计算机科学
COMPUTER SCIENCE

基于分步协作融合表示的情感分类方法

高龙, 李旸, 王素格

引用本文

高龙, 李旸, 王素格. 基于分步协作融合表示的情感分类方法[J]. 计算机科学, 2025, 52(9): 313-319.

GAO Long, LI Yang, WANG Suge. Sentiment Classification Method Based on Stepwise Cooperative Fusion Representation [J]. Computer Science, 2025, 52(9): 313-319.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700161

基于分步协作融合表示的情感分类方法

高 龙[1] 李 旸[2] 王素格[1,3]
1 山西大学计算机与信息技术学院 太原 030006
2 山西财经大学金融学院 太原 030006
3 山西大学计算智能与中文信息处理教育部重点实验室 太原 030006
(glong202202@163.com)

摘 要 多模态情感分析任务旨在通过各种异构模态(如语言、视频和音频)感知和理解人类的情感,但不同模态间存在着复杂的关联。现有的大多数方法将多个模态特征直接融合,忽略了不同步的模态融合表示在情感分析中的贡献不同。针对上述问题,提出了一种基于分步协作融合表示的情感分类方法。首先,利用降噪瓶颈模型对音视频中的噪声和冗余进行过滤,通过Transformer完成对音视频两种模态的交互融合,建立音视频融合的低级特征表示;进一步利用跨模态注意力机制,强化文本模态对音视频模态的低级融合表示,构建音视频融合的高级特征表示。其次,设计一个新颖的模态融合层将多级特征表示引入预训练模型T5中,建立以文本为中心的多模态融合表示。最后,将低级特征表示、高级特征表示以及以文本为中心的特征融合表示进行联合,实现了多模态数据的情感判别。在两个公开数据集CMU-MOSI和CMU-MOSEI上进行实验,结果表明所提出的方法相比已有基线模型ALMT在Acc-7指标上分别提高0.1和0.17,表明了分步协作融合表示能够提高多模态情感分类性能。

关键词:多模态融合;情感分析;瓶颈机制;注意力机制;预训练模型
中图分类号 TP391

Sentiment Classification Method Based on Stepwise Cooperative Fusion Representation

GAO Long[1], LI Yang[2] and WANG Suge[1,3]
1 School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
2 School of Finance, Shanxi University of Finance and Economics, Taiyuan 030006, China
3 Key Laboratory Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China

Abstract The goal of multimodal sentiment analysis is to perceive and understand human emotions through various heterogeneous modalities, such as language, video, and audio. However, there are complex correlations between different modalities. Most existing methods directly fuse multiple modality features, they overlook the fact that asynchronous modality fusion representations contribute differently to sentiment analysis. To address the above issues, this paper proposes a sentiment classification method based on stepwise collaborative fusion representation. Firstly, a denoising bottleneck model is used to filter out noise and redundancy in the audio and video, and the two modalities are fused through Transformer, establishing a low-level feature representation of the audio-video fusion. Then, a cross-modal attention mechanism is utilized to enhance the audio-video modalities with the text modality, constructing a high-level feature representation of the audio-video fusion. Secondly, a novel multimodal fusion layer is designed to incorporate multi-level feature representations into the pre-trained T5 model, establishing a text-centric multimodal fusion representation. Finally, the low-level feature representation, high-level feature representation, and text-centric feature fusion representation are combined to achieve sentiment classification of multimodal data. Experimental results on two public datasets, CMU-MOSI and CMU-MOSEI indicate that the proposed method improves the Acc-7 metric by 0.1 and 0.17 compared to the existing baseline model ALMT, demonstrating that stepwise collaborative fusion representation can enhance multimodal sentiment classification performance.

Keywords Multimodal fusion, Sentiment analysis, Bottleneck mechanism, Attention mechanism, Pre-trained model

到稿日期:2024-07-25 返修日期:2024-10-18
基金项目:国家自然科学基金(62106130,62376143,62076158);山西省基础研究计划(20210302124084);山西省高校科技创新计划(2021L284)
This work was supported by the National Natural Science Foundation of China(62106130, 62376143, 62076158), Basic Research Program in Shanxi(20210302124084) and Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi(2021L284).
通信作者:李旸(liyangprimrose@163.com)

1 引言

多模态情感分析(MSA)旨在从视频、音频和语言等多种类型的数据中识别人类的情感态度。如图 1 所示,文本是中性的描述且说话人语气平静,但人物的面部表情一直保持微笑,最终情感极性预测为积极的。因此,需要研究多种模态的有效融合技术。

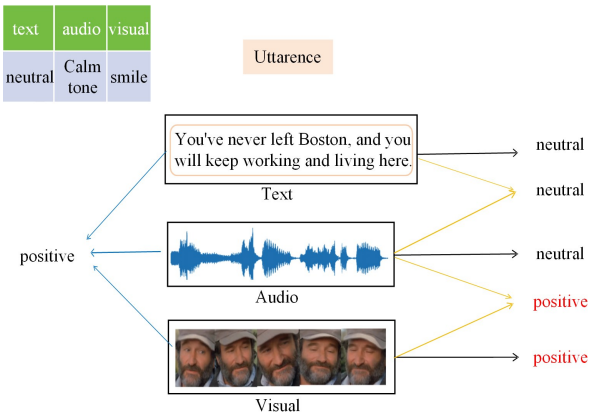

图 1 人物独白示例

Fig. 1 Example of monologue

模态融合的主要目标是缩小语义子空间中的分布差距,同时保持模态特定语义的完整性[1]。以多模态融合为中心的方法主要关注直接设计复杂的融合机制,从基于简单操作到张量融合模型,再到利用跨模态注意力机制获得多模态数据的联合表示。但在以文本为中心的模态融合任务中存在以下几个问题。1)音频和视频模态的信息存在大量的冗余和噪声[2],尽管 Wu 等[3]提出了基于视频多模态融合的降噪瓶颈模型,但提取的有效辅助模态信息未被文本充分利用。2)音频和视频模态对文本的影响是不同的,尽管 Zhang 等[4]利用不同尺度的语言特征引导模型学习较为有效的超模态表示,但并未考虑不同层次的辅助模态融合表示与不同尺度的语言特征的复杂关联与映射关系。3)现有的研究工作将先进的预训练模型如 T5 或 BERT 作为文本编码器,Yu 等[5]提出了基于模态对齐的语音文本对话预训练模型,Yang 等[6]通过对生成式多模态预训练模型提示调优,但它们都缺乏接收视频和音频模态的组件,无法得到以文本为中心的多模态融合表示。

