智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于自适应图学习权重的多模态情感分析

曲海成, 徐波

引用本文:
曲海成, 徐波. 基于自适应图学习权重的多模态情感分析[J]. 智能系统学报, 2025, 20(2): 516-528.
QU Haicheng, XU Bo. Multimodal sentiment analysis based on adaptive graph learning weight[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(2): 516-528.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202401001

DOI: 10.11992/tis.202401001
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250107.1710.010

基于自适应图学习权重的多模态情感分析

曲海成, 徐波
(辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105)

摘　要: 在多模态情感分析任务中, 由于不同模态表现方式的不一致性, 模态间的情感信息密度具有较大的差异。为了平衡情感信息在不同模态中分布的不均匀性并减少多模态特征表示的冗余性, 提出了一种基于自适应图学习权重的多模态情感分析方法。首先, 采用不同的特征提取方法捕获单一模态内的特定信息; 其次, 将不同模态通过公共编码器映射到同一空间中, 利用跨模态注意力机制来显式构建模态间的关联; 然后, 将每种模态对任务分类的预测值以及模态表示嵌入到自适应图中, 通过模态标签学习不同模态对最终分类任务的贡献度来动态调整不同模态之间的权重, 以适应主导模态的变化; 最后, 引入信息瓶颈机制进行去噪, 旨在学习一种无冗余的多模态特征表示进行情感预测。在公开的多模态情感分析数据集上对所提出的模型进行了评估。实验结果表明, 其有效提升了多模态情感分析的准确性。

关键词: 多模态; 情感分析; 模态差异性; 信息冗余; 自适应图学习; 跨模态注意力; 相似性约束; 信息瓶颈
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2025)02−0516−13

中文引用格式: 曲海成, 徐波. 基于自适应图学习权重的多模态情感分析 [J]. 智能系统学报, 2025, 20(2): 516–528.
英文引用格式: QU Haicheng, XU Bo. Multimodal sentiment analysis based on adaptive graph learning weight[J]. CAAI transactions on intelligent systems, 2025, 20(2): 516–528.

Multimodal sentiment analysis based on adaptive graph learning weight

QU Haicheng, XU Bo
(School of Software, Liaoning Technical University, Huludao 125105, China)

Abstract: The inconsistency in representing different modalities in multimodal sentiment analysis tasks results in significant differences in the density of emotional information between modalities. A multimodal sentiment analysis method based on adaptive graph learning weights is proposed to balance the uneven distribution of emotional information in different modalities and reduce the redundancy of multimodal feature representations. First, different feature extraction methods are used to capture specific information within each mode. Second, different modalities are mapped to the same space through a common encoder, and cross-modal attention mechanisms are used to explicitly construct correlations between modalities. Third, the predicted values and modal representations of each modality for task classification are embedded into the adaptive graph, and the contribution of different modalities to the final classification task is learned through modal labels to dynamically adjust the weights between different modalities for adapting to changes in the dominant modality. Finally, an information bottleneck mechanism is introduced for denoising, aiming to learn a nonredundant multimodal feature representation for sentiment prediction. The proposed model is evaluated on the publicly available multimodal sentiment analysis datasets. Experimental results show that its effectively improving the accuracy of multimodal sentiment analysis.

Keywords: multimodal; sentiment analysis; modal differences; information redundancy; adaptive graph learning; cross modal attention; similarity constraints; information bottleneck

收稿日期: 2024−01−02.　**网络出版日期**: 2025−01−08.
基金项目: 辽宁省高等学校基本科研项目 (LIKMZ20220699).
通信作者: 曲海成. E-mail: quhaicheng@lntu.edu.cn.

随着在线视频的普及和视频平台的日益增多, 多模态情感分析已经成为一个重要的研究领

域[1]。多模态情感分析旨在从视频片段中感知人类的情绪态度, 视频内容涉及来自各种模态的时间序列数据, 例如文本、视觉和音频, 将来自不同模态的情感信息结合起来, 能够提高情感分类的准确性和全面性[2]。

在多模态情感分析中, 同一视频片段中不同模态都包含丰富的情感信息, 且这些信息往往是互补的, 为语义和情感消岐提供了额外的线索。多模态情感分析的核心部分是多模态表征学习与融合, 通过对多模态表征进行编码与整合, 以理解原始数据背后的情感。先前的工作取得了一些成就[3-4], 但在多模态情感分析研究中仍然有 2 个具有挑战性的问题尚未解决。第 1 个问题是如何使得不同模态进行有效融合, 以减少模态间的异质性; 第 2 个问题是如何学习一种无冗余、无噪声信息且只与分类任务相关的多模态情感特征表示。

对于第 1 个问题, 现有的多模态情感分析方法大多通过设计复杂的融合机制对不同模态进行融合, 以促进模态内和模态间的交互[5-6], 但是模态间固有的异质性使得模态间不能进行很好融合。一种有效的缓解模态异质性的方式是将可靠的、可推广的情感信息从强势模态融合到弱势模态。由于文本模态在 3 种中占有更优的地位, 许多学者提出以文本模态为中心来增强非文本模态的语义特性[7-9]。然而, 模态间显著的特征分布不匹配使得直接的静态式跨模态融合并不是最优的方案。这是因为不同时刻主导情感判定的模态通常呈现动态变化, 使得模态间的交互方向具有不确定性。例如, 在某一时刻, 许多情绪更容易通过文本识别, 而有些情绪更容易通过视觉来识别。因此, 模型本身应该根据不同时刻模态间的交互权重动态调整交互方向; 另外一种缓解模态异质性的方式是学习模态间一种潜在相似性表示来减少模态间的差异性[10]。学习模态间潜在相似性需要从有关的各种模态中提取共同特征, 通常情况下, 不同的模态在数据结构方面是不同的, 但是由于它们具有相似的情感倾向, 所以可以共享一些共同的语义信息。

第 2 个问题目前受到关注较少, 这是因为许多先前的工作只关注于如何设计复杂的网络融合模型来学习具有单模态特征的多模态联合表示, 却忽视了产生的多模态嵌入可能是冗余的[11]。例如, Zadeh 等[12] 使用外积生成高阶多模态张量, 该张量产生冗余表示, 可能在预测过程中包含一些与分类任务无关的信息。Tsai 等[13] 提出利用跨模态编码器将不同模态进行两两融合, 最后将 6 种融合向量进行拼接得到最终的预测特征, 虽然这种拼接操作可以保留每一种融合向量的信息, 然而却并没有考虑视觉和音频模态中带有大量的与情感分类任务无关的冗余信息。最近的一些工作[14-15] 提出利用门控机制来去除模态融合过程中产生的冗余信息, 却依然没有考虑对多模态特征表示进行噪声消除。此外, 门控机制的不可解释性以及对参数的敏感性等问题使得模型在预测过程中难以做出正确的决策。因此, 学习一种多模态特征表示足以进行正确推理, 同时又没有冗余信息和噪声是多模态情感分析中的一项重大挑战。

针对上述 2 个问题, 本文提出了一种基于自适应图学习权重的多模态情感分析方法 (multimodal sentiment analysis method based on adaptive graph learning weight), 旨在平衡不同模态间情感信息差异性并减少多模态特征表示的冗余性。该方法首先将不同模态数据输入到单模态特征提取模块, 针对不同的模态采用不同的特征提取方法, 来捕获模态内的特定信息。将单模态特征通过公共编码器映射到同一空间中, 利用跨模态注意力机制对不同模态进行显式构建以捕获模态间的关联性。然后将各个模态特征输入到自适应图学习模块中, 将每种模态对任务分类的预测值以及模态表示嵌入到自适应图中, 通过模态标签学习不同模态对最终分类任务的贡献度来动态调整不同模态之间的权重, 以适应主导模态的变化。同时, 为了减少模态间的差异性, 在语义和时间 2 个维度来捕获模态间潜在相似性的不变表示。之后, 利用信息瓶颈机制对多模态融合特征进行去噪, 去除一些与任务分类无关的冗余信息, 使得多模态特征表示尽可能与情感分类任务相关。最后得到情感预测结果。

综上所述, 本文的主要工作总结如下:

