标签之间的误差损失:

$$\mathcal{L}_{\mathrm{ce}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} y_{i,j} \cdot \log(\hat{y}_{i,j}) + \lambda\|\Theta\|_2 \tag{25}$$

通过将分类损失和对比学习损失结合, 本文得到最终的损失函数:

$$\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{ce}} + \alpha\mathcal{L}_{\mathrm{scl}} + \beta\mathcal{L}_{\mathrm{ucl}} \tag{26}$$

其中, 超参数权重α/β用于控制有监督/无监督对比学习的重要性.

5 实验

5.1 数据集和评价指标

本文使用 IEMOCAP[33]和 MELD[34]两个数据集来评价模型的有效性, 这些数据集都包含了文本、音频和视觉模态的信息. 表 1 为两个数据集的数据分布.

表 1 数据集统计

数据集	会话		话语数量		类别
	训练集	测试集	训练集	测试集	
IEMOCAP	120	31	5 810	1 623	6
MELD	1 153	280	11 098	2 610	7

IEMOCAP: 该数据集中的每个对话都来自两位演员根据剧本进行的表演. IEMOCAP 中包含了 7 433 个话语和 151 个对话. 对话中的每一句话都标有 6 种情绪: 快乐 (happy)、悲伤 (sad)、中立 (neutral)、愤怒 (angry)、兴奋 (excited) 和沮丧 (frustrated).

MELD: 该数据集中的数据来自电视节目《老友记》, 总共包含了 13 708 个话语和 1 433 个对话. 与 IEMOCAP 数据集不同的是, MELD 在对话中有 3 个或者多个说话者. 对话中的每个话语都标有 7 类情绪: 中性 (neutral)、惊讶 (surprise)、恐惧 (fear)、悲伤 (sadness)、快乐 (joy)、厌恶 (disgust) 和生气 (angry).

5.2 实验设置

本文模型使用 PyTorch 实现并使用 Adam[35]优化器来训练模型. 在 IEMOCAP 数据集上的学习率和批量大小为 0.000 1 和 32. α和β的值分别设定为 0.07 和 0.009. MATransformer 的层数设定为 3 层. 在 MELD 数据集上的学习率和批量大小为 0.000 1 和 36. α和β的值分别设定为 0.05 和 0.005. MATransformer 的层数设定为 2 层. 本文模型在 GPU 型号为 RTX 2080Ti 上训练, 且结果均为随机运行 3 次结果的平均值.

5.3 对比实验

在本节中展示了本文的模型和以往基线模型的性能比较. 以下是对基线模型的介绍.

DialogueRNN[11]: 使用 3 个不同功能的 GRU (全局 GRU、说话者 GRU 和情感 GRU) 将对话中的说话者和顺序信息进行建模.

DialogueGCN[13]: 将 GCN 应用于 ERC 任务中, 生成的特征可以整合丰富的信息, 但是这样同样会引入冗余的信息.

MMGCN[21]: 使用 GCN 网络在考虑模态内部的全局上下文信息基础上, 还尝试把其他模态信息建模进去.

MM-DFN[22]: 使用一种新的基于图的动态融合模块来充分融合多模态上下文信息, 更好地解释话语中的情感.

SCMM[36]: 设计了模态交互模块, 该模块由 3 个子模块组成, 包括完全、部分和有偏交互, 以充分研究不同模态之间的相关性.

CTNet[9]: 用 Transformer 的结构来模拟不同模态的模态内和跨模态交互. 单模态 Transformer 被开发来捕获单模态特征之间的时间依赖性, 而跨模态 Transformer 被设计来学习在未对齐的多模态特征上的跨模态交互.

CMCF-SRNet[10]: 设计了两种 Transformer, 即跨模态局部约束 Transformer 和基于图的语义细化 Transformer, 用以探索话语间的多模态交互和语义关系信息.

MGLRA[8]: 引入记忆机制来迭代地对齐多个模态的语义信息并且使用跨模态多头注意力机制来探索模态间的交互语义信息从而扩大语境信息的感受域.

