文章编号： 1673-3193（2025）01-0001-09　　J. North Univ. China, Nat. Sci. Ed., 2025, 46(1): 1-9.

自适应门控解耦特征融合的多模态情感分析

李烽源[1]，蔺素珍[1]，王彦博[1]，李大威[2]，顾梦瑶[1]

（1. 中北大学 计算机科学与技术学院，山西 太原 030051；2. 中北大学 电气与控制工程学院，山西 太原 030051）

摘　要： 现有的多模态情感分析研究主要通过不同模态特征的整体交互进行不同模态信息的融合，未考虑不同模态包含的独有特征以及共有特征之间的联系，导致无法有效分析复杂的情感。针对上述问题，本文提出了自适应门控解耦特征融合的多模态情感分析模型（AGDF）。首先，利用预训练的BERT模型和Transformer模型进行不同模态的特征提取。其次，根据不同模态的共有特征相似而独有特征不相似的原理构造对比对，通过对比学习的方法，将不同模态的特征分解为独有特征和共有特征。然后，根据图像和语音模态在文本模态存在偏移的原理，设计了一种新的自适应门控机制进行特征融合，将其他模态信息融于文本模态。同时，设计了独有特征和共有特征的联系图，利用图注意力神经网络进行融合，以平衡模态之间的独有信息和共有信息。最后，对融合特征进行分类。在数据集CMU-MOSI、CMU-MOSEI上进行了实验，结果显示本文方法比基线方法在准确率和*F*1分数上均提高了约1百分点。此外，与其他特征分解方法相比，本文方法的准确率提高了1.23百分点，*F*1分数提高了1.37百分点，Corr提高了2.13百分点，MAE降低了4.83百分点。综合结果表明，本文提出的方法能够更加充分利用不同模态的异质信息，从而有效提高情感识别的效果。

关键词： 情感分析；对比学习；图神经网络；多模态信息融合；自适应门控

中图分类号： TP391.1　　**文献标识码：** A　　**doi:** 10.62756/jnuc.issn.1673-3193.2024.07.0005

引用格式： 李烽源，蔺素珍，王彦博，等. 自适应门控解耦特征融合的多模态情感分析[J]. 中北大学学报（自然科学版），2025，46(1)：1-9.

LI Fengyuan，LIN Suzhen，WANG Yanbo，et al. Multimodal sentiment analysis of adaptive gated decoupling feature fusion[J]. Journal of North University of China(Natural Science Edition)，2025，46(1)：1-9.

Multimodal Sentiment Analysis of Adaptive Gated Decoupling Feature Fusion

LI Fengyuan[1], LIN Suzhen[1], WANG Yanbo[1], LI Dawei[2], GU Mengyao[1]

(1. School of Computer Science and Technology, North University of China, Taiyuan 030051, China;
2. School of Electrical and Control Engineering, North University of China, Taiyuan 030051, China)

Abstract: In the existing research on multi-modal sentiment analysis, the fusion different modal information is mainly through the overall interaction of different modal features, but it doesn't consider the relationship between unique features and common features contained in different modes, so the complex emotions can't be analyzed effectively. To solve this problem, a multimodal sentiment analysis model based on adaptive gated decoupling feature fusion (AGDF) was proposed. Firstly, the pre-trained BERT model and Transformer model were used for feature extraction of different modes. Secondly, according to the principle that the common features of different modes were similar but the unique features were not similar, the contrast pair was constructed. By contrastive learning, the features of different modes were decomposed into unique features and

收稿日期： 2024-07-05

基金项目： 国家自然科学基金项目（62271453）；山西省自然科学基金项目（202303021211147）；山西省应用基础研究计划（20210302123025）；山西省知识产权局专利转化专项计划（202302001）

作者简介： 李烽源（1999—），男，硕士生，主要从事多模态融合和情感分析的研究。

通信作者： 蔺素珍（1966—），女，教授，博士，主要从事数字图像处理的研究。E-mail：lsz@nuc.edu.cn。

common features. Thirdly, according to the principle that the image and speech modes were offset in the text mode, a new adaptive gating mechanism was designed to fuse the features and integrate other modal information into the text mode. At the same time, the relation graph of unique features and common features was designed, and the fusion of the graph attention neural network was used to balance the unique information and common information among the modes. Finally, the fusion features were classified. Experiments on the datasets CMU-MOSI and CMU-MOSEI show that the accuracy and $F1$ score of the proposed method are improved by about 1 percentage point compared with the baseline method. In addition, compared with other feature decomposition methods, the proposed method improves accuracy by 1.23 percentage point, $F1$ score by 1.37 percentage point, Corr by 2.13 percentage point, and reduces MAE by 4.83 percentage point. Consequently, the proposed method can make full use of the heterogeneous information of different modes and effectively improve the effect of sentiment analysis.