1) 提出了一种基于自适应图学习权重的多模态情感分析方法来减少模态间的异质性, 使得情感信息均匀地分布在不同的模态中, 以增强情感分类的性能。

2) 提出了在语义和时间 2 个维度进行相似性约束来捕获模态间的潜在相似性不变表示, 以减少模态间的差异性。

3) 首次考虑多模态特征表示具有冗余性, 并利用信息瓶颈机制来学习一种与任务分类相关并且无冗余信息的多模态特征表示。

4) 在多模态情感分析基准数据集进行实验,

实验结果表明，本文提出的方法能有效提高多模态任务准确率。

1　相关工作

多模态情感分析旨在从嵌入在视频片段中的文本、视觉和音频信息中挖掘情感信息，探索模态间的内在相关性，减少单一模态情感信息偏差[16-18]。多模态情感分析的一个关键问题是使得不同模态之间有效地融合，以减小模态间的差异性。Zadeh 等[12] 提出一种张量融合网络模型，采用端到端的方式学习模态内和模态间的联系。Majumder 等[19] 提出了一种分层特征融合方法，首先融合 3 种模态中的 2 种，然后融合所有模态。Xu 等[20] 使用长短期记忆和视觉特征引导来融合多模态信息。Akhtar 等[21] 采用 3 个双向门控循环单元网络捕获每种模态的上下文信息，并通过多任务学习共同进行情绪和情绪分析。Hazarika 等[22] 提出了模态共性和特性的表示方法 (modality-invariant and-specific representations for multimodal sentiment analysis, MISA)，通过利用损失函数学习模态间的共性和特性并进行融合。Yu 等[23] 设计了一种自监督学习策略的标签生成模块，并以多任务学习的方式分别学习模态间的一致性表示和差异性表示。上述多模态情感分析方法旨在通过设计复杂的融合策略来实现不同模态之间的交互，却忽视了不同时刻主导情感判定的模态通常呈现动态变化，使得模态间的交互方向具有不确定性。模态间显著的特征分布不匹配使得这种直接的跨模态融合并不是最优的，一种更优的方式是模型本身应该可以根据不同例子动态调整交互方向，以促进不同模态更好地融合。

多模态情感分析的另一个关键问题是学习模态间不变表示来减少模态间的差异性，大多数方法是通过使用相似性约束来实现的。Yu 等[23] 在数值水平上约束模态级特征，从语言、声学和视觉模式中获得更多正相关的表征。Hazarika 等[22] 和 Zuo 等[24] 通过使用 (central moment discrepancy, CMD) 来约束不同模式之间的相似性，旨在最小化模态间的差异性。Liu 等[25] 提出了离散共享空间来捕获细粒度表示，以提高跨模态检索的准确性。以上研究表明，模态间的不变表示可以有效地弥补不同模态之间的差异。然而，上述方法却忽略了不同模态在语义层面以及时间层面含有不同的特征。在语义层面，来自相同情绪但不同模态的语义表征应该比来自相同模态但不同情绪的语义表征更相似。在时间层面，每一个时间信号都带有不同的情感信息，如果在时间序列上捕获模态间的相似性，则可以获得更加细粒度的特征表示。

近年来，基于注意力机制的模型框架在处理长序列表示方面表现出较好的性能，目前已成为大多数神经网络结构的重要组成部分[26-29]。Tsai 等[13] 提出一种跨模态编码器，该结构利用交叉模态注意力将文本、音频、图像 3 种模态进行交叉融合，使单个模态可以从其他的模态中获取信息。Lyu 等[30] 在此基础上提出了渐进式模态强化方法，通过设计一种公共信息池来交换不同模态之间的信息。Rahman 等[31] 利用 (bidirectional encoder representations from Transformers, BERT) 提出一种多模态自适应门，允许在微调的过程中接受多模态非语言数据，通过结合文本和音频模态来动态调整单词的权重。上述的多模态情感分析方法通过不同融合策略进行模态间的交互，却忽略了多模态特征联合表示可能存在信息冗余，即在多模态融合过程中，生成的多模态特征表示可能存在一些与分类任务无关的信息。为了解决这一问题，文献 [32-33] 等利用门控机制去除冗余信息。与其不同的是，本文使用信息瓶颈机制[34] 对多模态联合表示进行去噪。信息瓶颈基于互信息，旨在最大化编码表示与标签之间的互信息，同时最小化编码表示与输入之间的互信息。通过应用信息瓶颈原理，模型可以学习过滤掉可能干扰预测的噪声和冗余信息，从而获得最小预测充分表示。

综上所述，随着深度学习研究的不断深入，多模态情感分析取得了显著的进步和发展，但如何平衡异质模态之间信息差异性以及如何消除多模态联合表示中的冗余信息仍然具有挑战性。为此，本文提出了一种基于自适应图学习权重的多模态情感分析方法实现多模态特征的有效融合，同时减少与分类任务无关的冗余信息，提高多模态情感分类的准确率。

2　提出的方法

为了平衡情感信息在不同模态中分布的不均匀性并减少多模态特征表示的冗余性，提出了一种基于自适应图学习权重的多模态情感分析模型，模型结构如图 1 所示，其具体包含 4 个部分。1) 单模态特征提取：针对文本、视觉和音频 3 种模态采用不同的特征提取方法进行特征提取，捕获模态内的特定信息。2) 自适应图学习：将每种

模态对任务分类的预测值以及模态表示嵌入到自适应图中，通过模态标签学习不同模态对最终分类任务的贡献度来动态调整不同模态之间的权重，以适应主导模态的变化；在语义和时间 2 个维度来捕获模态间潜在相似性的不变表示，以减小模态间的差异性。3) 信息瓶颈机制：对多模态特征表示进行特征重选择，以学习一种与分类任务相关并无冗余信息的多模态融合表示。4) 情感分类：基于多模态融合表征，利用情感分类器，得到最终的情感预测结果。

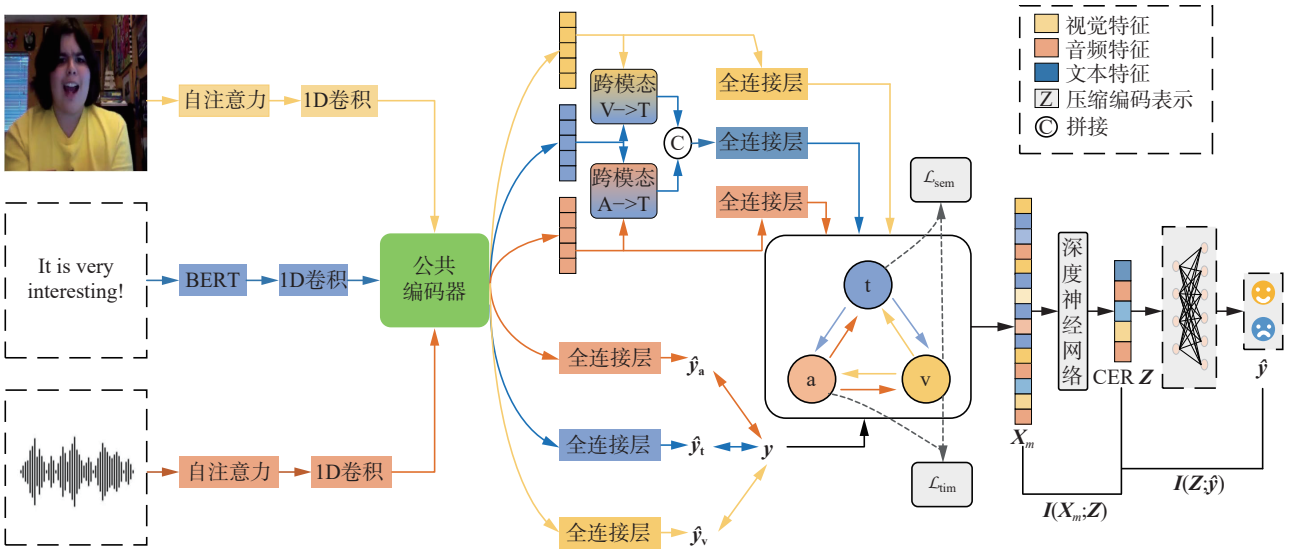


图 1　基于自适应图学习权重的多模态情感分析模型

Fig. 1　Multimodal sentiment analysis model based on adaptive graph learning weights

