非语言信息增强和对比学习的多模态情感分析模型

刘 佳[①②③④] 宋 泓[①②] 陈大鹏*[①②③④] 王 斌[①②] 张增伟[①②]

[①](南京信息工程大学 天长研究院 滁州 239356)

[②](南京信息工程大学 自动化学院 南京 210044)

[③](江苏省智能气象探测机器人工程研究中心 南京 210044)

[④](江苏省大气环境与装备技术协同创新中心 南京 210044)

摘 要：因具有突出的表征和融合能力，深度学习方法近年来越来越多地被应用于多模态情感分析领域。已有的研究大多利用文字、面部表情、语音语调等多模态信息对人物的情绪进行分析，并主要使用复杂的融合方法。然而，现有模型在长时间序列中未充分考虑情感的动态变化，导致情感分析性能不佳。针对这一问题，该文提出非语言信息增强和对比学习的多模态情感分析网络模型。首先，使用长程文本信息去促使模型学习音频和视频在长时间序列中的动态变化，然后，通过门控机制消除模态间的冗余信息和语义歧义。最后，使用对比学习加强模态间的交互，提升模型的泛化性。实验结果表明，在数据集CMU-MOSI上，该模型将皮尔逊相关系数(Corr)和F1值分别提高了3.7%和2.1%；而在数据集CMU-MOSEI上，该模型将“Corr”和“F1值”分别提高了1.4%和1.1%。因此，该文提出的模型可以有效利用模态间的交互信息，并去除信息冗余。

关键词：多模态情感分析；多模态融合；信息增强；MLP

中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2024)00-0001-10

DOI: 10.11999/JEIT231274

A Multimodal Sentiment Analysis Model Enhanced with Non-verbal Information and Contrastive Learning

LIU Jia[①②③④] SONG Hong[①②] CHEN Da-Peng[①②③④]

WANG Bin[①②] ZHANG Zeng-Wei[①②]

[①](*Tianchang Research Institute*, *Nanjing University of Information Science & Technology*, *Chuzhou* 239300, *China*)

[②](*School of Automation*, *Nanjing University of Information Science & Technology*, *Nanjing* 210044, *China*)

[③](*Jiangsu Province Engineering Research Center of Intelligent Meteorological Exploration Robot*, *Nanjing* 210044, *China*)

[④](*Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology*, *Nanjing* 210044, *China*)

Abstract: Deep learning methods have gained popularity in multimodal sentiment analysis due to their impressive representation and fusion capabilities in recent years. Existing studies often analyze the emotions of individuals using multimodal information such as text, facial expressions, and speech intonation, primarily employing complex fusion methods. However, existing models inadequately consider the dynamic changes in emotions over long time sequences, resulting in suboptimal performance in sentiment analysis. In response to this issue, a Multimodal Sentiment Analysis Model Enhanced with Non-verbal Information and Contrastive Learning is proposed in this paper. Firstly, the paper employs long-term textual information to enable the

收稿日期：2023-11-17；改回日期：2024-03-24；网络出版：2024-04-07

*通信作者： 陈大鹏 dpchen@nuist.edu.cn

基金项目：国家自然科学基金(61773219, 62003169)，江苏产业前瞻与关键技术重点项目(BE2020006-2)，江苏省自然科学基金青年基金(BK20200823)

Foundation Items: The National Natural Science Foundation of China (61773219, 62003169), The Key R&D Program of Jiangsu Province (Industry Prospects and Key Core Technologies) (BE2020006-2), The Natural Science Foundation of Jiangsu Province (BK20200823)

model to learn dynamic changes in audio and video across extended time sequences. Subsequently, a gating mechanism is employed to eliminate redundant information and semantic ambiguity between modalities. Finally, contrastive learning is applied to strengthen the interaction between modalities, enhancing the model's generalization. Experimental results demonstrate that on the CMU-MOSI dataset, the model improves the Pearson Correlation coefficient (Corr) and F1 score by 3.7% and 2.1%, respectively. On the CMU-MOSEI dataset, the model increases "Corr" and "F1 score" by 1.4% and 1.1%, respectively. Therefore, the proposed model effectively utilizes intermodal interaction information while eliminating information redundancy.