Key words: sentiment analysis; contrastive learning; graph neural network; multimodal information fusion; adaptive gating

0 引 言

情感是人类真实意图的流露。在互联网视频数量激增的今天，准确捕获网络视频所承载的复杂场景中互动多方的真实情感信息对于推荐系统研发[1]、社交媒体分析[2]和金融分析[3]等具有重要意义。

现阶段，多数研究都倾向于利用复杂的信息融合机制来提高情感分析的准确性。例如，基于量子的融合模型[4-5]和基于注意力机制的融合模型[6-7]等方法都是通过复杂的多模态特征融合在某些场景中提升情感分析的准确度。最新研究开始关注不同模态的共有信息和独有信息[8-9]，通过区分两种特征，可避免模型忽略关键信息，从而提高预测的准确性，这对增强多模态特征表示具有重大意义。共有信息是指在不同模态中常见的特征，通过共有信息，可以促进不同模态之间的相互作用。独有信息是指各种模态所独有的特征，可以为其他模态提供额外的上下文或补充信息，从而增强多模态表征。特征解耦主要通过投影法实现，如通过全连接层将多模态特征映射到两个公共子空间[8,10]，通过对比学习对模态的特征进行解耦[11]。这些研究虽注重对独有特征和共有特征的提取，但没有充分利用所提取到的独有和共有特征之间的关系，往往通过CONCAT操作连接解耦的特征进行情感分析，未能有针对性地制定更精细的融合策略。忽略不同模态信息中包含的独有信息和共有信息之间联系的后果是模型往往无法有效应对复杂的情感分析。

图1展示了不同模态的数据反应的情感信息及不同模态特征的二维可视化图像。

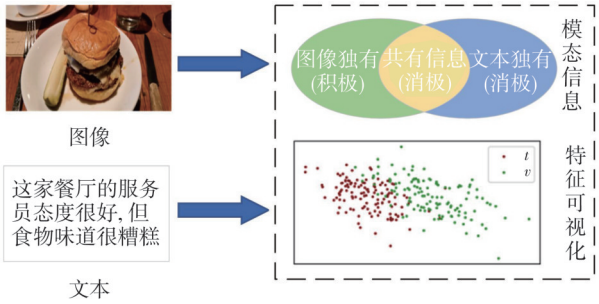

图1 共有特征和独有特征的差别示例

Fig. 1 Examples of differences between common and unique features

由图1模态信息中可以看出，文本中包含两种不同的情感倾向，但是可以推断出情绪是消极的，而由美味的食物图片可推断出的情绪是积极的。同时考虑图像和文本的共有特征和独有特征可以推断出图像中食物的情绪是消极的，表示对食物的不满，而不是对食物的喜爱。使用t-SNE算法可视化不同模态的提取特征，其中，t表示文本特征，v表示视觉特征，可以看出，两种模态特征在空间上具有相交部分也有不相交的部分。相交部分组成的公共空间可以表示其共有特征，而不相交部分组成的私有空间可以表示独有特征。从该示例中可以看到不同模态的信息可以互相影响，而模型往往会忽略这些特征之间的联系，从而影响了模型性能。

针对不同模态之间独有特征和共有特征之间的联系交互不足的问题，本文将图神经网络(Graph Neural Network，GNN)引入多模态情感分析领域，用于对其关系建模，提出了一种自适应门控解耦特征融合的多模态情感分析模型。首

先利用预训练的BERT 和 Transformer模型进行不同模态的特征提取，采用对比学习损失进行特征解耦；接着，使用一种新的自适应的门控融合网络融合已解耦的特征，同时引入GAT网络学习解耦特征之间的关系；最后，对融合的特征进行情感分类识别。本文的主要贡献如下：

1）针对模型未考虑模态间独有特征和共有特征而导致的复杂情感识别困难的问题，本文提出自适应门控解耦特征融合的多模态情感分析模型。该模型能够有效利用不同模态之间的互补信息，从而提高情感预测的准确性。

2）本文设计了一种新的融合策略来融合解耦的特征。通过图注意力网络和门控自适应的特征融合策略，充分学习模态间共有特征和独有特征的联系。

3）在两个基准数据集上的实验表明，本文提出的模型效果优于目前主流的方法。

1　相关工作

1.1　多模态情感分析

现有的模型按融合方式可分为特征级融合、决策级融合以及混合融合。

特征级融合方法通常将多模态的特征作为融合网络的输入，如使用CONCAT 拼接、相加等简单的方式进行跨模态融合[12-13]。这种方法在特征提取后，直接进行特征的合并，未能捕捉到模态间的层次关联信息，这导致不同模态间的动态交互被忽略，丢失了上下文信息。

决策级融合方法在多个模型进行独立情感分析的基础上进行结果的融合[14-15]，通过平均、加权以及投票等方式进行决策。例如：在特征提取之后采用专家模型进行决策，并通过门控机制选择整体的情感特征[14]，但这种方法既耗时又未能充分利用不同模态间的交互信息。

混合融合方法首先学习模态内表示，然后进行模态间融合。例如：通过双向注意力进行融合[16]，使用笛卡尔积显式地模拟单模态、双模态和三模态相互作用；采用多头注意力进行融合[17]，同时考虑了模态内以及模态间的相互作用；语言引导的多级关联融合[18]利用了从低级到高级的模态相关信息。尽管这些方法在多模态特征融合领域取得了显著的进展，但它们在有效利用不同模态之间的共有特征和独有特征方面仍存在不足，限制了融合模型的性能。

本文采用图注意力和自适应门控的策略来融合解耦的特征，以实现对解耦特征的充分交互。

1.2　对比学习

对比学习通过正样本和负样本之间的对比进行学习，其常用的对比损失函数为

$$l_c(p,q)=\sum_{b\in p}-\log\left(\frac{\exp(\mathrm{sim}(b)/\tau)}{\sum_{k\in p\cup q}\exp(\mathrm{sim}(k)/\tau)}\right),\quad(1)$$