2.1　问题的描述

多模态情感分析任务主要涉及文本 (t)、视觉 (v) 和音频 (a) 3 种模态。将从相同视频片段中提取的一组模态 (文本、视觉、音频) 表示为 $M=\{\mathrm{t},\mathrm{v},\mathrm{a}\}$，对于本文的其余部分，$m\in M$ 表示一个特定的模态。每一个模态输入的特征向量表示为 $\boldsymbol{I}_{\mathrm{t}}\in\mathbf{R}^{l_{\mathrm{t}}\times d_{\mathrm{t}}}$，$\boldsymbol{I}_{\mathrm{v}}\in\mathbf{R}^{l_{\mathrm{v}}\times d_{\mathrm{v}}}$，$\boldsymbol{I}_{\mathrm{a}}\in\mathbf{R}^{l_{\mathrm{a}}\times d_{\mathrm{a}}}$，其中 l_m 和 d_m 分别表示模态 m 的序列长度和特征维数。多模态情感分析的任务是从每个模态序列中提取任务相关信息并通过融合策略形成统一的表示进行情感预测，模型最终输出是一个情感强度结果 $\hat{\boldsymbol{y}}$。

2.2　单模态特征提取

为了更好地提取单模态数据特征，本文针对不同的模态采用了不同的特征提取方法来获取每个模态的特定信息。对于视觉和音频模态，采用自注意力机制网络来捕获全局信息，获得远距离依赖特征。对于文本模态，采用预训练语言双向编码器 BERT 来提取句子深层语义特征，3 种模态的初始特征信息 $\boldsymbol{X}_m$ 可表示为

$$\boldsymbol{X}_m=\begin{cases}\mathrm{BERT}(\mathrm{I}_m;\theta_m^{\mathrm{BERT}})\in\mathbf{R}^{l_m\times d_m}, & m=\mathrm{t}\\ \mathrm{Attention}(\mathrm{I}_m;\theta_m^{\mathrm{Attention}})\in\mathbf{R}^{l_m\times d_m}, & m\in\{\mathrm{v},\mathrm{a}\}\end{cases}$$

式中：$\boldsymbol{I}_m$ 表示输入模态的特征向量，l_m 和 d_m 分别表示模态 m 的序列长度和特征维度，θ_m^M 是模态 m 的特征提取网络模型 $M\in\{\mathrm{BERT},\mathrm{Attention}\}$ 的参数。为了使每个模态的输入序列元素可以获得相邻元素信息，以便更全面地理解整个序列的上下文信息，将 3 种模态输入到一维时间卷积层中，具体表示为

$$\boldsymbol{h}_m=\mathrm{Conv1D}(\boldsymbol{X}_m,k_m)\in\mathbf{R}^{l_m\times d_m},m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$$

式中：$\mathrm{Conv1D}(\cdot)$ 表示一维时间卷积层，k_m 表示卷积核的大小，$\boldsymbol{h}_m$ 是单模态特征提取的输出。

2.3　自适应图学习

为了缓解由于表示方式差异以及其他因素导致的模态间不一致性，将 3 种模态通过公共编码器映射到同一空间中，具体表示为

$$\hat{\boldsymbol{h}}_m=\mathrm{Encoder}(\boldsymbol{h}_m;\theta_{\mathrm{Encoder}})\in\mathbf{R}^{l_m\times d_m},m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$$

式中：$\mathrm{Encoder}(\cdot;\cdot)$ 表示公共编码器，$\theta_{\mathrm{Encoder}}$ 表示公共编码器的参数，$\hat{\boldsymbol{h}}_m$ 表示映射到同一空间中的模态特征表示。为了更好进行模态之间的交互，促进多模态信息的融合，利用跨模态注意力机制来显式构建模态间的关联，具体来说，2 种不同的模态 m_1 和 m_2 之间的跨模态注意力可表示为

$$\mathrm{CoAn}_{m_1\to m_2}(\boldsymbol{X}_{m_1},\boldsymbol{X}_{m_2})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_{m_2}\boldsymbol{K}_{m_1}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_{m_1}$$

式中：$\mathrm{CoAn}_{m_1\to m_2}(\cdot,\cdot)$ 表示 2 种模态间地跨模态注意力，$\boldsymbol{X}_{m1}$ 与 $\boldsymbol{X}_{m1}$ 是 CoAn 的输入；$\mathrm{softmax}(\cdot)$ 是一个归一化指数函数；$\boldsymbol{Q}_{m_2}=L(\boldsymbol{X}_{m_2})\boldsymbol{W}_Q$，$\boldsymbol{K}_{m_1}=L(\boldsymbol{X}_{m_1})\boldsymbol{W}_K$，$\boldsymbol{V}_{m_1}=L(\boldsymbol{X}_{m_1})\boldsymbol{W}_V$ 分别表示多头注意力机制的查

询、键和值，$L(\cdot)$ 表示层归一化操作。通常情况下，文本模态含有高阶的情感语义特征，视觉和音频模态含有低阶的情感语义特征，因此，将文本作为主要模态，并应用跨模态注意力机制得到文本与视觉和文本与音频模态的联合融合结果，具体可表示为

$$\boldsymbol{f}_t=\mathrm{ConV}\left(\left[\mathrm{CoAn}_{\hat{\boldsymbol{h}}_v\to\hat{\boldsymbol{h}}_t}\left(\hat{\boldsymbol{h}}_t,\hat{\boldsymbol{h}}_v\right);\mathrm{CoAn}_{\hat{\boldsymbol{h}}_a\to\hat{\boldsymbol{h}}_t}\left(\hat{\boldsymbol{h}}_t,\hat{\boldsymbol{h}}_a\right)\right]\right)\in\mathbf{R}^{l_m\times d_m}$$

式中：$\boldsymbol{f}_t$ 表示通过跨模态注意力机制后得到的文本模态的特征向量，$\mathrm{ConV}(\cdot)$ 表示卷积操作，$[\cdot;\cdot]$ 表示拼接操作。使用全连接层和 ReLU 函数将 3 种模态进行映射变化得到模态表示，具体可表示为

$$\boldsymbol{F}_m=\begin{cases}g_m(\boldsymbol{f}_m;\theta_{g_m})\in\mathbf{R}^{l_m\times d_m},m=\mathrm{t}\\ g_m(\hat{\boldsymbol{h}}_m;\theta_{g_m})\in\mathbf{R}^{l_m\times d_m},m\in\{\mathrm{v},\mathrm{a}\}\end{cases}$$

式中：$g_m(\cdot)$ 表示全连接神经网络，θ_{g_m} 表示模态 m 所对应 $g_m(\cdot)$ 的网络参数。

在多模态情感分析任务中，由于模态间固有的异质性，不同时刻主导情感判定的模态通常呈现动态变化，导致模态间的交互方式具有不确定性，为了适应不同时刻主导模态的变化，构建一种图学习单元以自适应的方式动态调整不同模态之间的权重。具体来说，$V_m=\{m|\{\mathrm{t},\mathrm{v},\mathrm{a}\}\}$表示每一种模态作为图中的顶点，$w_{m_1\to m_2}$ 表示模态 m_1 与模态 m_2 之间的交互权重。为了能够动态地调整模态之间的交互权重，将每种模态对任务分类的预测值以及模态表示嵌入到自适应图中，具体可表示为

$$\hat{\boldsymbol{y}}_m=P_m(\boldsymbol{h}_m,\theta_{p_m}),m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$$

$$\boldsymbol{e}_m=[\hat{\boldsymbol{y}}_m;\boldsymbol{F}_m],m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$$

式中：$\hat{\boldsymbol{y}}_m$表示每一种模态的预测值，$P_m(\cdot,\cdot)$表示带有线性层和激活函数 ReLU 的预测网络，θ_{P_m}则表示每一种模态预测网络的参数，$\boldsymbol{e}_m$表示每一种模态的自适应表示，则模态 m_1 与模态 m_2 间的交互权重则可以表示为

$$w_{m_1\to m_2}=H([\boldsymbol{e}_{m_1};\boldsymbol{e}_{m_2}],\theta_H)\tag{1}$$

式中：$H(\cdot)$ 表示带有可学习参数为 θ_H 的全连接层。整个图学习单元的边权重 $W_{ij}=w_{m_i\to m_j}(i\neq j)$ 可以通过式 (1) 来构建和学习。具体来说，通过将每种模态预测值以及模态表示嵌入到自适应图中，使得自适应图中的边权既可以保留每个模态的决策信息又可以保留每个模态的特征信息，通过整合不同信息层次，促使在学习过程中同时考虑底层和高层特征，提高对多模态数据的表达能力。然后，通过动态调整边的权重，在训练过程中适应不同模态之间的关系，实现模态之间的自适应学习。通过这种方式，模态间可以以一种更加灵活的方式进行信息交互，来适应主导模态的变化。