Key words: Multimodal emotion analysis; Multimodal fusion; Information enhancement; MLP

1 引言

近年来，随着互联网的发展，人们在社交平台上发表言论的方式逐渐多元化，主要包含文本(t)、音频(a)、视频(v)等形式。如何通过多模态的信息进行情感分析受到越来越多的关注。例如，多模态情感分析(Multimodal Sentiment Analysis, MSA)可以协助医疗人员通过病患的情绪变化诊断病情[1]，还可以在用户购物时根据用户的情绪反馈进行个性化的商品推荐[2]，因此其具有很大的社会价值。

与传统的单模态情绪分析任务不同，MSA通过融合两个或多个模态信息进行情感分析。多模态情感分析的主要关注点之一是整合来自多种模态的序列情感数据。先前研究人员[3–7]的关注点主要集中在开发具有复杂融合机制的模型，用于加强不同模态间情感信息的交互。最近，一些基于Transformer[8,9]和大型预训练模型[10,11]的方法被提出，用于执行模态特征之间的交互特征提取。这些方法已被证明在性能上优于卷积神经网络等传统框架。然而，在多模态研究中仍存在 两个具有挑战性的问题。一是如何有效地处理长时间序列以捕获模态中的重要上下文信息。二是如何准确提取模态内和模态间的特征，并有效去除特征提取过程中引入的噪声。

人们情绪通常以长时间序列的形式呈现。在第1个挑战中，为了捕获目标人物在长时间序列中的情感变化，先前的研究广泛使用了递归神经网络(Recurrent Neural Network, RNN)等传统结构。然而在处理过长的序列时，模型仍会出现梯度消失或梯度爆炸问题。

相对于第1个挑战，第2个挑战受到的关注较少，这是因为先前的研究通常独立地从各个模态中提取特征，并且在特征融合之前不共享信息，这限制了多模态特征的交互学习能力。此外，不同的模态之间也可能存在情感不一致等问题。如图1所示(“+”代表积极情绪，“？”代表因模态间的语义冲突而无法获知准确的情绪特征，“-”代表消极情绪，“--”代表情感强度更大的消极情绪)，当人们使用反语时，其语言内容与音频和视频情感信号可能出现不一致性。例如，一个人可能口头表达“我很好”(这通常被认为是积极的情感语境)，但低沉的音调与忧郁的面部表情可能更准确地反映出