式中：p为正样本对；q为负样本对；τ为温度参数，用于相似度调节；sim为相似度函数。

sim定义为

$$\mathrm{sim}(\boldsymbol{a},\boldsymbol{b})=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}/\|\boldsymbol{a}\|\|\boldsymbol{b}\|,\quad(2)$$

式中：$\mathrm{sim}(\boldsymbol{a},\boldsymbol{b})$为向量$\boldsymbol{a}$和$\boldsymbol{b}$的余弦相似度；$\|\cdot\|$为模长；T为转置运算。

对比学习在神经网络中已经取得了显著进展[19-22]，其工作的核心在于构造合适的正负样本对。例如：通过对数据增广获得更多的正负样本，可以大幅提高学习表征的质量[19]；通过不计算某些梯度的方式，可以实现在不需要负样本的情况下进行表征学习[20]；通过文本图像对比对进行无监督的预训练，使得模型性能达到了与有监督学习相竞争的程度[21]；通过采用多意图的对比学习，实现对用户细粒度意图的有效区分，并能够选择主要意图[22]。

本文采用文献[13]的对比损失进行不同模态的特征解耦。

1.3　图神经网络

近几年，GNN模型因其对图结构化数据的适用性而变得越来越流行[23]。图神经网络可分为两大类：基于频谱的方法和基于空间的方法。

基于频谱的方法通过以类似于图信号处理的方式定义滤波器来实现图的卷积。基于空间的方法通过信息传播来定义图的卷积，由于其更高的效率、灵活性和通用性，得到了广泛应用，如ROLAND[24]、RAHG[25]和GAT[26]等。尤其是GAT，它能够自适应地为图中的节点分配权重，提高了模型对复杂数据关系的处理能力，并且利用并行计算和参数效率高的特点，增强了模型的泛化能力和计算效率，因而受到了极大的欢迎。

本文将不同模态的独有特征和共有特征构造为图数据，然后采用图注意力的方式将其融合，

以学习解耦特征间的联系。

2 AGDF模型

AGDF的总体结构如图2所示。模型共包含3个主要模块：特征提取模块、特征解耦模块和解耦特征融合模块。

特征提取模块将多模态原始数据提取为3个特定的低维特征表示。使用预训练的BERT模型对文本模态T进行编码，使用两个独立的Transformer编码器分别对视觉模态V和语音模态A进行编码。

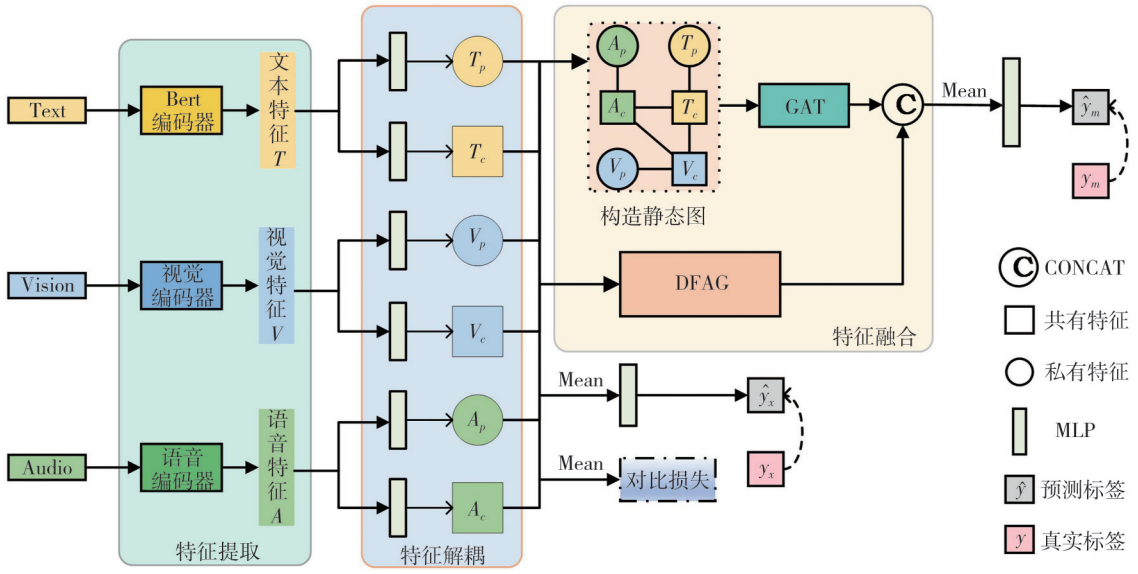

图2 AGDF结构图

Fig. 2 Structure diagram of AGDF

特征解耦模块采用对比学习的方式，将每一个模态的低维特征解耦为与模态相关的独有特征和与模态无关的共有特征。

解耦特征融合模块负责融合解耦的特征。融合过程如下：首先，设计了一个多模态解耦特征自适应融合模块，将非文本模态的信息融入文本中。其次，根据解耦特征之间的联系，引入图注意力网络进行融合，以强化特征间的关联性。