为了表示模态间的相似性，使用不同模态对不同类别的预测值之间的差值来衡量个模态间的差异性，具体可表示为

$$\varepsilon_{m_1\to m_2}=\left\|\hat{y}_{m_1}-\hat{y}_{m_2}\right\|_1$$

式中 $\|\cdot\|_1$ 表示 L_1 损失函数。具体来说，若 $\varepsilon_{m_1\to m_2}$ 的值越小，则表示模态 m_1 与模态 m_2 差异性越小；若 $\varepsilon_{m_1\to m_2}$ 的值越大，则表示模态 m_1 与模态 m_2 差异性越大。因此，整个图学习单元的损失可表示为

$$\mathcal{L}_{\mathrm{glu}}=\sum_{j=1}^{3}\sum_{i=1,i\neq j}^{3}w_{m_i\to m_j}\times\varepsilon_{m_i\to m_j}\tag{2}$$

2.4 信息瓶颈机制

经过自适应图学习阶段，不同模态间以一种更加灵活的方式进行情感信息的交互与转移，使得模态间的异质性得到了有效的缓解。考虑到在融合多个不同模态的情感信息时，一些信息在不同的模态中可能被重复编码，导致生成的多模态嵌入可能包含冗余的信息，对于最终的情感分类任务并不能提供额外的有益信息。为了使得最终的多模态嵌入表示尽可能与情感分类任务相关，引入信息瓶颈机制来学习一种强大且无冗余的多模态特征表示，提高多模态情感分类的准确率。具体来说，根据信息瓶颈理论，通过最大化表示和目标之间的互信息，同时约束表示和输入数据之间的互信息，来学习给定任务的最小充分表示。由文献 [35] 可知，目标函数可表示为

$$\mathcal{R}_{\mathrm{IB}}=I(\boldsymbol{Z};\boldsymbol{Y})-\beta I(\boldsymbol{Z};\boldsymbol{X})\tag{3}$$

式中：$\boldsymbol{X}=[\boldsymbol{F}_{\mathrm{t}};\boldsymbol{F}_{\mathrm{v}};\boldsymbol{F}_{\mathrm{a}}]$，这里 $\boldsymbol{F}_m,m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$为经过图学习之后的模态表示。通过拼接操作，$\boldsymbol{X}$可以保留来自 3 个不同模态的所有信息，用于学习最小充分的多模态压缩编码表示 $\boldsymbol{Z}$，$I(\boldsymbol{Z};\boldsymbol{Y})$表示压缩编码 $\boldsymbol{Z}$ 与真实标签之间的互信息，$I(\boldsymbol{Z};\boldsymbol{X})$表示压缩编码 $\boldsymbol{Z}$ 与输入之间的互信息，β是一个在优化过程中决定最小信息约束权重的标量。具体来说，目标函数 $\mathcal{R}_{\mathrm{IB}}$ 中有 2 个约束条件，第 1 个约束鼓励 $\boldsymbol{Z}$ 最大限度地预测目标 $\boldsymbol{Y}$，第 2 个约束鼓励 $\boldsymbol{Z}$ 尽可能“忘记”$\boldsymbol{X}$，本质上来说，它迫使 $\boldsymbol{Z}$ 表现得像 $\boldsymbol{X}$ 的最小充分表示来预测 $\boldsymbol{Y}$。通过这种方式，可以减少生成的多模态特征表示中的冗余信息，提供了更具有表达力的情感特征。

通过结合第一约束条件互信息下界和第二约

束条件互信息上界, 式 (3) 可表示为

$$\begin{aligned}&\mathcal{R}_{\mathrm{IB}}=I(\boldsymbol{Z};\boldsymbol{Y})-\beta I(\boldsymbol{Z};\boldsymbol{X})\geqslant\\&\int \mathrm{d}x\mathrm{d}y\mathrm{d}z p(x)p(y|x)p(z|x)\log q(y|z)-\\&\beta\int \mathrm{d}x\mathrm{d}z p(x)p(z|x)\log\left(\frac{p(z|x)}{q(z)}\right)=\mathrm{E}(x,y)\sim p(x,y),z\sim p(z|x)\\&[\log q(y|z)-\beta\cdot\mathrm{KL}(p(z|x)\|q(z))]=J_{\mathrm{IB}}\end{aligned}\tag{4}$$

式中: $q(y|z)$与$q(z)$分别为$p(y|z)$和$p(z)$的变分近似, $\mathrm{KL}(\cdot)$表示 2 个随机变量的 KL 散度, 通过最大化J_{IB}的下界, 从而对目标函数$\mathcal{R}_{\mathrm{IB}}$进行优化。为了优化$J_{\mathrm{IB}}$的下界, 将$p(y|z)$与$p(z)$视为高斯分布通过神经网络来学习期望$\mu_{\boldsymbol{Z}}$和方差$\Sigma_{\boldsymbol{Z}}$, 由于参数梯度的计算具有随机性, 使用重参数化技巧来获得$\boldsymbol{Z}$, 最后, 通过蒙特卡罗近似采样, 式 (4) 可表示为

$$\begin{aligned}&J_{\mathrm{IB}}=\\&\frac{1}{n}\sum_{i=1}^{n}\left[E_{\kappa\sim p(\kappa)}[\log q(y_i|z_i)]-\beta\cdot\mathrm{KL}\left(\mathcal{N}(\mu_{\boldsymbol{Z}_i},\sum \boldsymbol{z}_i)\|\mathcal{N}(0,1)\right)\right]\end{aligned}$$

式中: $\kappa\sim\mathcal{N}(0,1)$, n表示抽样大小, i为每个样本的下标。通过最大化J_{IB}, 使得$\boldsymbol{Z}$对分类任务最具判别性的同时, 遗忘多模态联合表示$\boldsymbol{X}$的冗余信息。本文提出的算法如算法 1 所示。

算法 1　自适应图学习算法

1) 获取单模态特征表示$\boldsymbol{X}_m, m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$;

2) 通过公共编码器获得$\hat{\boldsymbol{h}}_m, m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$;

3) 计算得到跨模态注意力 CoAn V−>T 和 CoAn A−>T;

4) 通过全连接层得到模态表示$\boldsymbol{F}_m, m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$;

5) 计算得到模态自适应表示$\boldsymbol{e}_m, m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$;

6) 通过式 (1) 获得模态间的交互权重$W_{ij}=w_{m_i\to m_j}$;

7) 通过图学习得到模态联合表示;

8) 通过信息瓶颈机制去冗余, 得到多模态联合特征;

9) 得到预测结果$\hat{\boldsymbol{y}}$;

10) 计算损失。

2.5　损失函数

模态间的不变表示可以有效地弥补不同模态之间的差异, 因此在自适应图学习中, 为了获得模态间潜在相似性的不变表示, 分别从语义层面和时间层面来对模态间的相互作用进行约束, 旨在捕获更加细粒度的特征, 减小模态间的差异。此外, 本文将每种模态对任务分类的预测值以及模态表示嵌入到自适应图中, 因此整个图学习单元的损失可表示为式 (2), 多模态特征表示所产生的情感预测值与真实值之间的差异表示为任务损失, 则整个模型的损失函数具体可表示为

$$\mathcal{L}_{\mathrm{total}}=\mathcal{L}_{\mathrm{task}}+\mathcal{L}_{\mathrm{glu}}+\mathcal{L}_{\mathrm{sem}}+\mathcal{L}_{\mathrm{tim}}$$

2.5.1　语义相似性约束

在模态间进行自适应权重融合时, 引入语义方面的约束, 以减小模态在交互时特征表示的差异。具体来说, 来自相同情绪但不同模态的语义表征比来自相同模态但不同情绪的语义表征更相似。基于此, 利用余弦相似度来衡量 2 种模态之间的相似性, 每种模态都具有一个特征表示, 在共享空间中, 如果 2 种不同模态对应的情感相同, 那么它们的特征表示在这个空间中应该更加接近, 表现为更相似。而如果是相同模态但不同情感, 那么它们的特征表示在这个空间中应该有一定的差异, 表现为一定的距离, 基于此, 将语义层面损失定义为

$$\begin{aligned}&\mathcal{L}_{\mathrm{sem}}=\\&\frac{1}{|V|}\sum_{(i,j,k)\in V}\max(0,\alpha-\cos(\boldsymbol{F}_{m[i]},\boldsymbol{F}_{m[j]})+\cos(\boldsymbol{F}_{m[i]},\boldsymbol{F}_{m[k]}))\end{aligned}\tag{5}$$