针对上述问题,本文提出了一种基于分步协作融合表示的多模态情感分类方法(Multimodal Sentiment Classification for Stepwise Collaborative Fusion Representations,SCFR),以改善情感分析的性能。首先,对于有效的辅助模态表示,本文引入了一个瓶颈模块,使不同模态之间的冗余和噪声信息尽可能被过滤,在此基础上,利用跨模态注意力强化文本模态对音视频的低级融合表示。其次,本文设计了一个模态融合层,将两步融合特征表示引入文本预训练模型,文本特征基于注意力机制与两步特征表示融合,捕获更多与音视频模态相关的信息。最后,经过全连接层,将音视频模态的低级特征表示、高级特征表示以及以文本为中心的多模态融合特征表示进行相加操作,实现多模态数据的情感预测。本文方法在两个数据集上进行了大量的实验,实验结果表明,该模型与当前先进的方法相比,取得了较好的性能。本文的主要贡献如下:

1)提出了一种基于分步协作融合表示的多模态情感分类方法 SCFR,通过分步协作的方法,建模了音视频模态的低级特征表示和高级特征表示,在此基础上,进一步建立了以文本为中心的多模态融合特征表示,提升了情感分析的性能。

2)设计了一个模态融合层,将多步融合特征表示引入文本预训练模型,建立了不同层次的辅助模态融合表示与不同尺度的语言特征的复杂关联与映射关系。

3)在 CMU-MOSEI 和 CMU-MOSI 两个公开数据集上进行了实验,实验结果表明,本文模型优于现有基线模型,验证了本文方法的有效性。

2 相关工作

MSA 通常利用 3 种模态(音频、视觉、文本)的信息来判别人类情感。

对于以多模态表示学习为中心的方法,Tsai 等[7]建立了一个推理网络和一个生成网络,后者具有中间模态特定因素,可以促进融合过程中的重建和判别损失。Hazarika 等[1]将多模态表示学习视为域自适应任务,利用对抗学习建立模态不变和特定的表示,在数据集中实现了其先进的性能。Han 等[8]提出一个名为 MMIM 的框架,将互信息的概念引入多模态情感分析中,建模了分层互信息最大化,提升了多模态融合的性能。Guo 等[9]结合语言与非语言信息之间的跨模态交互信息增强了语言表示,但忽略了音频和视频中存在与情感无关的冗余信息。Wu 等[3]提出了一种细粒度视频多模态融合的降噪瓶颈模型,该模型可以去除嘈杂和冗余的噪声,并在音频和文本输入中捕获显著特征。Sun 等[10]引入了一种基于元学习的方法学习更好的单模态表示,并将其用于随后的多模态融合。Sun 等[11]提出一个通用的、统一的框架 EMT-DLFR,实现了鲁棒多模态特征表示。

对于以多模态融合为中心的方法,Zadeh 等[12]提出了一种多模态融合方法(TFN),利用张量的笛卡尔积运算建模了不同模态之间的关系。Huang 等[13]提出基于对齐的方法,引入了基于多模态 Transformer 的对齐序列,并建模跨模态元素之间的长程依赖关系。然而,这些方法仅直接利用单一模态的信息融合,忽略了模态内的特征具有的特性。Rahman 等[14]提出了一种多模态门控组件,使得 BERT 模型在不改变结构的前提下可以动态融合多模态信息。Liang 等[15]尝试了基于跨模态注意力机制,实现了从一种模态到另一种模态的潜在适应。Luo 等[16]提出了一种多尺度融合方法,用于对齐来自多种模态的不同粒度信息。Sun 等[17]提出了一种促进模态独立的模型,并同时使用单模态标签和多模态标签进行训练。Shi 等[18]设计了一种基于双向多头交叉注意力层的多模态融合模型。Zhang 等[4]提出了自适应语言引导的多模态转换器,通过多模态融合获得互补和联合表示,从而提高情感分析的性能。

受以上工作的启发,本文提出一种基于分步协作融合表

示的多模态情感分类方法。首先，设计了聚合模块与跨模态注意力机制，建立了音视频模态融合的低级表示与高级特征表示进一步，设计了模态融合层将两步融合特征嵌入预训练模型中，建模了以文本为中心的融合特征表示；在此基础上，将音视频模态的原始特征表示、高级特征表示以及以文本为中心的多模态融合特征表示进行融合，实现多模态情感判别。

3　本文模型

多模态情感分析涉及 3 种主要模态，即文本(t)、视频(v)和音频(a)。本文提出的 SCFR 结构如图 2 所示。该方法的核心思想是，首先，将两个单模态特征音频和视频通过降噪瓶颈模块去除噪声和冗余，再基于 Transformer 进行融合，建立音视频模态的低级特征表示；其次，利用跨模态注意力机制，建模文本对于音视频低级特征的强化表示，再将多步的音视频融合特征和文本信息融合，通过门控机制将其注入 T5 模型中，为文本预训练模型提供更多与情感相关的信息。模型由 3 个模块组成：降噪瓶颈模块(见 3.1 节)、跨模态注意力模块(见 3.2 节)、三模态融合模块(见 3.3 节)。