2.1 特征提取模块

文本特征提取采用预训练的BERT模型进行编码。对于文本$T_0=\{w_1,w_2,\cdots,w_n\}$，BERT模型能够对其嵌入位置信息、[$cls$]分类信息等，并且通过双向编码更好地理解上下文信息，将句子T_0转化为二维矩阵$\boldsymbol{T}$。编码过程表示为

$$\boldsymbol{T}=\mathrm{BERT}(T_0,\theta_{\mathrm{BERT}})\in\mathbb{R}^{s\times h},\tag{3}$$

式中：θ为BERT模型的参数；s为词元数；h为词元的向量长度，在BERT模型中，h=768。

视觉特征提取，首先采用openface库提取人脸面部特征V_0，然后经过双层多头自注意力Transformer网络f_V进行特征编码。编码过程表示为

$$\boldsymbol{V}=f_V(V_0,\theta_{f_A}).\tag{4}$$

语音特征提取，首先采用librosa库提取视频中的语音特征A_0，然后经过双层多头自注意力Transformer网络f_A进行特征编码。编码过程表示为

$$\boldsymbol{A}=f_A(A_0,\theta_{f_A}).\tag{5}$$

2.2 特征解耦模块

对比学习用来实现多模态数据特征的解耦，且以样本内和样本间两种方式构建对比学习样本。选择文本相似性作为锚点，使得视觉和音频共有特征接近文本特征，而不同模态的独有特征远离文本特征。同时，根据余弦相似性选择正样本集和负样本集，使得模型可以更容易地识别困难样本。上述的特征解耦过程表示为

$$\boldsymbol{X}_c=MLP_1(\boldsymbol{X}),\tag{6}$$

$$\boldsymbol{X}_p=MLP_2(\boldsymbol{X}),\tag{7}$$

式中：$\boldsymbol{X}$为不同的模态特征；$\boldsymbol{X}_c$为不同模态的共有特征；$\boldsymbol{X}_p$为不同模态的独有特征；MLP_1，MLP_2分别为非共享参数的MLP模型。

对比损失

$$l_c=ContrastLoss(\boldsymbol{T}_c,\boldsymbol{T}_v,\boldsymbol{V}_c,\boldsymbol{V}_v,\boldsymbol{A}_c,\boldsymbol{A}_v).\tag{8}$$

2.3 解耦特征融合模块

本模块旨在探索以不同的方式对解耦特征进

行融合。一方面利用解耦特征之间的联系并采用图网络进行融合，另一方面利用视觉和语音模态的偏移进行融合[27]。

在构建静态图时，本文考虑了以下原则：共有特征之间的相似性、同一模态内特征之间的相似性、相邻帧之间共有特征的时序相似性。这是由于视频的情感在相邻帧以及不同模态数据的共有特征之间往往会有相同的情绪极性，而在同一模态的情感极性则可能不同。

1）边和节点的构造。给定一个样本不同模态的 m 个特征 $\{\boldsymbol{T}_p^t, \boldsymbol{T}_c^t, \boldsymbol{V}_p^t, \boldsymbol{V}_c^t, \boldsymbol{A}_p^t, \boldsymbol{A}_c^t\}$，将其作为图的节点，其中 $t\in[1,m]$。图数据的边可由以下几类组成：不同模态共有特征 $\{\boldsymbol{T}_c^t\boldsymbol{V}_c^t, \boldsymbol{A}_c^t\boldsymbol{T}_c^t, \boldsymbol{A}_c^t\boldsymbol{V}_c^t\}$；同模态解耦特征 $\{\boldsymbol{T}_c^t\boldsymbol{T}_p^t, \boldsymbol{A}_c^t\boldsymbol{A}_p^t, \boldsymbol{V}_c^t\boldsymbol{V}_p^t\}$；相邻共有特征 $\{\boldsymbol{V}_c^t\boldsymbol{V}_c^{t+1}, \boldsymbol{A}_c^t\boldsymbol{A}_c^{t+1}, \boldsymbol{T}_c^t\boldsymbol{T}_c^{t+1}\}$。三种关系的六关系图 G6 如图 3 所示。

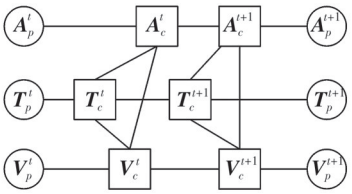

图 3　关系图

Fig. 3　Relational graph

2）图注意力融合。图注意力神经网络在分类任务中展现了显著的有效性，它能够捕获节点之间更细粒度的关系，从而实现更高效的通信和更高质量的信息聚合。基于此，采用图自注意力网络（Graph Attention Network，GAT）来聚合图中的邻域信息，具体为

$$\boldsymbol{h}_v=\sigma\left(\sum_{u\in\mathcal{N}(v)}\boldsymbol{\alpha}_{uv}\boldsymbol{W}\boldsymbol{h}_u\right),\tag{9}$$

式中：$\sigma(\cdot)$ 为 LeakyReLU 激活函数；$\mathcal{N}(v)$ 为节点 v 的相邻节点；$\boldsymbol{h}_u$ 和 $\boldsymbol{h}_v$ 分别为节点 u 和节点 v 的特征；$\boldsymbol{\alpha}_{uv}$ 为两个节点 u 和 v 的注意力分数。

$$\boldsymbol{\alpha}_{uv}=Softmax(\sigma(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_v,\boldsymbol{W}\boldsymbol{h}_u])),\tag{10}$$