SCMFN[7]: 设计了一个基于话语距离的注意力模块, 根据距离来区分信息融合的优先级. 该模块使得模型能够有效利用不同话语对情感识别的不同贡献.

表 2 和表 3 分别展现了在 IEMOCAP 数据集和 MELD 数据集上本文提出的模型和基线模型的性能的比较. 其中最优和次优的结果分别进行了加粗和下划线标注, “*”表示提升是具有显著性的 (t-test 下 p 值小于 0.05), “—”表示原文并未给出结果.

表 2 本文模型与基线模型在 IEMOCAP 数据集上的总体结果 (%)

基线模型	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	W-F1
DialogueRNN	32.20	80.48	57.89	62.82	73.87	59.76	63.52	62.89
DialogueGCN	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89
MMGCN	39.66	76.89	62.81	**71.43**	75.40	63.43	65.43	66.25
CTNet	51.30	79.90	65.80	67.20	**78.70**	58.80	68.00	67.50
MM-DFN	42.22	78.98	66.42	69.77	75.56	66.33	68.21	68.18
SCMM	45.37	78.76	63.54	66.05	76.70	66.18	—	67.53
CMCF-SRNet	52.20	81.40	68.80	70.30	76.70	66.18	70.50	69.60
MGLRA	**63.50**	81.80	71.50	61.10	76.30	67.80	71.30	70.10
SCMFN	—	—	—	—	—	—	71.23	71.21
本文模型	58.53	**85.74**	**74.79**	66.28	77.29	**71.34**	**73.82***	**73.76***

表 3 本文模型与基线模型在 MELD 数据集上的总体结果 (%)

基线模型	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Angry	Acc	W-F1
DialogueRNN	76.97	47.69	—	20.41	50.92	—	45.23	—	57.66
DialogueGCN	75.97	46.05	—	19.60	51.20	—	40.83	—	56.36
MMGCN	76.33	48.15	—	22.93	53.02	—	46.09	60.42	58.31
CTNet	77.40	52.70	10.00	32.50	56.00	11.20	44.60	62.00	60.50
MM-DFN	77.76	50.69	—	22.93	54.76	—	47.82	62.49	59.46
SCMM	—	—	—	—	—	—	—	—	59.44
CMCF-SRNet	—	—	—	—	—	—	—	62.80	62.30
MGLRA	**80.80**	**59.50**	00.00	27.80	**66.50**	00.00	**48.40**	66.40	64.90
SCMFN	—	—	—	—	—	—	—	67.01	66.25
本文模型	79.52	58.93	**25.13**	**44.48**	63.85	**23.39**	**57.51**	**67.58***	**66.67***

从表 2 和表 3 中可以得出本文提出的模型在所有的数据集的总体指标上达到了最优的水平. 具体而言, 与最优基线 SCMFN 相比, 本文模型在 IEMOCAP 数据集的总体准确率 (Acc) 和 W-F1 分别提升了 2.59% 和 2.55%. 同时在“Sad”“Neutral”和“Frustrated”这 3 个情感类比上面达到了最优水平, 在“Happy”和“Excited”情感上面达到了次优的效果. 在 MELD 数据集中, 与 SCMFN 相比, 本文提出的模型在 Acc 和 W-F1 上分别提升了 0.57% 和 0.42%, 而在“Fear”“Sadness”“Disgust”和“Angry”这 4 个情感类别上面表现出了最优的性能, 在剩余的情感上也都表现出了不弱的效果. 从上述的实验结果可以发现本文所提出的模型总体上展现出了具有竞争力的性能表现, 这主要归功于模型能够较好地解决模态异质性的问题, 且在融合多模态信息方面

能考虑到模态自身的重要性. 相反 SCMFN 和 MGLRA 在对多模态信息融合的时候没有考虑到多模态数据之间的异质性问题, 从而在 Acc 和 W-*F*1 方面实现了次优的效果.