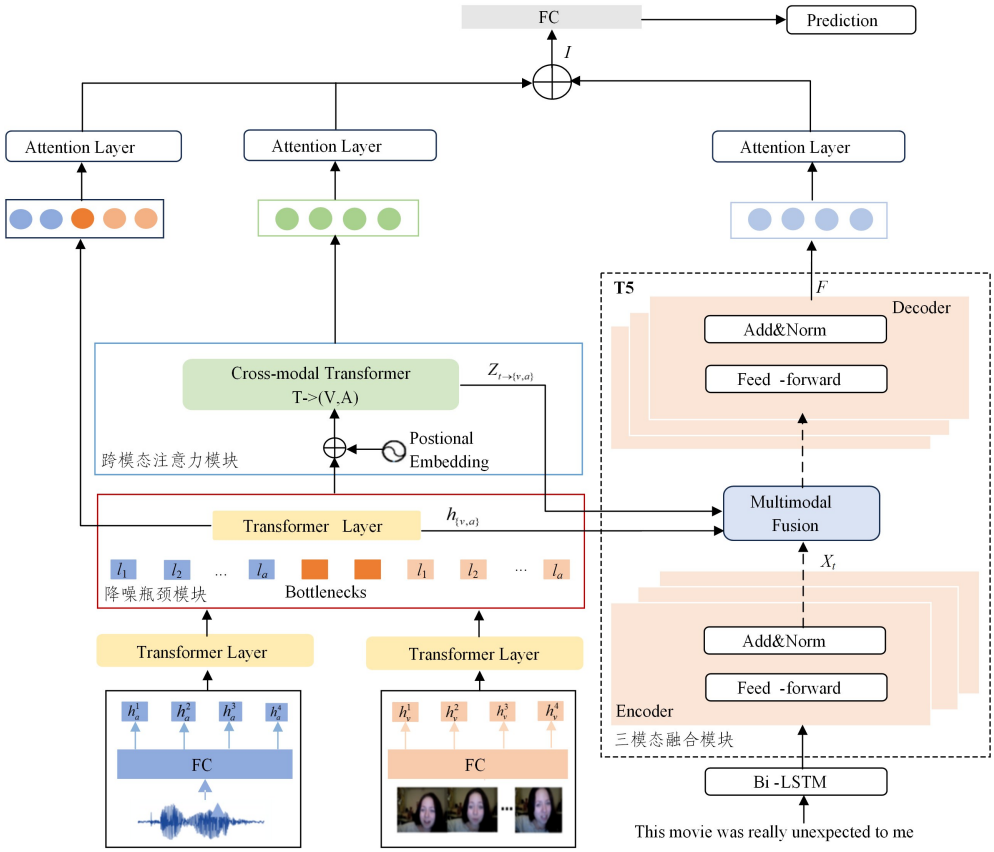

图 2　SCFR 整体架构图

Fig. 2　Overall architecture of SCFR

3.1　辅助模态低级特征表示

对于音频模态 $\boldsymbol{X}_a \in R^{l_a \times d_a}$ 和视频模态 $\boldsymbol{X}_v \in R^{l_v \times d_v}$，本文使用 COVAREP[19] 与 Openface/Facet[20] 提取音频与视频浅层特征后利用全连接层，可以获得每种模态的特征表示，具体如下：

$$\begin{aligned}\boldsymbol{h}_v &= W_v X_v + b_v \\ \boldsymbol{h}_a &= W_a X_a + b_a\end{aligned} \tag{1}$$

将 $\boldsymbol{h}_a$ 和 $\boldsymbol{h}_v$ 作为 Transformer 编码器的输入。一个编码器由 L 个 Transformer 层组成，每个 Transformer 层包括多头自注意力机制(MSA)、层归一化(LN)和多层感知机(MLP)块，并应用残差连接，分别定义音频和视频的 Transformer 的第 $l+1$ 层，具体如下：

$$\boldsymbol{h}_a^{l+1} = Transformer(\boldsymbol{h}_a^l) \tag{2}$$

$$\boldsymbol{h}_v^{l+1} = Transformer(\boldsymbol{h}_v^l) \tag{3}$$

为了过滤音频和视频中的噪声和冗余信息，采用降噪瓶颈模块进行音频和视频的底层特征融合。降噪瓶颈模块具体表示为引入 B 个瓶颈模块 $\boldsymbol{R}_f = [\boldsymbol{R}_f^1, \boldsymbol{R}_f^2, \cdots, \boldsymbol{R}_f^B]$。在该模块中，单模态特征 $\boldsymbol{h}_a$ 和 $\boldsymbol{h}_v$ 需要与降噪瓶颈 $\boldsymbol{R}_f$ 拼接，再经过 Transformer 进行信息交互融合。对于第 l 层，融合结果计算如下：

$$[\boldsymbol{h}_v^{l+1} \parallel \overset{\wedge}{\boldsymbol{R}}_f^{l+1}] = Transformer([\boldsymbol{h}_v^l \parallel \boldsymbol{R}_f^l]; \theta_v) \tag{4}$$

$$[\boldsymbol{h}_a^{l+1} \parallel \overset{\wedge}{\boldsymbol{R}}_f^{l+1}] = Transformer([\boldsymbol{h}_a^l \parallel \boldsymbol{R}_f^l]; \theta_a) \tag{5}$$

$$\overset{\wedge}{\boldsymbol{R}}_f^{l+1} = Avg_i(\boldsymbol{R}_f^{l+1}) \tag{6}$$