式中: $V=\{(i,j,k)|m[i]\neq m[j],m[i]=m[k],c[i]=c[j],c[i]\neq c[k]\}$, $m[i]$表示模态i, $c[i]$表示模态i所对应的情感类别; $\cos(\cdot,\cdot)$表示 2 个特征向量之间的余弦相似度; α是一个距离参数, 用来调节同一情感但不同模态的特征表示与同一模态但不同情感的特征表示之间的距离。

2.5.2　时间相似性约束

在时间序列数据中, 不同时间点上的模态输入可能呈现出情感变化的趋势。通过在时间方面进行约束, 模型可以随着时间演变更好地捕获情感信息。为了不同模态在同一时间点上的特征表示有一致性, 引入多元高斯分布的 KL 散度来捕获更细粒度的相似性约束。具体来说, 将每个模态的时间戳特征视为高斯分布, 因此, 每个具有时间序列的模态都可以视为多元高斯分布, 利用多元分布的 KL 散度来约束时间分布水平上的相似性, 不仅可以利用顺序信息, 还可以防止在此过程中丢失特定信息, 在时间级别上约束模型可以使其学习到更细粒度的分布相似性。

对于每一个模态, 利用线性投影在时间层面上施加相似性约束, 模态中的每一个时间点, 将其视为一个独立的期望为μ_m、方差σ_m为的高斯分布$\mathcal{N}(\mu_m,\sigma_m^2)$, 因此, 所有模态序列是一个多元高斯分布$\mathcal{N}(\mu_m,\Sigma_m^2)$则对于 2 个不同模态$m_1$和$m_2$的分布形态分别为$P_{m_1}(x)\sim\mathcal{N}\left(\mu_{m_1},\Sigma_{m_1}^2\right)$和$P_{m_2}(x)\sim\mathcal{N}\left(\mu_{m_2},\Sigma_{m_2}^2\right)$。对多元高斯分布应用 KL 散度来计算相似度, 可表示为

$$D_{\mathrm{KL}}(P_{m_1}(x)\|P_{m_2}(x))=\int\log(P_{m_1}(x))P_{m_1}(x)\mathrm{d}x-\int\log(P_{m_2}(x))P_{m_1}(x)\mathrm{d}x \tag{6}$$

将高斯公式一般形式带入式 (6), 最后经过化简得

$$D_{\mathrm{KL}}(P_{m_1}(x)\|P_{m_2}(x))=\frac{1}{2}\log\left(\frac{|\Sigma_{m_2}|}{|\Sigma_{m_1}|}\right)+\frac{1}{2}E_{p_{m_1}(x)}(\Gamma_2-\Gamma_1)$$

式中: $|\Sigma_{m_1}|$和$|\Sigma_{m_2}|$分别表示分布$P_{m_1}(x)$和$P_{m_2}(x)$的方差, $E_{p_{m_1}(x)}$表示分布$P_{m_1}(x)$的期望值, Γ_1和Γ_2分别表示分布$P_{m_1}(x)$和$P_{m_2}(x)$的指数部分。模态间在时间序列上的损失最终可表示为

$$\mathcal{L}_{\mathrm{tim}}=D_{\mathrm{KL}}(\boldsymbol{F}_{m_1}(x)\|\boldsymbol{F}_{m_2}(x))$$

2.5.3 任务损失

在模型训练期间, 使用均方误差来计算模型预测值$\hat{\boldsymbol{y}}$和真实值$\boldsymbol{y}$之间的差异, 具体来说, 在一个批次大小为N_{b}的训练过程中, 损失可表示为

$$\mathcal{L}_{\mathrm{task}}=\frac{1}{N_{\mathrm{b}}}\sum_{i=0}^{N_{\mathrm{b}}}\|\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i\|_2^2$$

3 实验及结果分析

本文提出模型的实现基于 PyTorch 框架, 利用 NVIDIA 3050 GPU 进行训练。实验训练时批量大小为 16, 训练轮数为 50, 基础学习率为 10^{-5}, 采用 Adam 作为优化器。

3.1 数据集

本文在 2 个可公开访问的数据集 CMU_MOSI[36] 和 CMU_MOSEI[37] 上进行了实验, 以测试本文模型对多模态情感分析的有效性, 本文使用的数据集分割方案遵循原始数据集的方案, 具体见表 1。

表 1　CMU_MOSI 和 CMU_MOSEI 数据集统计

Table 1　CMU_MOSI and CMU_MOSEI datasets statistics

数据集	训练集	验证集	测试集	总计
CMU_MOSI	1 281	229	685	2 195
CMU_MOSEI	16 326	1 871	4 659	22 856

CMU_MOSI　CMU-MOSI 数据集包含了人们对各种话题的短评, 数据集中共包含 2 195 个标注的样本, 每个样本都包括 3 种模态信息: 文本、视觉和音频, 情感标注范围为−3 ~ 3, 表示从强烈负面到强烈正面的情感强度。

CMU_MOSEI　CMU_MOSEI 是一个比 CMU_MOSI 规模更大、情感标注更为丰富和细致的情感分类数据集, 包含来自 1 000 个不同说话者的 22 856 个带注释的视频片段。

3.2 基准模型

为了充分验证本文模型的性能, 本文选取了部分常用的多模态情感分类方法作为对比。

1)TFN[12]: 基于张量融合的多模态网络, 采用三倍笛卡尔积以端到端的方式学习模态内和模态间的动态信息。

2)LFM[38]: 基于张量融合的低秩多模态网络, 利用低秩权值张量来降低张量融合的复杂度, 并获取模态内和模态间的信息。

3)MULT[13]: 该方法利用交叉模态注意机制获取多模态序列之间的远程交互, 并使单个模态可以从其余模态中获取辅助信息。

4)PMR[30]: 基于渐进式网络融合, 利用公共消息池来实现 3 种模态之间的交互, 获得具有更深层次的情感特征。

5)MISA[22]: 该方法利用损失函数学习各个单模态表示之间的共性特征和特性特征, 利用 Transformer 将学习到的特征进行融合。

6)SELF_MM[23]: 基于自监督学习策略获得的单模态监督, 采用联合训练多模态任务和单模态任务的方式来学习模态的一致性和差异性。

7)MMIM[39]: 该模型通过分层最大化互信息 (mutual information, MI) 来维护关键任务的相关信息。

8)MAG-BERT[31]: 提出了一个基于 BERT 和 XLNET 微调模型的多模态自适应门, 允许在微调的过程中接受多模态非语言数据。

9)TETFN[40]: 以文本为中心的跨模态注意力融合网络, 它是一种利用跨模态注意力对未对齐的多模态时间信息建模的多模态融合网络。

10)DMD[41]: 针对多模态间异构的本质属性, 根据同质和异质使用了共享编码器和私有编码器, 利用图蒸馏单元编码器来融合同质和异质特征。

3.3 结果与分析

为充分验证本文模型的有效性, 利用上述模型进行对比实验, 在 2 个数据集上的实验结果见表 2, 其中, Acc7 为七分类正确率, Acc2 为二分类正确率, F_1 为精确度和召回率的调和平均数, MAE 为平均绝对误差, Corr 为相关系数。从表 2 数据中可以看出: 本文模型 2 个数据集中的结果均优于对比模型。相比于经典模型 TFN 与 LMF, 本文模型在数据集 CMU_MOSEI 上 Acc7、Acc2、$F_1$3 个指标分别提升了 4.0、6.3、6.0 百分点和 3.8、5.1、4.7 百分点, 原因在于 TFN 与 LMF 无法捕捉各个模态数据内部存在的上下文语义信息, 不能有效地提升异质特征之间的融合效果。MULT、

PMR、MAG-BERT、TETFN 模型都引入了跨模态注意力机制, 并提升了模型的性能, 这说明通过跨模态交互机制, 可以使不同模态将关注点集中在更有价值的信息上。本文模型相比上述 4 个模型在 Acc7、Acc2、F_1 上均有所提升, 原因在于跨模态注意力机制的使用并不能保证模态间交互的信息是与分类任务密切相关的, 在一定程度上会来带冗余信息。MISA 通过挖掘多模态特征间存在的共性特征和独立性特征促进多模态信息融合, 在一定程度上提升了模型的综合性能, 本文模型在 3 个指标上的性能分别提升了 1.8、1.1、1.1 百分点, 原因在于 MISA 仅仅通过距离度量来约束模态间的相似性, 却忽略了时间序列中带有更细粒度的情感特征。MMIM 通过在单模态输入对以及多模态融合结果和单模态输入之间分层最大化互信息, 通过多模态融合保持任务相关信息, 本文模型在 3 个指标上的性能分别提升了 1.0、1.0、1.2 百分点, 原因在于 MMIM 虽然降低了关键信息丢失的概率, 但是并没有充分挖掘多模态数据间的情感交互信息。与 SELF_MM 相比, 本文模型在 CMU_MOSI上, Acc7、Acc2、$F_1$3 个指标上提升了 1.0、0.8、0.7 百分点, 在 CMO_MOSEI 上 Acc2 和 F_1 指标上的性能分别提升了 0.6、0.7 百分点, 原因在于 Self-MM 标签都是通过自监督学习过程生成的, 可能会给实验的预测结果带来一定的负面影响。DMD 旨在通过蒸馏单元来学习模态间的交互作用, 旨在平衡模态间的异质性, 与其相比, 本文模型在 CMU_MOSEI 上 Acc7、Acc2、$F_1$3 个指标提升了 0.9、1.1、1.1 百分点, 原因在于 DMD 在模态间进行信息交互时无法过滤一些与情感信息无关的特征, 使得多模态特征表示含有冗余信息。