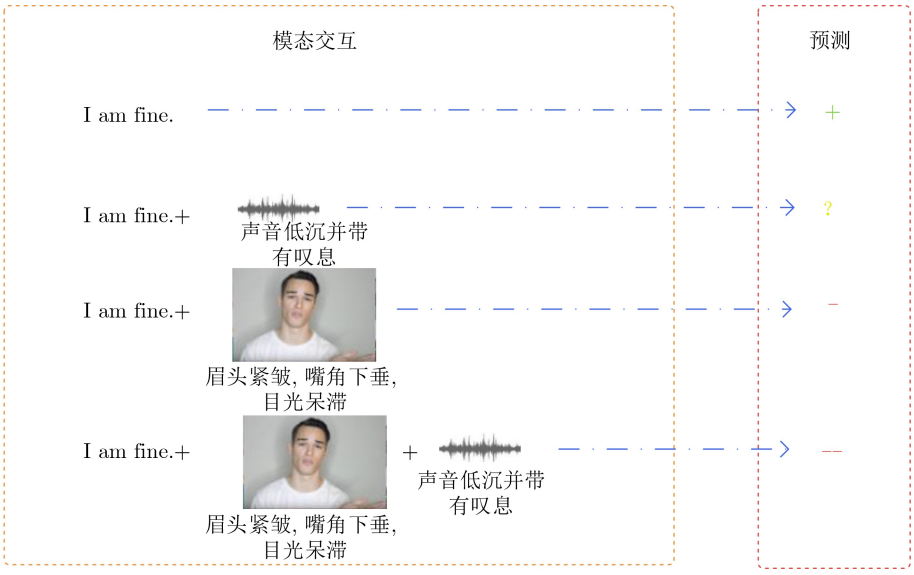


图 1 不同模态组合的情绪预测

该个体实际上经历的是负面的情绪状态。因此，提取模态间有效的特征并过滤掉无效的歧义以避免不一致，这是一个非常具有挑战性的任务。

为应对上述挑战，本文提出了一种简单而有效的多模态情绪识别模型，称为“非语言信息增强和对比学习的多模态情感分析模型”(A Multimodal Sentiment Analysis Model Enhanced with Non-verbal Information and Contrastive Learning)。可以有效的利用多种模态的互补信息，以实现更加稳健和准确的情感分析。

本文的主要贡献如下：

(1)本文提出了一种新的非语言信息增强方法，旨在捕捉不同模态情感之间的差异性，以进一步增强目标模态的情感表示。

(2)考虑到非语言数据的不稳定性，本文采用多模态对比学习使得模型可以提取更加细粒度的信息，并通过MLP融合显著降低计算负担。

(3)本文在两个公开可用的多模态数据集CMU-MOSI和CMU-MOSEI上进行实验。证明了本文提出的模型能够更好地关注模态互补信息和相似信息。与先前的模型相比，该模型在处理人物在长时间序列中情感变化和情感歧义问题上具有显著的优势。

2　相关工作

根据是否使用预训练的语言模型，已有的MSA工作可以分为以下两类。

2.1　不使用预训练模型

不使用预训练模型的网络通常使用WordVec2进行文本特征的语义学习。Cao等人[12]使用文本和图像结果进行加权获得结果。这种方法属于后期融合，无法有效利用模态信息进行互补。为了解决这一问题，情感分析领域的后续工作主要集中在设计复杂的融合网络上。例如，张量融合网络(Tensor Fusion Network, TFN)[13]以端到端方式对模态内和模态间动力学建模。Tsai等人[15]通过跨模态注意力机制实现了不同模态之间的信息映射和交互。韩虎等人[16]则通过双向长短期记忆网络和注意力机制对上下文信息建模，但这同时也带来计算负担过重等问题。为了解决该类问题，Sun等人[17]采用完全基于MLP的架构用来融合三模态间的信息，在保持性能的同时大幅度降低了计算机的计算负担。

2.2　使用预训练模型

第2种方法是利用预先训练好的语言模型提取文本特征。相较于第1种方法，这种方式虽然可以获得更好的文本特征，但是同时也增加了模型的计算负担。Hazarika等人[18]将模态映射入两个不同的特征子空间，然后将表示输入Transformer结构中进行融合来预测情绪。蔡宇扬等人[19]使用BERT (Bidirectional Encoder Representation from Transformers)[20]提取特征，随后通过细粒度注意力机制提取模态间的局部关联特征，弱化非情感特征的影响。Wang等人[21]通过特征转换策略将非言语模态信息转换为文本信息，以减少模态间的差异性。然而这种方法的结果依赖于文本特征信息的挖掘，当用作提取文本语义信息的模型性能不佳或模态间存在语义歧义等问题时，该方法很难做出准确的判断。

3　多模态情感分析模型

为了更好的捕捉人物在长时间序列中情感变化并消除情感歧义，本文利用跨模态注意力机制来增强音频和视频的特征表示。随后，将增强后的信息送入自注意力机制中，以捕获人物在长时间序列中的情感倾向。最后，模型使用一个门控机制来判断加强后的特征是否存在噪声，并以此选择使用原始的音频和视频特征，或经过文本增强的特征。为进一步加强模态间的交互性，模型引入了对比学习策略。在多模态特征融合阶段，本文采用了基于MLP的融合框架，实现了在时序、模态及通道3个维度上的混合融合。这不仅保持了模型的高性能，还显著降低了计算复杂度。模型结构图如图2所示。