5.4 消融实验

在本节中, 本文进一步深入分析模型中各个模块的有效性以及每个模态的重要性. 下文介绍了本文的各种消融设置. 消融实验结果如表 4 所示.

表 4 消融实验结果展示 (%)

消融操作	IEMOCAP		MELD	
	Acc	W-*F*1	Acc	W-*F*1
w/o 信息注入操作	73.34	73.27	67.38	66.34
w/o 对齐操作	72.99	72.95	67.43	66.20
w/o MAT	69.47	69.53	66.95	64.38
w/o T	50.57	52.9	57.19	51.08
w/o A	71.59	71.36	66.93	65.97
w/o V	72.89	72.97	67.23	66.32
本文模型	**73.82**	**73.76**	**67.58**	**66.67**

w/o 注入操作: 移除了模型中信息注入操作用以验证保持模态多样性是否有效.

w/o 对齐操作: 移除了模型中对齐操作用以验证对齐模态语义信息是否有效.

w/o MAT: 移除了 MATransformer 用以验证考虑模态自身重要性是否有效.

w/o T/A/V: 去除文本/音频/视觉模态信息用于验证文本/音频/视觉模态重要性.

从模块消融实验中可以看出, 移除任何一个模块都导致模型性能的下降, 这说明每个模块在模型中都发挥了自身作用. 具体而言, 在移除了 MATransformer 模块后, 下降的程度是最为明显的, 尤其在 IEMOCAP 中 Acc 和 W-*F*1 分别下降了 3.87% 和 3.74%. 这说明考虑模态重要性的情况下实现多模态信息融合对于模型准确识别情感是至关重要的. 而移除了信息注入操作和对齐操作都有着一定程度的性能下降, 这证明保持模态特征多样性是有利于模型的性能. 而跨模态语义信息对齐能够加强不同模态下相同话语特征之间的相关性以及相同模态内话语情感之间的相关性, 从而突出了它们之间的共有信息.

在模态消融实验展现的结果可以发现移除任何一个模态效果都有下降, 说明每个模态都发挥着其自身的作用. 移除掉文本模态性能下降是最剧烈的, 其次是音频模态和视觉模态. 出现该情况是因为文本模态中所包含的语义信息最为丰富, 能够为情感识别提供更多的情感线索. 而音频和视觉模态的语义信息相对较少, 因此下降的程度相对较轻.

5.5 参数敏感性实验

为了验证对齐操作中对比学习损失的α、β以及不同层数的 MATransformer 对模型性能的影响, 本节对这些参数进行了参数敏感性实验.

图 3 展现的是 IEMOCAP 数据集下的实验结果, 从图中可以观察到, α的值和β的值为 0.007 和 0.09 的时候模型达到了最优的性能. 从实验结果可以看出, 适当的增大对齐力度是有利于提高模型的性能. 对于 MATransformer 的层数来说层数越深, 代表多模态信息融合深度越深. 从实验结果可以看出模型性能在第 3 层的时候达到了最优, 而后在第 4 层的时候模型效果急剧下降. 出现该情况可能是由于层数过大则会导致每个模态特征出现过度平滑的现象, 因此影响了模型的性能.