表 2　基于 CMU-MOSI、CMU-MOSEI 数据集的模型对比结果

Table 2　Model comparison results based on CMU-MOSI and CMU-MOSEI datasets

模型	CMU_MOSI					CMU_MOSEI				
	Acc7/%	Acc2/%	F_1/%	MAE	Corr	Acc7/%	Acc2/%	F_1/%	MAE	Corr
TFN	32.2	76.4	76.3	1.017	0.604	49.8	79.4	79.7	0.610	0.671
LMF	30.6	73.8	73.7	1.026	0.602	50.0	80.6	81.0	0.608	0.677
MULT	39.1	81.1	81.0	0.889	0.686	50.7	81.6	81.6	0.591	0.694
PMR	40.6	82.4	82.1	—	—	51.8	83.1	82.8	—	—
MISA	41.3	83.5	83.5	0.776	0.778	52.0	84.6	84.6	0.557	0.751
SELF_MM	46.6	85.4	85.4	**0.708**	0.796	53.8	85.1	84.9	**0.530**	0.764
MMIM	45.7	83.0	83.1	0.751	0.764	52.8	84.7	84.5	0.549	0.747
MAG-BERT	42.9	83.5	83.5	0.790	0.769	51.9	85.0	85.0	0.602	0.778
TETFN	—	84.0	83.8	0.717	0.800	—	84.2	84.1	0.551	0.748
DMD	40.2	81.3	81.2	0.897	0.691	52.9	84.6	84.6	0.558	0.737
本文模型	**47.6**	**86.2**	**86.1**	0.726	**0.801**	**53.8**	**85.7**	**85.7**	0.593	**0.785**

通过对上述模型的对比与分析, 充分说明了本文提出的模型在多模态情感分类任务上的优越性。通过图学习的方式自适应地调整不同时刻模态间的交互权重, 来应对主导模态的变化。在语义和时间 2 个维度上的约束使得模态间的交互可以到学习潜在相似性的不变表示, 捕获更加细粒度的特征, 有效降低了模态间的异质性问题。信息瓶颈的引入使得多模态特征表示与分类任务密切相关, 降低了信息冗余性, 提高了情感任务分类的准确率。

3.4　消融实验

为了进一步探究模型中各个子模块的具体作用及其对于最终结果的影响, 对本文模型进行了消融实验, 在 5 项指标中, 更为重要的 3 项指标为 Acc7、Acc2、F_1, 七分类指标是在二分类指标的基础上更为细粒度的划分, 而 F_1 指标的变化情况与 Acc2 指标的变化情况大体一致, 因此在接下来的分析中, 主要关注各组实验在 Acc2 指标上的变化。

3.4.1　不同模态融合对比实验

为了进一步评估每种模态在模型中的贡献度以及图学习模块的作用, 在 CMU_MOSI 数据集上设计了 3 组消融实验, 实验结果见表 3。

表 3　不同模态输入的消融研究结果

Table 3　Ablation study results for different modal inputs

编号	模态	Acc7/%	Acc2/%	F_1/%	MAE	Corr
1	文本	44.5	83.8	83.7	0.758	0.773
2	音频	18.1	60.8	60.8	1.429	0.087
3	视觉	19.7	62.4	62.4	1.420	0.061
4	文本、图学习	45.7	84.3	84.3	0.751	0.774
5	音频、图学习	42.1	81.7	81.7	0.788	0.736
6	视觉、图学习	42.4	81.6	81.6	0.784	0.748
7	文本、音频	47.5	82.6	82.6	0.762	0.750
8	文本、视觉	47.7	83.2	83.2	0.744	0.774
9	视觉、音频	21.0	62.9	62.8	1.404	0.073
10	文本、音频、图学习	45.8	84.1	84.1	0.738	0.768
11	文本、视觉、图学习	46.4	85.0	85.0	0.733	0.781
12	视觉、音频、图学习	46.1	84.7	84.7	0.746	0.775
13	文本、音频、视觉	46.3	85.1	85.1	0.751	0.783
14	文本、音频、视觉、图学习	**47.6**	**86.2**	**86.1**	**0.726**	**0.801**

第 1 组实验，主要探索单模态的个体作用，对应于表 3 中编号为 1 ~ 6 的实验，只使用一种模态特征作为预测，而排除其他模态。结果表明：1) 在 3 种模态，文本模态效果最好，而视觉和音频模态效果比较差，这充分说明了文本模态相对另外 2 种模态具有更高阶的特征。2) 在引入图学习模块后，各个模态的预测效果均有所提升，尤其是音频和视觉模态在 Acc2 指标上分别提升了 20.9 百分点和 19.2 百分点，这充分说明了通过图学习，各个模态分别从其他模态吸取了丰富的情感特征，弥补了自身模态的不足，这也证明了图学习模块的有效性。

第 2 组实验，主要探索了 2 种模态之间的相互作用，3 种模态的相互组合对应于表 3 中 7 ~ 12 号的实验，结果表明：1) 双模态总体上优于单模态，这初步证明了不同模态下的情感信息的互补性。2) 图学习融合模块的引入进一步促进了双模态之间的相互作用，有效提升了情感分类结果。

3.4.2　各模块对模型性能的影响

为了探究模型中各个子模块以及损失函数的作用，在 CMU_MOSI 数据集上设计了 2 组消融实验，表 4 中编号为 1 的实验表示完整模型，w/o 表示在完整模型上去除一些内容，CoAn 表示跨模态注意力机制，GFM 表示图学习单元模块，IB 表示信息瓶颈机制。

表 4　不同模块对模型性能的影响

Table 4　Impact of different modules on model performance

编号	状态	Acc7/%	Acc2/%	F_1/%	MAE	Corr
1	本文模型	**47.6**	**86.2**	**86.1**	**0.726**	**0.801**
2	w/o CoAn	46.5	85.2	85.2	0.725	0.786
3	w/o GFM	44.8	83.9	83.9	0.768	0.766
4	w/o IB	45.3	84.7	84.7	0.745	0.779
5	w/o CoAn and IB	46.3	83.3	83.3	0.761	0.776
6	w/o GFM and IB	43.6	82.7	82.7	0.768	0.767
7	w/o $\mathcal{L}_{\text{sem}}$	46.2	84.2	84.1	0.729	0.781
8	w/o $\mathcal{L}_{\text{tim}}$	45.6	84.6	84.5	0.731	0.788
9	w/o $\mathcal{L}_{\text{sem}}$、$\mathcal{L}_{\text{tim}}$	44.9	83.8	83.8	0.745	0.775

第 1 组实验, 主要探究了各个子模块对模型性能的影响, 观察表 4 中编号为 2 ~ 6 的实验现象可以得出: 1) 在去除跨模态注意力 CoAn 后, 模型性能下降了 1.0 百分点, 这说明 CoAn 可以有效捕获模态间的关联和语义对齐。2) 在去除图学习模块 GFM 后, 对指标影响最大, Acc2 下降 2.3 百分点, Acc7 下降了 2.8 百分点, 这说明通过 GFM, 模态间可以进行充分的交互, 来适应不同时刻主导模态的变化, 使得彼此间获得最为丰富的情感特征, 来弥补自身模态的不足。3) 在去除信息瓶颈 IB 后, 模型性能下降 1.5 百分点, 这说明 IB 可以有效去除多模态特征表示中的冗余信息, 使得多模态表示只与分类任务相关。4) 在去除双模块 CoAn-IB 和 GFM-IB 后, 模型性能分别下降了 2.9 百分点和 3.5 百分点, 表明子模块间合作的有效性和必要性。