3.1　特征提取

在进行模态信息增强之前，本文采用LSTM[22]来提取音频和视频特征。而对于文本特征，模型使用多模态情感数据集CMU-MOSEI[23]对BERT模型进行了微调。得到的三模态特征表示为$X_{s\in(\mathrm{t,a,v})}$，X_{t}表示文本模态信息，X_{a}表示听觉模态信息，X_{v}表示视觉模态信息。其中，$t\in\boldsymbol{R}^{L_t\times d_{\mathrm{t}}}$，$a\in\boldsymbol{R}^{L_a\times d_{\mathrm{a}}}$，$v\in\boldsymbol{R}^{L_v\times d_v}$，$L_{s\in(\mathrm{t,a,v})}$表示数据序列的长度，$d_{s\in(\mathrm{t,a,v})}$代表各自特征的维数。

3.2　非语言信息增强模块

非语言信息增强模块(Non-verbal Information Enhancement Module, NE)通过跨模态注意力机制产生基于视觉信息和听觉信息的语言嵌入。随后，使用自注意力机制动态对长序列中的元素施加不同的注意力权重。最后，模块引用门控机制用以消除因模态增强而产生的模态间噪声。

为赋予序列时间信息，本文采用如式(1)所示方式将位置信息嵌入到序列$s\in(a,v)$中，这一操作有助于非语言特征在时间上保持有序。

$$\boldsymbol{H}(s)=X_s+PE(T_s,d) \tag{1}$$

文献[24]的研究表明，在多模态情感分析中，文本模态包含丰富的情感线索，但同时也可能掩盖

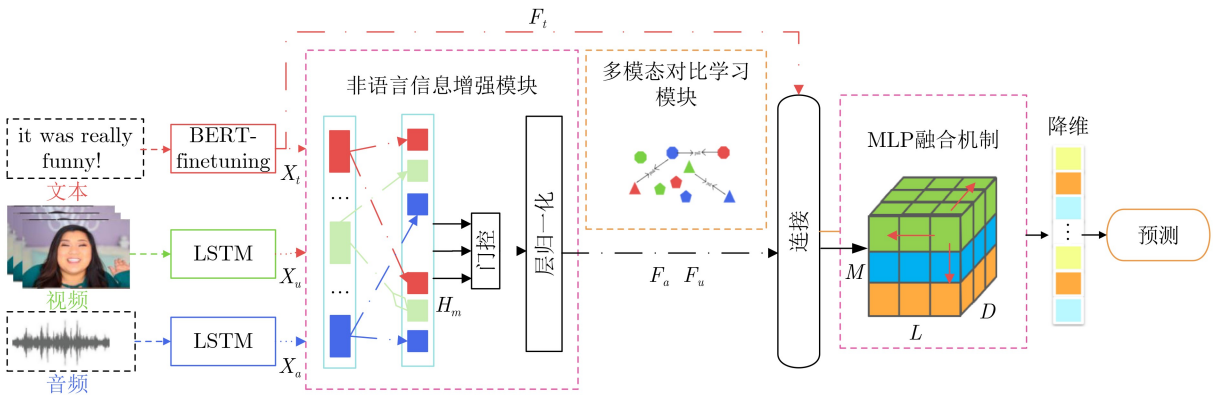


图 2 模型结构图

个体真实情感的表达。受这一观点的启发，本文采用跨模态注意力机制将文本信息用于增强音频和视频信息。随后，对增强后的信息使用自注意力机制[25](Self-Attention)，利用自注意力机制捕捉人物在长序列中的情感倾向。此举有助于提高模型对情感变化迟缓的人物情感分析的准确性。最后，根据增强后的特征与原始特征的相似性来判断跨模态增强的有效性。如果增强后的特征被认定为有效，模型使用增强后的特征。否则，模型将增强后的特征视为噪声，并使用未增强特征，如图3所示。