其中，l 表示 Transformer 层数，$\parallel$ 表示拼接操作。如式(4)和式(5)所示，$\boldsymbol{R}_f$ 分别与音频和视频特征同时更新，同样 $\boldsymbol{h}_a$ 和 $\boldsymbol{h}_v$ 只能通过 $\boldsymbol{R}_f$ 交换信息。$\boldsymbol{R}_f^l$，$\boldsymbol{h}_a^l$ 和 $\boldsymbol{h}_v^l$ 分别为第 l 层的降噪瓶颈模块、音频特征和视频特征表示。

在跨模态更新中，将每个跨模态的最终融合表示在式(6)中进行平均。这种方法可以提高或保持多模态融合性能，同

时降低计算复杂性。最终，输出最后一层的表示 $\boldsymbol{h}_{\{v,a\}}$ 为融合结果。

3.2 辅助模态高级特征表示

现有的方法主要是多个模态特征直接融合，未考虑模态间的相关性和交互作用。跨模态注意机制利用来自源模态的信息，通过学习源模态与目标模态之间的成对注意力来增强目标模态表示，可以实现文本模态对于视频音频模态融合特征的强化表示。

为了保留时间信息，将位置信息 PE 增强到 3.1 节得到的音视频模态低级特征表示 $\boldsymbol{h}_{\{v,a\}}$ 中。

$$\boldsymbol{Z}_{\{v,a\}}^{[0]}=h_{\{v,a\}}+PE(T_{\{v,a\}},d) \tag{7}$$

其中，$PE(T_{\{v,a\}},d)$ 表示为每个索引的位置进行向量编码，$\boldsymbol{Z}_{\{v,a\}}^{[0]}$ 表示不同模态下获取的底层特征中的位置信息。

为了让音视频融合表示能够接收来自文本模态的信息，本文基于跨模态注意力机制，设计了跨模态的 Transformer。具体计算如下：

$$\boldsymbol{Z}_{t\to[v,a]}^{[0]}=\boldsymbol{X}_t^{[0]} \tag{8}$$

$$\bar{\boldsymbol{Z}}_{t\to[v,a]}^{[l]}=Cross-Transformer(LN(\boldsymbol{Z}_{t\to[v,a]}^{[l-1]}),LN(\boldsymbol{Z}_{\{v,a\}}^{[0]}))+LN(\boldsymbol{Z}_{t\to[v,a]}^{[l-1]}) \tag{9}$$

$$\boldsymbol{Z}_{t\to\{v,a\}}^{[l]}=f_{\theta_{t\to\{v,a\}}^{[l]}}(LN(\bar{\boldsymbol{Z}}_{t\to\{v,a\}}^{[l]})^{\mathrm{FFN}}+LN(\bar{\boldsymbol{Z}}_{t\to\{v,a\}}^{[l]})) \tag{10}$$

其中，$\boldsymbol{X}_t^{[0]}$ 为文本的表示特征，f_θ 是由参数 θ 表示的位置前馈层，LN 表示层归一化，$\bar{\boldsymbol{Z}}$ 表示中间状态，$(\bar{\boldsymbol{Z}}_{t\to\{v,a\}}^{[l]})^{\mathrm{FFN}}$ 表示位置前馈层的变换，$\boldsymbol{Z}_{t\to\{v,a\}}^{[l-1]}$ 表示第 $l-1$ 层获得的被文本强化后的音视频模态融合特征。

此过程中，跨模态的 Transformer 学习在不同的模态之间找到关联，有效地将文本模态的特征增强到音视频融合特征中。

3.3 三模态融合的预训练模型

在以文本为中心的多模态融合任务中，音频和视频模态信息会对文本的含义提供更多的信息，从而提升其在语义空间中表示的性能[6]。目前，预训练模型(例如 T5)主要用于文本编码，其自身未考虑音频和视频模态信息的融合。本文基于 3.1 和 3.2 节得到的两步融合特征表示，引入了一个多模态融合层(Multimodal Fusion)，将模态信息融合到预训练模型中，使得多个级别的音频-视频融合特征与文本信息进行融合，以更好地适应多模态输入。

3.3.1 多模态融合层

如图 3 所示，首先利用 3.1 节得到的音频-视频低级融合特征表示 $\boldsymbol{h}_{\{v,a\}}$ 和 3.2 节得到的音频-视频高级融合特征表示 $\boldsymbol{Z}_{t\to\{v,a\}}$ 分别与文本向量 $\boldsymbol{X}_t$ 的编码进行拼接，获得双模态因子 $[\boldsymbol{X}_t:\boldsymbol{h}_{\{v,a\}}]$ 和 $[\boldsymbol{X}_t:\boldsymbol{Z}_{t\to\{v,a\}}]$，然后利用它们构建两个门控向量 $\boldsymbol{g}_{va}$ 和 $\boldsymbol{g}_{t\to\{v,a\}}$，具体公式如下：

$$\boldsymbol{g}_{va}=R(\boldsymbol{W}_{va}[\boldsymbol{X}_t:\boldsymbol{h}_{\{v,a\}}]+b_{va}) \tag{11}$$

$$\boldsymbol{g}_{t\to\{v,a\}}=R(\boldsymbol{W}_{t\to\{v,a\}}[\boldsymbol{X}_t:\boldsymbol{Z}_{t\to\{v,a\}}]+b_{t\to\{v,a\}}) \tag{12}$$

其中，$\boldsymbol{W}_{va}$ 和 $\boldsymbol{W}_{t\to\{v,a\}}$ 分别是低级与高级融合特征表示的权重矩阵，$b_{t\to\{v,a\}}$ 和 b_{va} 是标量偏差，$R(X)$ 是激活函数。