第 2 组实验, 主要探究了图学习模块中的损失约束的作用, 观察表 4 中编号为 7 ~ 8 的实验结果可以得出, 在单独去除对图学习过程中空间和时间上的约束后, 模型指标有所下降, 尤其是当 2 种约束都去除后, 对指标影响最大, Acc2 下降了 2.4 百分点。这说明在时间和空间上的约束可以使得模态间捕获潜在相似性的不变表示, 进一步缓解模态间的差异性。

3.4.3　超参数敏感性分析

在语义层面相似性约束函数和信息瓶颈机制中, 引入了用于调节权重的标量α和β, 具体如式 (5) 和式 (3) 所示。为了探究α和β对于模型性能的影响, α和β取值为 0.1, 0.01 和 0.001, 实验结果见表 5。从表 5 中可以看出: 对于 CMU_MOSI 数据集, α取 0.01、β取 0.001 时, 效果最好。对于 CMU_MOSEI 数据集, 各项指标差异不大, 在α为 0.01、β为 0.01 时, Acc7 准确率达到最高, 但在综合考虑下, 取α为 0.01、β为 0.001 作为本文模型的最终取值。

表 5　超参数 α 和 β 对模型性能的影响

Table 5　Impact of hyperparameters α and β on model performance

参数取值	CMU_MOSI					CMU_MOSEI				
	Acc7/%	Acc2/%	F_1/%	MAE	Corr	Acc7/%	Acc2/%	F_1/%	MAE	Corr
α=0.1, β=0.1	46.8	84.5	84.4	0.738	0.777	52.4	84.9	84.9	0.606	0.778
α=0.1, β=0.01	43.9	83.7	83.7	0.747	0.773	53.1	85.3	85.2	0.596	0.782
α=0.1, β=0.001	46.9	84.6	84.6	**0.716**	0.786	52.5	85.1	85.1	0.614	0.776
α=0.01, β=0.1	46.0	84.5	84.5	0.720	0.781	52.9	84.9	84.9	0.598	0.778
α=0.01, β=0.01	47.1	85.4	85.4	0.731	0.792	**54.1**	85.2	85.2	0.592	0.764
α=0.01, β=0.001	**47.6**	**86.2**	**86.1**	0.726	**0.801**	53.8	**85.7**	**85.7**	**0.593**	**0.785**
α=0.001, β=0.1	46.7	84.6	84.6	0.746	0.784	52.1	84.6	84.5	0.612	0.776
α=0.001, β=0.01	47.5	85.6	85.6	0.730	0.783	53.4	85.1	85.0	0.592	0.763
α=0.001,β=0.001	45.7	84.9	84.9	0.732	0.774	53.3	85.4	85.2	0.594	0.781

对于参数α, 当 α=0.1 时, 模型会过于关注每个模态的特定细节, 而忽略了它们之间共享的情感信息。当α=0.001 时, 同一情感在不同模态下的表示过于相似, 使得模型无法很好地区分不同模态之间的细微差异, 导致信息冗余。而α=0.01 时, 模型能够更好地平衡同一情感不同模态之间的相似性和不同情感相同模态之间的差异性, 从而提高模型的性能。

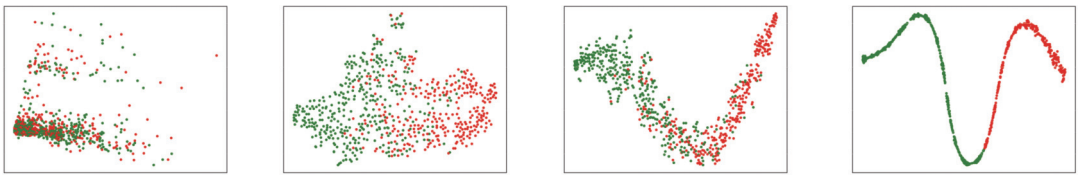
对于参数β, 它控制了多模态表示向量与输入数据之间互信息值, 当β取值较小时, 模型性能最佳。这是因为较小的β会使得模型更加关注多模态特征表示与真实标签之间的互信息最大值, 同时过滤掉了多模态表示中的冗余信息, 又保留了其中的关键信息。

3.5　结果可视化

3.5.1　分类结果可视化

为了更加直观地查看分类结果, 本节提供了多模态情感分类可视化结果, 以分析本文模型在各个阶段的处理过程。图 2 给出了在数据集 CMU_MOSI 上二分类任务中的可视化结果。图 2(a) ~ (d) 分别为多模态特征表示从初始状态、特征提取状态、图学习融合状态、情感分类状态。从图 2(a) 中可以看出, 每一个点凌乱地散布在二维平面中; 经过特征提取后, 不同类别的情感状态得到了有效的区分; 在经过图学习融合阶段后, 分类结果得到进一步改善, 但是仍然有一

些点因为冗余信息的存在而不能被正确地分类(图 2(c) 所示); 最后通过信息机制去除冗余信息，使得多模态表示只与分类相关 (图 2(d) 所示), 分类效果达到最佳。



(a) 原始特征　(b) 模态特征提取　(c) 自适应图学习　(d) 信息瓶颈去冗余

图 2　CMU_MOSI 二分类过程

Fig. 2　Binary classification process on CMU_MOSI

3.5.2　权重结果可视化

图 3 给出了在图学习单元中, 图中边权重的变化情况, 其中 6 条曲线表示 6 个不同的交互方向。从图 3 中可以看出: 模态间的交互方向主要是以 L−>V 和 L−>A 为主, 而其他 4 个方向的交互强度相对较弱, 这是因为文本模态相对视觉和音频模态, 仍然在模态间交互时发挥着关键作用, 并且比视觉和音频模态更具有优势。在表 3 中编号为 1 ~ 3 的实验中可以得出, 文本模态具有高阶的语义特征, 而视觉和音频模态具有低阶语义特征, 且含有大量的冗余信息。这也充分表明了本文提出的模型能够有效地实现模态间的交互, 使得情感信息能够均匀地分布在不同模态中, 进而有效地缓解了模态间的异质性。

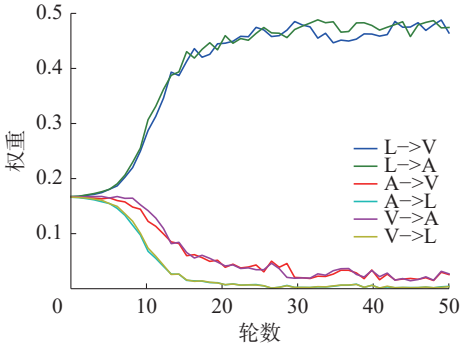


图 3　模态交互权重可视化

Fig. 3　Visualization of modal interaction weights

4　结束语

为了使情感信息均匀分布在每个模态并减少多模态特征表示的冗余性, 本文提出了一种基于自适应图学习权重的多模态情感分析方法。通过不同的特征提取方法对不同模态进行特征提取, 有效获得了模态内的特定信息。将 3 种模态对任务分类的预测值以及模态表示嵌入到自适应图中, 通过模态标签学习不同模态对最终分类任务的贡献度来动态调整模态间的权重, 使得模态间的情感信息得到了有效的平衡, 在语义层面和时间层面的相似性约束有效地缓解了模态间的差异性, 获得了模态间潜在的相似性不变表示。将多模态特征表示通过信息瓶颈机制进行去噪, 使得多模态情感表示只与任务分类相关, 有效地去除了其中的冗余信息。将本文模型与现有的多模态情感分析模型在情感分析进行对比, 结果显示本文模型在任务测试结果均优于对比模型, 验证了本文模型在情感分析任务中的有效性。当前, 多模态有效融合仍然是一个具有挑战性的任务, 不同的融合策略和融合方法都会导致分类任务结果的不同。在未来的工作中, 将探索一种更加高效的融合方法, 并进一步研究影响不同模态间有效融合的因素。

参考文献:

[1] PEÑA D, AGUILERA A, DONGO I, et al. A framework to evaluate fusion methods for multimodal emotion recognition[J]. IEEE access, 2023, 11: 10218−10237.

[2] ZHANG Junling, WU Xuemei, HUANG Changqin. AdaMoW: multimodal sentiment analysis based on adaptive modality-specific weight fusion network[J]. IEEE access, 2023, 11: 48410−48420.

[3] 张亚洲, 戎璐, 宋大为, 等. 多模态情感分析研究综述[J]. 模式识别与人工智能, 2020, 33(5): 426−438.
ZHANG Yazhou, RONG Lu, SONG Dawei, et al. A survey on multimodal sentiment analysis[J]. Pattern recognition and artificial intelligence, 2020, 33(5): 426−438.