模型通过两种注意力机制得到含有文本情感特征线索的音频特征$\boldsymbol{H}_{\mathrm{a}}^{*}$和视频特征$\boldsymbol{H}_{\mathrm{v}}^{*}$。随后，本文使用式(2)中的计算方式来量化模态增强后特征与原始特征之间的相似度sim。模型通过sim来判定是否采用增强后的特征。

$$\begin{aligned}\mathrm{sim}&=\cos(\theta)=\frac{\boldsymbol{H}_s^{*}\cdot\boldsymbol{H}_s}{\|\boldsymbol{H}_s^{*}\|\,\|\boldsymbol{H}_s\|}\\&=\frac{\sum_{i=1}^{n}\boldsymbol{H}_{s_i}^{*}\times\boldsymbol{H}_{s_i}}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{H}_{s_i}^{*})^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{H}_{s_i})^2}}\end{aligned}\tag{2}$$

在上述公式中，$\cos(\theta)$表示余弦距离；θ表示向量$\boldsymbol{H}_{s_i}^{*}$和$\boldsymbol{H}_{s_i}$间夹角；其中$s\in(a,v)$。$\|\cdot\|$表示向量的模；$\boldsymbol{H}_s$表示经Bert和LSTM提取的特征；n表示向量中的元素数量，表示特征的第i个分量。

为了解决模态特征不匹配所导致的噪声问题，本文通过计算向量之间的距离评估增强特征与原始特征的相似度。向量间的距离越接近于1，则表明非语言特征的情感极性在增强后未发生变化，这意味着文本模态成功增强了非语言模态特征，使得音频和视频特征获得了更多的上下文信息。相反，向量间的距离越接近于–1，则说经增强后的非语言特征情感极性发生了改变，表明出了情感歧义等问题，应当去除。

在非语言信息增强模块中，模块设定了阈值limit(limit是一个超参数)。当余弦距离大于我们预先设置的limit时，选择增强后的非语言特征。反之，使用原始特征。

3.3 多模态融合和预测

模型将增强后的音频特征和视频特征F_s与文本特征X_{t}在第二个维度上进行扩展，形成维度为$\boldsymbol{R}^{L\times1\times D}$。随后，模型在这3个模态特征的第2个维度上进行拼接，组合成多模态特征$U\in\boldsymbol{R}^{L\times M\times D}$。

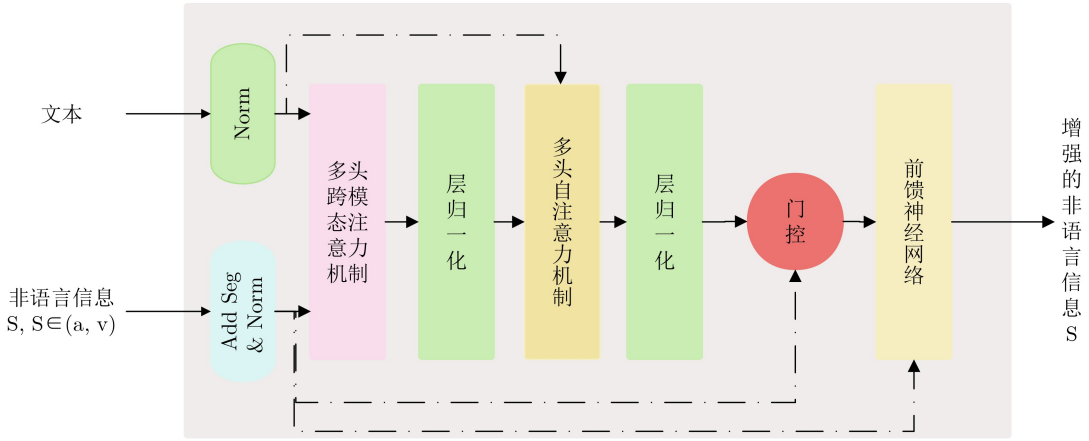


图 3 非语言信息增强机制结构图

然后将组合的特征$\boldsymbol{U}$传递给堆叠的MLP层进行融合，如图1所示，MLP融合模块由3个MLP单元组成，每个单元在各自的维度上进行混合融合。每个MLP单元由3个线性层和1个非线性激活函数组成，结构如图4所示。