式中：$\boldsymbol{a}$ 为权重注意向量；$\boldsymbol{W}$ 为共享权重矩阵。

3）自适应融合。基于语音和视觉信息的表示可以转换的思想被用于了融合文本、语音和视觉数据的方法[27]中。为了融合解耦之后的不同模态的特征，本文设计了多模态解耦特征自适应门控融合模块（Multimodal Decoupling Feature Adaptive Gated Fusion Module，DFAG），如图 4 所示，计算语音和视觉数据在文本语义空间中发生的偏移。偏移过程表示为

$$\boldsymbol{E}^t=\boldsymbol{g}_a(\boldsymbol{W}_a\boldsymbol{A}_p^t)+\boldsymbol{g}_v(\boldsymbol{W}_v\boldsymbol{V}_p^t)+\boldsymbol{b},\tag{11}$$

式中：$\boldsymbol{W}_a$，$\boldsymbol{W}_v$ 分别为语音和视觉特征的权重向量；$\boldsymbol{b}$ 为偏置偏移；$\boldsymbol{g}_a$，$\boldsymbol{g}_v$ 分别为语音和视觉特征在文本语义空间中的偏移向量。

$$\boldsymbol{g}_a=ReLU(\boldsymbol{W}_{ga}[\boldsymbol{A}_p^t,\boldsymbol{T}_p^t])+\boldsymbol{b}_a,\tag{12}$$

$$\boldsymbol{g}_v=ReLU(\boldsymbol{W}_{gv}[\boldsymbol{V}_p^t,\boldsymbol{T}_p^t])+\boldsymbol{b}_v,\tag{13}$$

式中：$\boldsymbol{W}_{ga}$，$\boldsymbol{W}_{gv}$ 分别为语音和视觉门控机制的权重矩阵；$\boldsymbol{b}_a$，$\boldsymbol{b}_v$ 分别为语音和视觉门控机制的偏置向量。

独有特征的融合向量

$$\boldsymbol{F}_p^t=\boldsymbol{T}_p^t+\lambda\boldsymbol{E}^t,\tag{14}$$

式中：λ 为一个超参数。

$$\lambda=\min\left(\frac{\|\boldsymbol{T}_p^t\|_2}{\|\boldsymbol{E}^t\|_2},1\right)。\tag{15}$$

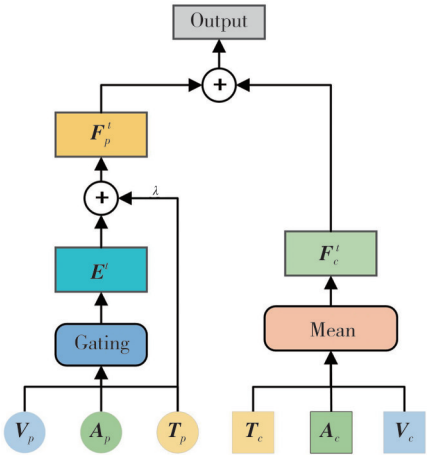

图 4　多模态解耦特征自适应门控融合模块

Fig. 4　Adaptive gated fusion module of multimodal decoupling feature

4）多模态特征融合。以特征均值作为多模态融合的特征 f_m。

$$f_m=Mean(\boldsymbol{h}_v,\boldsymbol{F}_c^t+\boldsymbol{F}_p^t),\tag{16}$$

式中：$\boldsymbol{F}_c^t$ 为共有特征的均值。

$$\boldsymbol{F}_c^t=Mean(\boldsymbol{T}_c^t,\boldsymbol{A}_c^t,\boldsymbol{V}_c^t)。\tag{17}$$

2.4　损失函数

为了训练 AGDF 模型，本文采用目标函数

$$loss=\beta_1 l_c+l_m,\tag{18}$$

式中：l_c 为对比损失；l_m 为多模态预测损失；β_1 为对比学习损失在总损失 $loss$ 中的贡献程度。

多模态预测损失 l_m。利用均方误差计算多模态预测损失，计算公式为

$$l_m=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_m^i-y_m^i)^2,\tag{19}$$

式中：$\hat{y}_m^i$ 为样本 i 的多模态预测结果；y_m^i 为样本 i 的多模态真实结果。

3 实 验

3.1 实验设置

3.1.1 数据集

CMU-MOSI[28]：该数据集包含了来自电影评论、面部表情、语音和文本的多模态数据，用于评估情感倾向的强度。该数据集包含93个电影片段，每个片段长度在1~3 min之间，共有2万多句话。在数据集中，每个片段中的每个句子都被标记为积极、中性或消极情感倾向，并有强度评分。