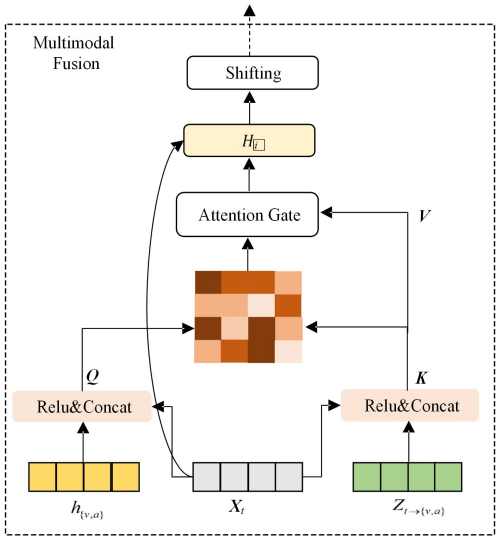

图 3 多模态融合层

Fig. 3 Multimodal fusion layer

本文定义 $\boldsymbol{Q}_{va}=\boldsymbol{W}_Q g_{va}$，$\boldsymbol{K}_{t\to\{v,a\}}=\boldsymbol{W}_K g_{t\to\{v,a\}}$ 和 $\boldsymbol{V}_{t\to\{v,a\}}=\boldsymbol{W}_V g_{t\to\{v,a\}}$ 分别为查询、键和值。其中，$\boldsymbol{W}_Q\in R^{d_{va}\times d_k}$，$\boldsymbol{W}_K\in R^{d_{t\to\{v,a\}}\times d_k}$ 和 $\boldsymbol{W}_V\in R^{d_{t\to\{v,a\}}\times d_v}$。经过多头注意力机制计算得到对文本含义产生影响的偏移向量 $\boldsymbol{H}$：

$$\begin{aligned}\boldsymbol{H}&=MH\text{-}ATT(\boldsymbol{Q}_{va},\boldsymbol{K}_{t\to\{v,a\}},\boldsymbol{V}_{t\to\{v,a\}})\\&=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_{va}(\boldsymbol{K}_{t\to\{v,a\}})^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V}_{t\to\{v,a\}}\\&=\mathrm{softmax}\left(\frac{\boldsymbol{W}_Q g_{va}\boldsymbol{W}_K^{\mathrm{T}}(g_{t\to\{v,a\}})^{\mathrm{T}}}{\sqrt{d_K}}\right)_v g\boldsymbol{W}_{t\to\{v,a\}}\end{aligned} \tag{13}$$

其中，$MH\text{-}ATT$ 表示多头注意力，Softmax(·)表示权重归一化操作。

为了获得文本在语义空间中的新位置，将文本向量 $\boldsymbol{X}_t$ 和非语言向量 $\boldsymbol{H}$ 加权求和得到多模态向量 $\bar{\boldsymbol{X}}_t$，计算如下：

$$\bar{\boldsymbol{X}}_t=\boldsymbol{X}_t+\alpha\boldsymbol{H} \tag{14}$$

$$\alpha=\min\left(\frac{\|\boldsymbol{X}_t\|_2}{\|\boldsymbol{H}\|_2}\beta,1\right) \tag{15}$$

其中，β 是交叉验证过程选择的超参数，$\|\boldsymbol{X}_t\|_2$ 和 $\|\boldsymbol{H}\|_2$ 分别表示 $\boldsymbol{X}_t$ 和 $\boldsymbol{H}$ 的 L2 范数，比例因子 α 用于控制偏移向量 $\boldsymbol{H}$ 的效果保持在理想的范围之内。

3.3.2 多模态融合层嵌入预训练模型过程

在多步的辅助模态特征表示与文本特征表示进行融合后，本节提出将多模态融合层嵌入预训练模型中。将音频和视频信息注入 T5 中，可以探测到大量与预训练文本知识相关的信息，获得以文本为中心的多模态融合表示。T5 作为 SCFR 结构的骨干，包含多个堆叠的 Transformer 层，每个用于编码器和解码器的 Transformer 层中都包含一个前馈层，多模态融合层设置在前馈层之后。T5 的第一个 Transformer 层中的多模态融合层接收一个三元组 $M=\{\boldsymbol{h}_{\{v,a\}},\boldsymbol{Z}_{t\to\{v,a\}},\bar{\boldsymbol{X}}_t\}$ 作为输入。多模态融合层接收到模态表示的三元组，并将得到的多模态表示映射到下一层的输入。第 j 个 Transformer 层的表示如下：

$$\bar{\boldsymbol{F}}=Cross\text{-}Attention[\boldsymbol{F}^{(j-1)},\boldsymbol{h}_{\{v,a\}},\boldsymbol{Z}_{t\to\{v,a\}}] \tag{16}$$

$$(\bar{\boldsymbol{F}})^d=\sigma(W^d\bar{\boldsymbol{F}})+b^u \tag{17}$$

$$(\bar{\boldsymbol{F}})^u=W^u\ (\bar{\boldsymbol{F}})^d+b^u \tag{18}$$

$$\boldsymbol{F}^j=W((\bar{\boldsymbol{F}})^u\odot\boldsymbol{F}^{(j-1)}) \tag{19}$$

其中，σ 为 Sigmoid 函数；$\{W^d,W^u,W,b^d,b^u\}$ 是可学习的参数；$\boldsymbol{F}^0=\bar{\boldsymbol{X}}_t^0$，$\bar{\boldsymbol{X}}_t^0$ 是 T5 的第一层 Transformer 编码的文本表示；$\boldsymbol{F}^{(j-1)}$ 表示经过 $j-1$ 层 Transformer 后的融合表示；$\odot$ 表示元素相加，融合层的输出直接传递到归一化层。