[4] GANDHI A, ADHVARYU K, PORIA S, et al. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Information fusion, 2023, 91: 424−444.

[5] MAI Sijie, HU Haifeng, XING Songlong. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 164−172.

[6] ZADEH A, LIANG P P, PORIA S, et al. Multi-attention recurrent network for human communication comprehension[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 5642−5649.

[7] HAN Wei, CHEN Hui, GELBUKH A, et al. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis[C]//Proceedings of the 2021 International Conference on Multimodal Interaction. Montréal: ACM, 2021: 6−15.

[8] 刘颖, 王哲, 房杰, 等. 基于图文融合的多模态舆情分析[J]. 计算机科学与探索, 2022, 16(6): 1260−1278.
LIU Ying, WANG Zhe, FANG Jie, et al. Multi-modal public opinion analysis based on image and text fusion[J]. Journal of frontiers of computer science and technology, 2022, 16(6): 1260−1278.

[9] HUANG Changqin, ZHANG Junling, WU Xuemei, et al. TeFNA: text-centered fusion network with crossmodal attention for multimodal sentiment analysis[J]. Knowledge-based systems, 2023, 269: 110502.

[10] SUN Hao, LIU Jiaqing, CHEN Y W, et al. Modality-invariant temporal representation learning for multimodal sentiment classification[J]. Information fusion, 2023, 91: 504−514.

[11] MAI Sijie, ZENG Ying, HU Haifeng. Multimodal information bottleneck: learning minimal sufficient unimodal and multimodal representations[J]. IEEE transactions on multimedia, 2023, 25: 4121−4134.

[12] ZADEH A, CHEN Minghai, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[EB/OL]. (2017−07−23)[2024−01−02]. https://arxiv.org/abs/1707.07250.

[13] TSAI Y H H, BAI Shaojie, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the Conference Association for Computational Linguistics Meeting. Florence: ACL, 2019: 6558−6569.

[14] SU Guixin, HE Junyi, LI Xia, et al. NFCMF: noise filtering and CrossModal fusion for multimodal sentiment analysis[C]//2021 International Conference on Asian Language Processing. Singapore: IEEE, 2021: 316−321.

[15] YANG Shuo, XU Zhaopan, WANG Kai, et al. BiCro: noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 19883−19892.

[16] 孙杰, 车文刚, 高盛祥. 面向多模态情感分析的多通道时序卷积融合[J]. 计算机科学与探索, 2024, 18(11): 3041−3050.
SUN Jie, CHE Wengang, GAO Shengxiang. Multi-channel temporal convolution fusion for multimodal sentiment analysis[J]. Journal of frontiers of computer science and technology, 2024, 18(11): 3041−3050.

[17] 鲍小异, 姜晓彤, 王中卿, 等. 基于跨语言图神经网络模型的属性级情感分类[J]. 软件学报, 2023, 34(2): 676−689.
BAO Xiaoyi, JIANG Xiaotong, WANG Zhongqing, et al. Cross-lingual aspect-level sentiment classification with graph neural network[J]. Journal of software, 2023, 34(2): 676−689.

[18] JANGRA A, MUKHERJEE S, JATOWT A, et al. A survey on multi-modal summarization[J]. ACM computing surveys, 2023, 55(13s): 1−36.

[19] MAJUMDER N, HAZARIKA D, GELBUKH A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-based systems, 2018, 161: 124−133.

[20] XU Nan, MAO Wenji. MultiSentiNet: a deep semantic network for multimodal sentiment analysis[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017: 2399−2402.

[21] AKHTAR M S, CHAUHAN D S, GHOSAL D, et al. Multi-task learning for multi-modal emotion recognition and sentiment analysis[EB/OL]. (2019−05−14) [2024−01−02]. https://arxiv.org/abs/1905.05812.

[22] HAZARIKA D, ZIMMERMANN R, PORIA S. MISA: modality-invariant and-specific representations for multimodal sentiment analysis[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020: 1122−1131.

[23] YU Wenmeng, XU Hua, YUAN Ziqi, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2021: 10790−10797.

[24] ZUO Haolin, LIU Rui, ZHAO Jinming, et al. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023: 1−5.

[25] LIU A H, JIN S, LAI C I J, et al. Cross-modal discrete representation learning[EB/OL]. (2021−06−10) [2024−01−02]. https://arxiv.org/abs/2106.05438.

[26] 胡文彬, 陈龙, 黄贤波, 等. 融合交叉注意力的突发事件多模态中文反讽识别模型[J]. 智能系统学报, 2024, 19(2): 392−400.
HU Wenbin, CHEN Long, HUANG Xianbo, et al. A mul-

timodal Chinese sarcasm detection model for emergencies based on cross attention[J]. CAAI transactions on intelligent systems, 2024, 19(2): 392−400.

[27] 李梦云, 张景, 张换香, 等. 基于跨模态语义信息增强的多模态情感分析[J]. 计算机科学与探索, 2024, 18(9): 2476−2486.
LI Mengyun, ZHANG Jing, ZHANG Huanxiang, et al. Multimodal sentiment analysis based on cross-modal semantic information enhancement[J]. Journal of frontiers of computer science and technology, 2024, 18(9): 2476−2486.

[28] 包广斌, 李港乐, 王国雄. 面向多模态情感分析的双模态交互注意力[J]. 计算机科学与探索, 2022, 16(4): 909−916.
BAO Guangbin, LI Gangle, WANG Guoxiong. Bimodal interactive attention for multimodal sentiment analysis[J]. Journal of frontiers of computer science and technology, 2022, 16(4): 909−916.

[29] TANG Jiajia, LIU Dongjun, JIN Xuanyu, et al. BAFN: bi-direction attention based fusion network for multimodal sentiment analysis[J]. IEEE transactions on circuits and systems for video technology, 2023, 33(4): 1966−1978.

[30] LYU Fengmao, CHEN Xiang, HUANG Yanyong, et al. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 2554−2562.

[31] RAHMAN W, HASAN M K, LEE Sangwu, et al. Integrating multimodal information in large pretrained transformers[C]//Proceedings of the conference. Association for Computational Linguistics. Online: ACL, 2020: 2359−2369.

[32] SANGWAN S, CHAUHAN D S, AKHTAR M S, et al. Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis[C]//Neural Information Processing. Sydney: Springer, 2019: 662−669.

[33] HUANG Yanping, PENG Hong, LIU Qian, et al. Attention-enabled gated spiking neural P model for aspect-level sentiment classification[J]. Neural networks, 2023, 157: 437−443.

[34] TISHBY N, PEREIRA F C, BIALEK W. The information bottleneck method[EB/OL]. (2000−04−24) [2024−01−02]. https://arxiv.org/abs/physics/0004057.

[35] ALEMI A A, FISCHER I, DILLON J V, et al. Deep variational information bottleneck[EB/OL]. (2016−12−01) [2024−01−02]. https://arxiv.org/abs/1612.00410.

[36] ZADEH A, ZELLERS R, PINCUS E, et al. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages[J]. IEEE intelligent systems, 2016, 31(6): 82−88.

[37] BAGHER 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 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: ACL, 2018: 2236−2246.

[38] LIU Zhun, SHEN Ying, LAKSHMINARASIMHAN V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[EB/OL]. (2018−05−31) [2024−01−02]. https://arxiv.org/abs/1806.00064.

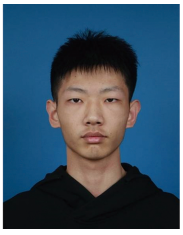
[39] HAN Wei, CHEN Hui, PORIA S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis[EB/OL]. (2021−09−01) [2024−01−02]. https://arxiv.org/abs/2109.00412.

[40] WANG Di, GUO Xutong, TIAN Yumin, et al. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis[J]. Pattern recognition, 2023, 136: 109259.

[41] LI Yong, WANG Yuanzhi, CUI Zhen. Decoupled multimodal distilling for emotion recognition[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 6631−6640.

作者简介：

曲海成，副教授，辽宁工程技术大学软件学院副院长，中国计算机学会会员。主要研究方向为遥感影像高性能计算、视觉信息计算、目标检测与识别。主持辽宁省自然科学基金项目 1 项、辽宁省教育厅面上项目 2 项，发表学术论文 60 余篇。E-mail：quhaicheng@lntu.edu.cn。



徐波，硕士研究生，主要研究方向为多模态情感分析。E-mail：lntu_xubo@163.com。