CMU-MOSEI[29]：该数据集是对CMU-MOSI的改进，包含了更多的样本和更多的演讲者和主题。该数据集包含来自3 227个不同视频的2万多个视频片段。

表1给出了不同的数据集用于训练、验证、测试的统计数据。

表1 数据集的统计数据

Tab. 1 Statistics of the datasets 个

数据集	训练集	验证集	测试集
CMU-MOSI	1 284	229	686
CMU-MOSEI	4 659	1 871	16 326

3.1.2 基线模型

为了全面评估本文提出的模型AGDF，选择多种基线模型进行比较。

1) MULT[30]。利用成对的跨模态注意来学习多模态序列之间的相互作用，并潜在地将流从一个模态调整到另一个模态。

2) SELF_MM[31]。采用自监督学习策略进行标签生成，并实施单模态监督。同时，通过权重调整策略来平衡不同子任务间的学习进度。

3) MMIM[32]。在不同层次上最大化互信息，包括模态间的互信息、多模态融合结果与单模态输入之间的互信息。

4) GraphCAGE[33]。通过将序列数据转换成图结构，避免了数据对齐的要求，克服了循环神经网络中梯度消失或爆炸的问题。

5) ConFEDE[11]。提出一种新的对比损失函数用于学习跨模态一致的共有特征，通过特征分解的方式来增强多模态信息的特征。

6) MVIR[26]。通过学习多视图交互捕捉不同交互状态下的独有和共有信息，增强了多模态情绪表征的表达能力。

3.1.3 实验参数

本文的实验均在配备NVIDIA RTX 3090 GPU的硬件环境中执行。训练模型所采用的超参数设置如表2中所示。其中，optimizer表示选用的优化算法；learning rate表示学习速率的大小；hidden dim、head num和layers分别表示图神经网络中输入的特征向量维度、注意力机制头的数量以及网络的层级数量。

表2 超参数设置

Tab. 2 Hyperparameter settings

超参数	值
optimizer	AdamW
learning rate	0.0005
hidden dim	2701
head num	4
layers	3
β_1	0.1

本文利用网格搜索和经验设置的参数范围来进行多个实验，以寻找潜在的最优参数。learning rate选择{0.000 1, 0.000 5, 0.001}，hidden dim选择{32, 64, 128, 256}，head num的范围为{1－12}，layers的范围为{1－4}，β_1选择{1, 0.5, 0.1, 0.05, 0.01}。

3.1.4 评价指标

遵循先前的工作[11,31]，分别计算了分类和回归的结果以评估模型的性能。在分类任务中，本文报告了多分类的准确性和$F1$评分。准确性可分为二分类精度(Binary Accuracy，Acc-2)和七分类精度(7-class Accuracy，Acc-7)。Acc-2和$F1$评分结果有两种形式：阴性和非阴性二类计算所得结果，阴性和阳性二类计算所得结果[11,34]。在回归任务中，本文报告了平均绝对误差(Mean Absolute Error，MAE)和皮尔逊相关性(Pearson Correlation，Corr)。

3.2 实验结果

本文在数据集CMU-MOSI和CMU-MOSEI上进行了对比实验，对比模型的实验数据来源于文献原文或者其他文献。实验结果详见表3和表4。

表3和表4的实验结果表明：在CMU-MOSI数据集上，AGDF模型在$F1$分数和Acc2上均达到了最佳性能，分别为86.89%和86.75%。类似地，在CMU-MOSEI数据集上，AGDF模型也展现出了最佳性能，$F1$分数和Acc2分别为86.14%和86.35%。

与MULT模型相比，本文模型在所有指标上均表现更好，这说明预训练的方式可以得到更优

的模态特征。与 SELF_MM 模型和 MMIM 模型相比，AGDF 模型在 Acc7 和 MAE 上基本相同，但是 Acc2 和 $F1$ 分数取得了明显的进步，这说明解耦特征能够增强不同模态之间的联系。

表 3　数据集 CMU-MOSI 上的实验结果

Tab. 3　Experimental results in the dataset of CMU-MOSI　%

模型	Acc2 ↑	$F1$ ↑	Acc7 ↑	MAE ↓	Corr ↑
MULT	79.71/80.98	79.63/80.95	36.91	87.99	70.22
SELF_MM	83.44/85.46	83.36/85.43	46.67	70.80	79.63
MMIM	83.67/85.37	83.6/85.37	45.34	75.50	77.30
CONFEDE	84.17/85.52	84.13/85.52	42.27	74.20	78.40
GraphCAGE	—/82.1	—/82.1	35.40	93.30	68.40
MVIR	84.3/85.5	83.9/85.5	—	71.4	79.9
AGDF	**84.48/86.75**	**84.69/86.89**	**46.94**	**69.37**	**80.53**

注：Acc-2 和 $F1$ 中，“/”左侧的值为阴性和非阴性二类计算所得结果，“/”右侧的值为阴性和阳性二类计算所得结果，“↑”表示值越大越好，“↓”表示值越小越好。

表 4　数据集 CMU-MOSEI 上的实验结果

Tab. 4　Experimental results in the dataset of CMU-MOSEI　%

模型	Acc2 ↑	$F1$ ↑	Acc7 ↑	MAE ↓	Corr ↑
MULT	81.15/84.63	81.56/84.52	52.84	55.93	73.31
SELF_MM	83.76/85.15	83.82/84.9	53.87	53.09	76.49
MMIM	82.24/85.97	82.66/85.94	54.24	52.60	77.20
CONFEDE	81.65/85.82	82.17/85.83	54.86	52.20	78.00
GraphCAGE	—/81.7	—/81.8	48.90	60.90	67.00
MVIR	83.9/85.8	84.2/85.6	—	53.1	77.0
AGDF	**84.42/86.35**	**84.49/86.14**	**53.23**	**53.22**	**78.00**