3.4　情感预测以及损失函数

为了将多模态信息进行进一步融合，本文通过自注意力机制得到视觉-音频低级特征表示 $\boldsymbol{h}_{\{v,a\}}$，通过跨模态注意力机制得到视觉-音频高级特征表示 $\boldsymbol{Z}_{t\to\{v,a\}}$，通过 T5 模型预训练得到多模态融合特征 $\boldsymbol{F}$，将三者进行相加，获得多模态的表示 $\boldsymbol{I}$，具体如式(20)所示：

$$\boldsymbol{I}=\boldsymbol{h}_{\{v,a\}}+\boldsymbol{Z}_{t\to\{v,a\}}+\boldsymbol{F} \tag{20}$$

最终情感预测通过一个全连接层获得：

$$pre=\mathrm{softmax}(W_1\boldsymbol{I}+b_1) \tag{21}$$

其中，softmax 函数用于捕获输入句子的情感类别分布表示，W_1 和 b_1 为可调整的权重和偏置。

最后，SCFR 模型在分类中采用交叉熵作为损失函数，其损失计算如下：

$$\zeta=-\frac{1}{N}\sum_{i=1}^{n}(pre_i\log y_i+(1-pre_i)\log(1-y_i)) \tag{22}$$

其中，y_i 代表第 i 个样本的真实标签。

4　实验

4.1　数据集

本文在 CMU-MOSI[21] 和 CMU-MOSEI[22] 两个公开可用的数据集上评估所提模型的性能。CMU-MOSI 是一个包含 2199 个短视频片段的数据集，包括视觉、音频和语言模态。每个片段都用－3(表示强烈负面)到＋3(表示强烈正面)的情感评分手工标注。CMU-MOSEI 数据集包括从 YouTube 收集的来自 1000 个不同演讲者的 22852 个注释视频片段(话语)和来自在线视频分享的 250 个主题。同样，每个实例都标记了从－3 到＋3 的情绪分数。从－3 到＋3 的情绪得分表示从最消极到最积极。在数据集划分方面，分别按照 6∶1∶3 和 7∶1∶2 的比例将 CMU-MOSI 和 CMU-MOSEI 划分为训练集。验证集和测试集，数据统计如表 1 所列。

表 1　MOSI 和 CMU-MOSEI 数据集统计

Table 1　MOSI and CMU-MOSEI datasets statistics

Dataset	# Train	# Valid	# Test	# All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856

4.2　实验设置

本文提出的 SCFR 模型使用 Adam 优化器进行训练，学习率设置为 0.001，批大小为 32，epoches 设置为 150。

本文模型采用 F1-Score、二分类准确率 Acc-2、七分类准确率 Acc-7、平均绝对误差 MAE，以及皮尔逊相关系数 Corr 作为评估指标。除 MAE 外，值越高表示这项指标的性能越好。

4.3　基线模型

为了验证本文模型的有效性，将其与现有的多模态情感分析方法进行比较。由于这些方法较多，本文将其分为以模态交互为中心的方法(TFN[12]，LMF[23]，MulT[24]，ICCN[2]，PMR[25] 和 DBF[3])、以多模态融合为中心的方法(MFM[7]，MISA[1] 和 Self-MM[26])、以特征融合为中心的方法(MAG-BERT[14]，MMIM[8]，UniMSE[27] 和 ALMT[10])。

4.4　实验结果与分析

将本文方法与基线模型在 CMU-MOSI 和 CMU-MOSEI 数据集上进行比较实验，结果如表 2 和表 3 所列。

表 2　CMU-MOSI 数据集上的对比实验

Table 2　Comparison of experimental on CMU-MOSI dataset

method	*MAE*	*Corr*	*F1-Score*	*Acc-7*	*Acc-2*
TFN	0.901	0698	－/80.70	34.90	－/80.80
LMF	0.917	0.695	－/82.40	33.20	－/82.50
MFM	0.877	0.706	－/81.60	35.40	－/81.70
ICCN	0.862	0.714	－/83.00	39.00	－/83.00
MulT	0.871	0.698	－/82.80	40.00	－/83.00
MISA	0.783	0.761	81.70/83.60	42.30	81.80/83.40
PMR	－	－	－/83.40	40.60	－/83.60
MAG-Bert	0.727	0.781	82.50/84.61	43.62	82.37/84.43
Self-MM	0.712	0.795	83.68/84.91	45.79	82.54/84.77
MMIM	0.700	0.800	84.00/85.98	46.65	84.14/86.06
DBF	0.693	0.801	85.10/86.90	44.8	85.10/86.90
Uni-MSE	0.691	**0.809**	85.83/86.42	48.68	85.85/86.90
ALMT	0.683	0.805	84.57/86.47	49.42	84.55/86.43
SCFR	**0.680**	0.801	**85.92/87.12**	**49.52**	**85.91/87.10**

表 3　CMU-MOSEI 数据集上的对比实验

Table 3　Comparison of experimental on CMU-MOSEI dataset

method	*MAE*	*Corr*	*F1-Score*	*Acc-7*	*Acc-2*
TFN	0.593	0.700	－/82.10	50.20	－/82.50
LMF	0.677	0.695	－/82.10	48.00	－/82.00
MFM	0.717	0.706	－/84.40	51.30	－/84.30
ICCN	0.565	0.713	－/84.20	51.60	－/84.20
MulT	0.580	0.703	－/82.30	51.80	－/82.50
PMR	－	－	－/82.60	52.50	－/83.30
MISA	0.555	0.756	83.80/85.30	52.20	83.60/85.50
MAG-Bert	0.543	0.755	82.77/84.71	52.67	82.51/84.82
Self-MM	0.529	0.767	83.00/84.90	53.50	82.70/85.00
MMIM	0.526	0.772	82.70/85.99	54.24	82.20/86.00
DBF	0.523	0.772	84.80/86.20	54.2	84.30/86.40
Uni-MSE	**0.523**	0.773	85.79/87.46	54.39	**85.86/87.50**
ALMT	0.526	0.779	85.19/86.86	54.28	84.78/86.79
SCFR	0.528	**0.783**	**85.92/87.64**	**54.45**	85.37/87.21