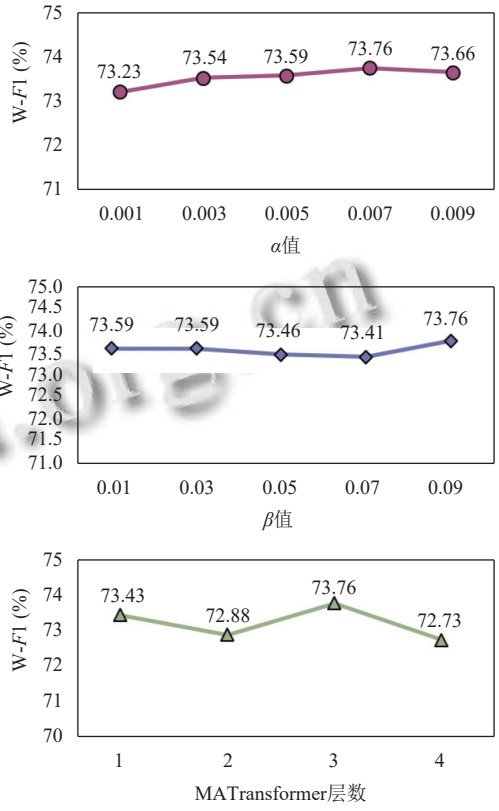

图 3 IEMOCAP 数据集的参数敏感性实验

图 4 展现的是 MELD 数据集下的实验结果, 从实验结果中可以观察到随着α和β的值增大, 模型的性能也随之增大, 直到α和β值分别在 0.005 和 0.05 的时候达到了最优效果. 该实验结果同样说明了适当增大对

齐力度是有利于性能提升. 而在 MATransformer 的层数实验结果可以看出, MELD 数据集在层数为 2 的时候达到了最优. 此后随着层数增大, 由于过度平滑的问题导致模型的性能持续下降.

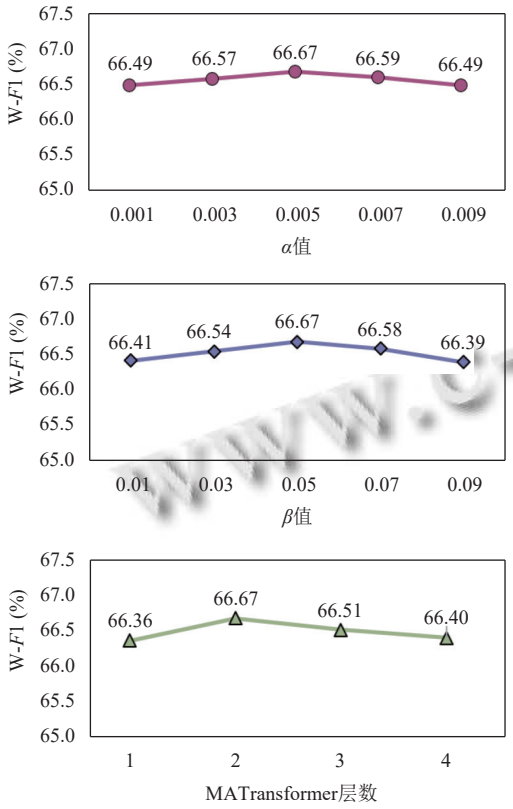

图 4 MELD 数据集的参数敏感性实验

5.6 可视化实验

为了进一步验证本文所提模型的有效性, 本文在 MELD 数据集上可视化了本文模型和基线模型 MGLRA 的多模态话语特征.

从图 5 所展现的可视化结果可以看出 MGLRA 学习到的多模态话语特征在可视化空间中混在了一起, 其中每个情感类别之间存在模糊的分割界限并且情感类别内部出现了松散的情况. 因此 MGLRA 无法很好地对情感类别进行区分. 相比较 MGLRA, 本文所提出的模型在每个情感类别上面展现出了更为清晰的分割界限并且每个情感类别内部展现出了更为紧凑的现象. 这说明本文提出的模型能够学习到更易于区分的多模态话语表示.

5.7 模态重要性实验

为了分析话语中不同模态所展现的主导作用, 本文在测试集中选取了其中的一个对话并将该对话中所有话语的模态重要性权重进行了展示. 实验结果展示在图 6 当中, 根据图中所呈现的结果, 本文得出以下结论.