注：Acc-2 和 $F1$ 中，“/”左侧的值为阴性和非阴性二类计算所得结果，“/”右侧的值为阴性和阳性二类计算所得结果。

与 TETFN 模型相比，AGDF 模型通过对比学习更好地保留了模态间的差异性，而不仅仅是依赖单模态的预测分类。与 COnFEDE 模型相比，AGDF 模型在 $F1$ 分数和 Acc2 上均提高了约 1 百分点，这表明直接进行特征拼接来分类并不是最佳策略，而采用图注意力神经网络和门控自适应网络可以对解耦特征进一步融合，可以更有效地利用不同模态之间的独有特征和共有特征，从而使模型的情感信息更加完整。

与采用图神经网络的 GraphCAGE 模型相比，AGDF 模型在所有指标上都取得了最优的结果，并得到了较大的提升，这表明通过图神经网络对解耦特征进行建模，可以在保留一致性信息的同时更好地处理模态独有的信息。与对共有信息和独有信息进行融合的 MVIR 模型相比，通过本文设计的图融合网络能够更好地处理共有信息和独有信息的关系。

这些实验结果共同证明了本文提出的模型在多模态情感分析任务中的有效性和优越性。

3.3　消融实验

为了分析各模块的不同影响，本文在数据集 CMU-MOSI 上进行以下消融实验。

3.3.1　模型结构消融实验

为了验证本文提出的对解耦特征的利用效果，按照表 5 中前 3 列的设置来设计消融实验。

表 5　模块消融实验结果

Tab. 5　Experimental results of module ablation　%

对比学习	GAT	DFAG	Acc2	$F1$	Acc7	MAE	Corr
√			82.56/84.73	82.65/84.76	43.29	74.45	78.47
√	√		83.61/85.54	83.84/85.43	46.65	71.36	79.86
√		√	82.94/85.06	83.09/84.96	**45.92**	72.21	79.40
√	**√**	**√**	**84.48/86.75**	**84.69/86.89**	**46.94**	**69.37**	**80.53**

注：Acc-2 和 $F1$ 中，“/”左侧的值为阴性和非阴性二类计算所得结果，“/”右侧的值为阴性和阳性二类计算所得结果。

表 5 展示了本文提出的模块对模型的影响。实验结果表明：两种解耦特征融合方式都能够有效利用其解耦的特征，使用图注意力网络融合较门控机制融合结果更优，结合两种特征融合方式可以取得最优性能。

为了验证图数据的不同构造方式对实验结果的影响，设计了基准模型 G0 和 G3，并与 2.3 节提出的图数据构造方法（G6）进行比较。G0 模型不采用图结构数据，而是通过 CONCAT 操作将所有特征合并后直接进行分类。G3 模型中图数据采用 3 种模态的共有特征，依据其时间帧的顺序进行构造，如图 5 所示。实验结果如表 6 所示。

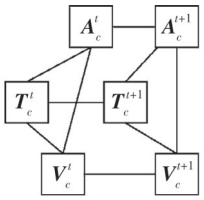

图 5　共有特征关系图

Fig. 5　Relational graph of common features

表 6 实验结果表明：与简单的拼接操作相比，图结构在处理解耦的特征时更为有效；尽管共有特征已捕捉大部分情感信息，但独有特征中包含的情感相关信息对于提升模型性能同样关键。

表6 图数据构造实验结果

Tab. 6 Experimental results of graph data construction %

模型	Acc2	$F1$	Acc7	MAE	Corr
G0	82.94/85.06	83.09/84.96	45.92	72.21	79.40
G3	84.32/86.56	84.4/86.59	46.36	70.01	79.83
G6	**84.48/86.75**	**84.69/86.89**	**46.94**	**69.37**	**80.53**

注：Acc-2和$F1$中，“/”左侧的值为阴性和非阴性二类计算所得结果，“/”右侧的值为阴性和阳性二类计算所得结果。

3.3.2 参数消融实验

为了进一步验证模型参数的影响，本文在数据集CMU-MOSI上进行注意力层数量的影响实验。实验结果如表7所示。

表7 注意力层数实验结果

Tab. 7 Experimental results of the number of attention layers %

层数	Acc2	$F1$	Acc7	MAE	Corr
1	83.67/85.52	83.61/85.48	46.79	71.52	78.92
2	82.94/85.06	83.09/84.96	45.92	72.21	79.40
3	**84.48/86.75**	**84.69/86.89**	46.94	**69.37**	**80.53**
4	83.38/85.37	83.33/85.37	**47.52**	72.28	79.38

注：Acc-2和$F1$中，“/”左侧的值为阴性和非阴性二类计算所得结果，“/”右侧的值为阴性和阳性二类计算所得结果。

表7展示了图注意力层数量对模型性能的影响。实验结果表明：随着注意力层数的增加，模型的性能呈现出先上升后下降的趋势。当注意力层数设置为3时，模型达到了最佳性能。这一现象的成因在于，注意力机制能够针对不同的节点动态地分配权重，从而有效提升了共有信息和独有信息的聚合能力。然而，注意力层数过多时，模型会过度拟合数据中的特定特征，导致其泛化能力降低。

4 结 论

针对现有研究在处理多模态数据时忽视不同模态的共有特征和独有特征联系的问题，本文提出了自适应门控解耦特征融合的多模态情感分析模型。通过对比学习分解特征，并结合自适应门控机制与图注意力网络进行融合，模型能够有效平衡不同模态的独有信息与共有信息。

数据集上的实验结果表明，与其他特征分解方法相比，本文方法的准确率提高了1.23百分点，$F1$分数提高了1.37百分点，Corr提高了2.13百分点，MAE降低了4.83百分点，说明本文模型能够充分利用多模态异质信息，从而有效提升情感分析的性能。

参考文献：

[1] LIU N, ZHAO J. Recommendation system based on deep sentiment analysis and matrix factorization [J]. IEEE Access, 2023, 11: 16994-17001.