由表 2 和表 3 可以得出：

1)本文 SCFR 模型与基线方法相比，几乎在所有指标上都取得了相当或最好的结果。在较难和细粒度的情感分类任务(Acc-7)上，本文模型与基线模型相比，性能有显著提升，表明了分步协作融合表示在多模态情感分析任务中是有效的。

2) SCFR 相比基于表示学习的模型如 MISA，在 CMU-MOSI 数据集上的所有评估指标均有较大提升，这表明 MISA 学习的多模态表示可能不够精细，而 SCFR 可以减小不同模态中的分布差异。

3)以 CMU-MOSEI 数据集为例，SCFR 与 DBF 方法相比，分别在 Acc-7 和 Corr 上实现了 0.46%和 1.42%的提升，这表明在过滤音频和视频模态噪声的同时，进一步强化模态特定的表示是必要的。

4)本文模型 SCFR 与 ALMT 方法相比,在两个数据集上的评估指标 Acc-2 上分别实现了 1.61/0.78 和 0.70/0.48 的提升,与 Uni-MSE 方法相比取得了相当的性能,表明了预训练模型引入其他模态信息可以提升情感分析的性能。

4.5 消融实验

为了验证各组件对 SCFR 的影响,本文在 CMU-MOSEI 数据集上设计一系列消融实验,结果如表 4 所列,其中-w/o 表示删除单个模态的影响。最后 3 行描述了去除瓶颈模块、跨模态注意力和模态融合层的影响。

表 4 消融实验结果

Table 4 Ablation experiment results

Model	*Acc-7*	*Acc-2*	*F1*	*MAE*	*Corr*
Ours	**54.45**	**85.37/87.21**	**85.92/87.64**	**0.528**	**0.783**
-w/o A	51.92	83.62/85.10	82.78/85.11	0.538	0.780
-w/o V	51.73	83.39/85.02	82.36/84.97	0.547	0.778
-w/o L	45.24	72.60/74.27	71.69/73.64	0.806	0.516
-bottleneck	52.06	82.76/84.72	83.09/84.96	0.533	0.759
-FvcA	52.85	83.63/85.07	83.17/85.32	0.541	0.756
-Multimodal Fusion	53.35	84.02/86.22	83.26/85.21	0.545	0.769

从多模态信号中一次删除一种模态,观察其对模型性能的影响。观察表 4 中结果可得,删除任何一种模态都会导致性能的下降,表明 SCFR 模型可以从不同模态中学习互补信息。此外,在去除音频和视频输入后,SCFR 的性能仍然相对较高,验证了文本具有更高的信息密度和辅助模态的特定表示对情感判别的积极作用。

1)-bottleneck 模型相比 SCFR 在 CMU-MOSEI 数据集上的 Corr 下降了 0.024,MAE 增加了 0.005,表明了在保持模态特定表示的同时去除冗余信息的必要性。

2)-FvcA 模型相比 SCFR 在 CMU-MOSEI 数据集上的 Acc-2 下降了 1.74/2.14,MAE 增加了 0.013,表明了跨模态注意对多模态学习有明显的好处,具有较好的表征学习效果。

3)-Multimodal Fusion 模型相比 SCFR 在 CMU-MOSEI 数据集上的 Acc-7 下降了 1.1,Acc-2 下降了 1.35/0.99,F1 下降了 2.66/2.43,MAE 增加了 0.017,Corr 降低了 0.014,表明 PMF 模块可以探测到大量与文本知识相关的信息,在模态交互中起到了重要作用。

4.6 实例分析

为了进一步展示本文模型 SCFR 的性能,表 5 列出了在 CMU-MOSI 数据集上的一些示例,与 Ground Truth 相比,可以看出本文 SCFR 模型相比基线模型 DBF,情感预测的性能较为准确。

在样例 2 中,原始数据的文本包含积极的词汇(如 love),说话人语调快速且表情惊讶,3 种模态信息的极性表达均为正面,SCFR 和 DBF 因此准确预测出情感的强度。在样例 1 和样例 4 中,文本中蕴含着积极的描述,但紧张的声音和矛盾的表情帮助 SCFR 预测出更加接近参考值的情感强度。在样例 3 中,原始数据的文本是中性的描述,说话人语调平稳且表情平淡,DBF 最终给出错误的情感预测,但 SCFR 综合考虑多个模态信息,成功预测出正确的情感极性,进一步证明了 SCFR 模型采用多步的模态融合特征进行情感分析的策略是有效的。

表 5 实例分析

Table 5 Case study

样例	Spoken words+acoustic+visual	Ground Truth	DBF	SCFR
1	"Except their eyes are kind of like this welcome to the polar express"+tense voice+frown expression	−0.6	−0.2	−0.5
2	"I think you will really love this movie if you are 8,"+emphatic voice+shocked expression	2.0	2.0	2.0
3	"and I think its predictable up an to point"+smooth voice+flat expression	−1.0	0.2	−0.7
4	"All I can say is he's a pretty good-looking guy"+disappointed voice+contradictory smile.	−1.2	−0.8	−1.0