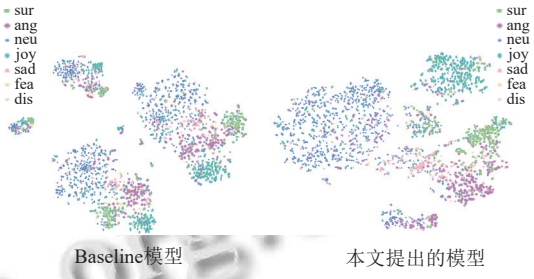

图 5 特征可视化实验

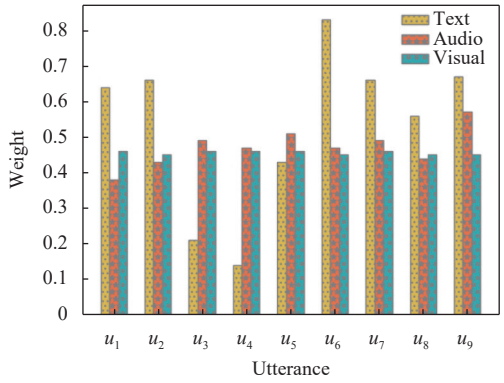

图 6 模态重要性实验

(1) 从实验的总体结果来看, 在大部分话语中, 文本模态的权重在文本、音频和视觉这 3 个模态当中居于首位. 这充分表明, 在总体层面上, 文本模态对于情感表达起着主导性作用. 这是因为文本承载着丰富且精确的语义信息, 通过词汇、语句结构以及上下文关联, 能够细腻地传达说话者的情感态度, 使得文本在情感表达方面具有较高的准确性和稳定性.

(2) 从实验的细节内容来看, u_3、u_4和u_5这 3 句话语的文本模态权重反而低于其他两个模态的权重, 这是因为在这些话语场景下, 话语文本内容中存在不含情感色彩的陈述性内容, 而音频中高低变换的声调或者是清晰的面部表情能够较完整的传达出情感.

5.8 错误分析

在本节中, 我们把所提出模型的预测结果以混淆矩阵的形式进行可视化, 从而来观察模型的实际分类情况. 从图 7 和图 8 中可以得出如下结论: (1) 在 IEMOCAP 数据集的预测结果可以发现, 本文提出的模型错误地

把“Happy”分类为“Excited”, 同时也会把“Excited”错误的分类为“Happy”. 这是因为“Happy”和“Excited”是相似的情感, 模型无法很好地区分它们之间的差异. (2) 在 MELD 数据集的预测结果可以发现除“Neutral”之外其他的情感都会被模型错误地分类为“Neutral”, 这是因为“Neutral”情感类别的话语数量在 MELD 数据集中占据了大部分, 这会导致模型学习会出现偏差从而错误地将其他情感识别为“Neutral”. 此外“Fear”和“Disgust”是属于样本数量偏少的情感类别, 容易受到数据不平衡的影响从而错误地分类成训练数量偏多的“Neutral”.