[2] PARK J, SEO Y S. Twitter sentiment analysis-based adjustment of cryptocurrency action recommendation model for profit maximization [J]. IEEE Access, 2023, 11: 44828-44841.

[3] WANG J, CHEN Z. SPCM: A machine learning approach for sentiment-based stock recommendation system[J]. IEEE Access, 2024, 12: 14116-14129.

[4] LI Q, GKOUMAS D, LIOMA C, et al. Quantum-inspired multimodal fusion for video sentiment analysis [J]. Information Fusion, 2021, 65: 58-71.

[5] ZHANG Y, SONG D, LI X, et al. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis[J]. Information Fusion, 2020, 62: 14-31.

[6] WANG F, TIAN S, YU L, et al. TEDT: Transformer-based encoding - decoding translation network for multimodal sentiment analysis[J]. Cognitive Computation, 2022, 15(1): 289-303.

[7] JI M Y, ZHOU J W, WEI N. AFR-BERT: Attention-based mechanism feature relevance fusion multimodal sentiment analysis model[J]. PLoS One, 2022, 17(9): 1-20.

[8] 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. New York, NY, USA, 2020: 1122-1131.

[9] VAN AMSTERDAM B, KADKHODAMOHAMMADI A, LUENGO I, et al. ASPnet: Action segmentation with shared-private representation of multiple data sources [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 2384-2393.

[10] LUO Y, WU R, LIU J, et al. Balanced sentimental information via multimodal interaction model[J]. Multimedia Systems, 2024, 30(1): 10.

[11] YANG J, YU Y, NIU D, et al. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 7617-7630.

[12] HUANG J, TAO J, LIU B, et al. Multimodal transformer fusion for continuous emotion recognition[C]// ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 3507-3511.

[13] PRAVEEN R G, GRANGER E, CARDINAL P. Cross attentional audio-visual fusion for dimensional

emotion recognition [C]//2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 2021: 1-8.

[14] FARHOUDI Z, SETAYESHI S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition [J]. Speech Communication, 2021, 127: 92-103.

[15] GKOUMAS D, LI Q, LIOMA C, et al. What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis [J]. Information Fusion, 2021, 66: 184-197.

[16] TANG J, LIU D, JIN X, et al. BAFN: Bi-direction attention based fusion network for multimodal sentiment analysis [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(4): 1966-1978.

[17] WU T, PENG J, ZHANG W, et al. Video sentiment analysis with bimodal information-augmented multi-head attention [J]. Knowledge-Based Systems, 2022, 235: 107676.

[18] LI Z, GUO Q, PAN Y, et al. Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis [J]. Information Fusion, 2023, 99: 101891.

[19] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [DB/OL]. (2020-03-30) [2024-07-05]. http://arxiv.org/abs/2002.05709v2.

[20] CHEN X, HE K. Exploring Simple siamese representation learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 15745-15753.

[21] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.

[22] LI X, SUN A, ZHAO M, et al. Multi-intention oriented contrastive learning for sequential recommendation [C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. New York, NY, USA, 2023: 411-419.

[23] BRODY S, ALON U, YAHAV E. How attentive are graph attention networks? [DB/OL]. (2022-01-31) [2024-07-05]. https://arxiv.org/abs/2105.14491.

[24] YOU J, DU T, LESKOVEC J. ROLAND: Graph learning framework for dynamic graphs [DB/OL]. (2022-08-15) [2024-07-05]. https://arxiv.org/abs/2208.07239.

[25] LI K, HUANG Z, JIA Z. RAHG: A role-aware hypergraph neural network for node classification in graphs [J]. IEEE Transactions on Network Science and Engineering, 2023, 10(4): 2098-2108.

[26] TANG Z, XIAO Q, QIN Y, et al. Multi-view interactive representations for multimodal sentiment analysis [J]. IEEE Transactions on Consumer Electronics, 2024, 70(1): 4095-4107.

[27] RAHMAN W, HASAN M K, LEE S, et al. Integrating Multimodal Information in Large Pretrained Transformers [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 2359-2369.

[28] ZADEH A, ZELLERS R, PINCUS E, et al. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos [DB/OL]. (2016-08-12) [2024-07-05]. https://arxiv.org/abs/1606.06259.

[29] 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. Melbourne, Australia, 2018: 2236-2246.

[30] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 6558-6569.

[31] YU W, XU H, YUAN Z, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis [DB/OL]. (2021-02-09) [2024-07-05]. https://arxiv.org/abs/2102.04830.

[32] HAN W, CHEN H, PORIA S. 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.

[33] WU J, MAI S, HU H. Graph capsule aggregation for unaligned multimodal sequences [C]//Proceedings of the 2021 International Conference on Multimodal Interaction. New York, NY, USA, 2021: 521-529.

[34] XU M, LIANG F, SU X, et al. CMJRT: Cross-modal joint representation transformer for multimodal sentiment analysis [J]. IEEE Access, 2022, 10: 131671-131679.