结束语 对于多模态情感分析任务,现有的方法一方面忽略了不同步的模态融合表示在情感分析中的贡献不同,另一方面忽略了去噪之后音视频模态表示与文本的相关性。针对这一问题,本文提出了基于分步协作融合表示的情感分析方法。该模型首先利用瓶颈机制获得去除冗余信息后的音视频低级特征,提升模态间互补信息的有效集成能力。然后利用跨模态注意力机制得到文本强化后的音视频模态的高级特征;同时,多步的音视频模态特征表示通过门控机制与文本模态融合,在预训练模型中更新获得更多与文本相关的信息。最后利用三步模态融合特征实现情感分析。本文在两个公开数据集上进行了大量的实验,结果表明本文方法优于一系列基线。未来可以考虑视频模态提取与学习音频信息,更好地将辅助模态的内容应用到以文本为中心的多模态融合的研究任务中。此外,我们也应考虑在音频视频模态过滤噪声的同时最大化保留关键信息,同时从细粒度的方面进一步探究多模态情感分析[28]。

参 考 文 献

[1] HAZARIKA D, ZIMMERMANN, PORIA S, et al. Misa: Modality-invariant and specific representations for multimodal sentiment analysis[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1122-1131.

[2] NAGRANI A, YANG S, ARNAB A, et al. Attention bottlenecks for multimodal fusion[C]// Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2021: 14200-14213.

[3] WU S X, DAI D M, QIN Z W, et al. Denoising bottleneck with mutual information maximization for video multimodal fusion [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 756-767.

[4] ZHANG H Y, W Y, YIN G H, et al. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023: 2231-2243.

[5] YU T S, GAO H Y, YANG M, et al. Speech-text dialog pretraining for spoken dialog understanding with explicit cross-modal alignment[C]// Proceedings of the 61st Annual Meeting of

the Association for Computational Linguistics. ACL，2023：7900-7913.

[6] YANG H，LIN J Y，YANG A，et al. Prompt tuning for unified multimodal pretrained models[C]// Findings of the Association for Computational Linguistics. 2023：402-416.

[7] TSAI Y H，LIANG P P，ZADEH A，et al. Learning factorized multimodal representations[C]// International Conference on Representation Learning. 2018：53-69.

[8] HAN W，CHEN H，PORIA S，et al. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021：9180-9192

[9] GUO J W，TANG J J，DAI W C，et al. Dynamically adjust word representations using unaligned multimodal information[C]// Proceedings of the 30th ACM International Conference on Multimedia. 2022：3394-3402.

[10] SUN Y，MAI S J，HU H F，et al. Learning to learn better unimodal representations via adaptive multimodal meta-learning[J]. IEEE Transactions on Affective Computing，2023，14(3)：2209-2223.

[11] SUN L C，ZHENG L，LIU B，et al. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing，2024，15(1)：309-325.

[12] ZADEH A，CHEN M H，PORIA S，et al. Tensor fusion network for multimodal sentiment analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017：1103-1114.

[13] HUANG J H，LIU B，NIU M Y. Multimodal transformer fusion for continuous emotion recognition[C]// Proceedings of the IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). 2020：3507-3511.

[14] RAHMAN W，HASAN M K，LEE S，et al. Integrating multimodal information in large pretrained transformers[C]// Proceedings of the Conference Association for Computational Linguistics. 2020：2359-2373.

[15] LIANG T，LIN G S，FENG L，et al. Attention Is not enough：mitigating the distribution discrepancy in asynchronous multimodal sequence fusion[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：8148-8156.

[16] LUO H S，JI L，HUANG Y Y，et al. Scalevlad：Improving multimodal sentiment analysis via multiscale fusion of locally descriptors[J]. arXiv：2112.01368，2021.

[17] SUN J，HAN S K，RUAN Y P，et al. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023：658-670.

[18] SHI T，HUANG S L. MultiEMO：An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023：658-670.

[19] GILLES D，KANE J，DRUGMAN T，et al. COVAREP：A collaborative voice analysis repository for speech technologies[C]// Proceedings of the IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). 2014：978-986.

[20] AMOS L，LUDWICZUK B，SATYANARAYANAN M. OpenFace：A general-purpose face recognition library with mobile applications[J/OL]. https：//elijah. cs. cmu. edu/DOCS/CMU-CS-16-118. pdf.

[21] ZADEH A，ZELLERS R，PINCU S，et al. Multimodal sentiment-intensity analysis in videos：Facial gestures and verbal messages[J]. IEEE Intelligent Systems，2016，31(6)：82-88.

[22] ZADEH A，LIANG P P，PORIA S，et al. Multimodal language analysis in the wild：CMU-MOSEI dataset and interpretable dynamic fusion graph[C]// Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2018：2236-2246.

[23] LIU Z，SHEN Y，LIANG P P，et al. Efficient low-rank multimodal fusion with modality-specific factors[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018：2247-2256.

[24] TSAI Y H，BAI S，LIANG P P，et al. Multimodal transformer for unaligned multimodal language sequences[C]// Proceedings of the Conference Association for Computational Linguistics Meeting. 2019：6558-6571.

[25] LYU F M，CHEN X，HUANG Y Y，et al. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021：2554-2562.

[26] YU W M，XU H，YUAN Z Q，et al. Learning modality-specific representations withself supervised multi-task learning for multimodal sentiment analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021：10790-10797.

[27] HU G M，LIN T E，ZHAO Y，et al. UniMSE：Towards unified multimodal sentiment analysis and emotion recognition[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022：7837-7851.

[28] GUO R R，GAO J L，XU R J. Aspect-based Sentiment Analysis by Fusing Multi-feature Graph Convolutional Network[J]. Journal of Chinese Computer Systems，2024，45(5)：1039-1045.

GAO Long，born in 2000，postgraduate. His main research interests include natural language processing and so on.

LI Yang，born in 1988，Ph.D，associate professor，is a member of CCF(No. P6278M). Her main research interests include text sentiment analysis and text mining.

（责任编辑：柯颖）