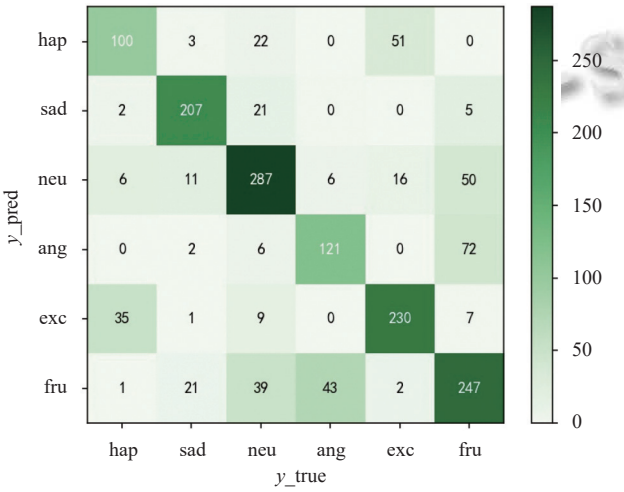

图 7 IEMOCAP 数据集的混淆矩阵

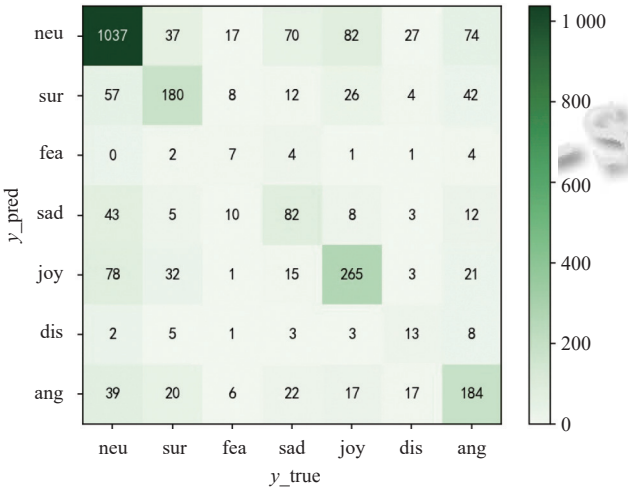

图 8 MELD 数据集的混淆矩阵

6 总 结

本文提出了一个更为全面的多模态融合框架, 名字为基于模态去异质性和自适应融合的多模态会话情感识别模型. 本文首先通过使用共享的编码器将不同模态的特征映射到共享的语义空间中来初步实现缓解多模态数据的异质性问题, 然后在共享语义空间中使用共享卷积网络来进一步学习出每个模态之间的共享语义信息, 以此来消除多模态特征之间的差距. 此外本文还会使用私有的编码器将不同模态特征映射到私有的语义空间中, 从而来学习每个模态私有的语义信息并将其注入共享语义空间中用来保持模态特征的多样性. 之后本文使用 MATransformer 来学习出每个模态的重要性从而合理的融合多个模态的信息并以此来增强单一模态表示.

在未来的工作中, 将讨论如何更好地识别数据中相似情感的问题, 以及在面对数据不平衡问题时能够保持情感识别的准确率, 从而更好地完善本文所提出的模型. 此外, 模型涉及了较为复杂的建模过程, 在推理效率方面存在不足. 如何在良好的性能的前提下, 提高模型在推理方面的效率也是本文工作未来的研究方向.

参考文献

1 DeSteno D, Gross JJ, Kubzansky L. Affective science and health: The importance of emotion and emotion regulation. Health Psychology, 2013, 32(5): 474–486. [doi: 10.1037/a0030259]

2 Chatterjee A, Narahari KN, Joshi M, *et al*. SemEval-2019 Task 3: EmoContext contextual emotion detection in text. Proceedings of the 13th International Workshop on Semantic Evaluation. Minnesota: ACL, 2019. 39–48. [doi: 10.18653/v1/S19-2005]

3 Majumder N, Hong PF, Peng SS, *et al*. MIME: Mimicking emotions for empathetic response generation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Florence: ACL, 2020. 8968–8979. [doi: 10.18653/v1/2020.emnlp-main.721]

4 Mehrabian A. A semantic space for nonverbal behavior. Journal of Consulting and Clinical Psychology, 1970, 35(2): 248–257. [doi: 10.1037/h0030083]

5 孙鹏, 彭敦陆. E2E-CER: 一种基于端到端的对话情感识别分类模型. 小型微型计算机系统, 2021, 42(2): 235–240. [doi: 10.3969/j.issn.1000-1220.2021.02.003]

6 刘若尘, 冯广, 罗良语, 等. 结合模态表征学习的多模态情感分析. 计算机系统应用, 2024, 33(5): 280–287. [doi: 10.15888/j.cnki.csa.009492]

7 Yao BY, Shi WZ. Speaker-centric multimodal fusion

networks for emotion recognition in conversations. Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul: IEEE, 2024: 8441–8445. [doi: 10.1109/icassp48485.2024.1044772]

8 Meng T, Zhang FC, Shou YT, *et al*. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 4298–4312. [doi: 10.1109/TASLP.2024.3434495]

9 Lian Z, Liu B, Tao JH. CTNet: Conversational Transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 985–1000. [doi: 10.1109/TASLP.2021.3049898]

10 Zhang XH, Li Y. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: ACL, 2023. 13099–13110. [doi: 10.18653/v1/2023.acl-long.732]

11 Majumder N, Poria S, Hazarika D, *et al*. DialogueRNN: An attentive rnn for emotion detection in conversations. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 6818–6825. [doi: 10.1609/aaai.v33i01.33016818]

12 Cho K, van Merriënboer B, Bahdanau D, *et al*. On the properties of neural machine translation: Encoder-decoder approaches. Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha: ACL, 2014. 103–111.

13 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 (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 154–164. [doi: 10.18653/v1/D19-1015]

14 Ishiwatari T, Yasuda Y, Miyazaki T, *et al*. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Punta: ACL, 2020. 7360–7370. [doi: 10.18653/v1/2020.emnlp-main.597]

15 Shen WZ, Wu SY, Yang YY, *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. Brussels: ACL, 2021. 1551–1560. [doi: 10.18653/v1/2021.acl-long.123]

16 Zhu LX, Pergola G, Gui L, *et al*. Topic-driven and knowledge-aware Transformer for dialogue emotion detection. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Brussels: ACL, 2021. 1571–1582. [doi: 10.18653/v1/2021.acl-long.125]

17 Jiao WX, Yang HQ, King I, *et al*. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 397–406. [doi: 10.18653/v1/N19-1037]

18 Ma H, Wang J, Lin HF, *et al*. A multi-view network for real-time emotion recognition in conversations. Knowledge-Based Systems, 2022, 236: 107751. [doi: 10.1016/j.knosys.2021.107751]

19 Huang Y, Du CZ, Xue ZH, *et al*. What makes multi-odal learning better than single (provably). Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021. 837.

20 Mao YZ, Liu G, Wang XJ, *et al*. DialogueTRM: Exploring multi-modal emotional dynamics in a conversation. Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana: ACL, 2021. 2694–2704. [doi: 10.18653/v1/2021.findings-emnlp.229]

21 Hu JW, Liu YC, Zhao JM, *et al*. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Florence: ACL, 2021. 5666–5675. [doi: 10.18653/v1/2021.acl-long.440]

22 Hu D, Hou XL, Wei LW, *et al*. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 7037–7041. [doi: 10.1109/ICASSP43922.2022.9747397]

23 宗林林, 周佳慧, 谢秋婕, 等. 基于超图的多模态情绪识别. 计算机学报, 2023, 46(12): 2520–2534. [doi: 10.11897/SP.J.1016.2023.02520]

24 Hu GM, Lin TE, Zhao Y, *et al*. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 7837–7851. [doi: 10.18653/v1/2022.emnlp-main.534]

25 Tsai YHH, Bai SJ, Liang PP, *et al*. Multimodal Transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6558–6569. [doi: 10.18653/v1/p19-1656]

26 Yuan ZQ, Li W, Xu H, *et al*. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. Proceedings of the 29th ACM International Conference on Multimedia. Chengdu: ACM, 2021. 4400–4407.

27 Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]

28 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186.

29 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

30 Eyben F, Wöllmer M, Schuller B. OpenSMILE: The munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia. Firenze: ACM, 2010. 1459–1462. [doi: 10.1145/1873951.1874246]

31 Iandola F, Moskewicz M, Karayev S, *et al*. DenseNET: Implementing efficient convnet descriptor pyramids. arXiv: 1404.1869, 2014.

32 Ong D, Su J, Chen B, *et al*. Is discourse role important for emotion recognition in conversation? Proceedings of the 36th AAAI Conference on Artificial Intelligence. Pennsylvania: AAAI, 2022. 11121–11129. [doi: 10.1609/aaai.v36i10.21361]

33 Busso C, Bulut M, Lee CC, *et al*. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335–359. [doi: 10.1007/s10579-008-9076-6]

34 Poria S, Hazarika D, Majumder N, *et al*. MELD: A multimodal multi-party dataset for emotion recognition in conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 527–536. [doi: 10.18653/v1/P19-1050]

35 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

36 Yang HZ, Gao XQ, Wu JL, *et al*. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. Findings of the Association for Computational Linguistics: ACL 2023. Toronto: ACL, 2023. 6267–6281. [doi: 10.18653/v1/2023.findings-acl.390]

(校对责编: 张重毅)