线性层公式为

$$\mathrm{MF_t} = \boldsymbol{W_t}\boldsymbol{U}_{*\times \mathrm{m}\times \mathrm{d}} + \boldsymbol{B_t} \tag{3}$$

其中，$\boldsymbol{W_t} \in \boldsymbol{R}^{L\times L'}$为可学习矩阵，$L'$是$L$轴上的降维，是一个超参数。式(4)表示模态在$L$轴上融合，将$L$轴替换为$M$或$D$则表示模态在$M$轴或$D$轴上融合。

$$\begin{aligned}\boldsymbol{U}^*_{*\times m\times d} = &\mathrm{LN}(\mathrm{MF}_t\left(\sigma\left(\mathrm{MF}_t\left(\mathrm{MF}_t\left(\boldsymbol{U}_{*\times m\times d}\right)\right)\right)\right)\\ &+\boldsymbol{U}_{*\times m\times d}),\\ &(m,d)\in\{(1,1),(1,2),(2,1),\cdots,(M,D)\}\end{aligned} \tag{4}$$

最终，得到多模态特征$\boldsymbol{U}^* \in \boldsymbol{R}^{L'\times M'\times D'}$。根据先前的研究[17,18]，模型将这些融合特征扁平化为$\boldsymbol{U}^{L'M'D'}$，然后将扁平化的特征传送到分类器中，以进行情感分析。

本文使用交叉熵损失(Binary CrossEntropy)和均方差损失(MeanSquaredErrorLoss)进行优化。计算方式为

$$L_{\mathrm{task}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \ln \widehat{y}_i \tag{5}$$

$$L_{\mathrm{task}} = \frac{1}{N}\sum_{i=1}^{N} \|y_i - \widehat{y}_i\|_2^2 \tag{6}$$

其中，N代表样本数量，y和$\hat{y}$分别表示样本标签和预测结果。

交叉熵损失函数和均方差损失函数度量预测结果与真实值之间的误差。交叉熵损失函数对错误的分类有较大的误差，均方差损失函数对于较小的误差有较大的惩罚，这有助于模型更准确的捕捉和预测细微的情感差异。因此，交叉熵损失函数和均方差损失函数经常被用来优化多模态情感析模型。

3.4 模态间的对比学习

为了将对比学习应用于多模态情感分析，Hycon[26]将类似于监督的对比损失学习与多模态嵌入相互结合，有效地利用了标签信息。与Hycon不同的是，本文的多模态对比学习策略旨在通过探索模态间的深层交互信息。具体来说，本文在对比学习策略中引入了正样本对和负样本对的概念。正样本对由来自同一标签下不同样本、不同模态的单峰表示组成，负样本对由两个不同标签下不同样本、不同模态的单峰表示组成。这样设计正负样本对有助于模型更好的理解和利用不同模态之间的信息。

对于一个大小为K的batch_size，每个锚点的每个模态s生成一个集合U。其中包含N个正样本对和M个负样本对。批量大小K是固定的，但正负样本对的数量是随机的(即N和M的数量是随机的)。对得到的特征表示执行L2-normalization，以确保每对样本的相似性在0到1之间。模态间对比损失函数表达式如式(7)所示。

$$L_{\mathrm{inter}} = -E_s\left\{\frac{\sum_{i=1}^{N}(a^s)^T p_i}{\sum_{i=1}^{N}(a^s)^T p_i + \sum_{j=1}^{M}(a^s)^T n_j}\right\} \tag{7}$$

式中，$s\in(t,a,v)$，a^s代表锚点，p_i和n_j分别代表正样本对和负样本对。模型的损失等于预测损失加上对比学习损失，如式(8)所示。

$$L_{\mathrm{all}} = L_{\mathrm{task}} + \alpha L_{\mathrm{inter}} \tag{8}$$

式中，L_{all}表示整体的损失函数，L_{task}表示模型的预测损失，α是一个超参数，用于调整两种损失之间的权重。通过调整α的值，以实现更好的性能。

4　实验设置与结果分析

4.1 实验数据集

本文分别在两个公开的多模态数据集CMU-MOSI[27]和CMU-MOSEI[23]上进行情感分析实验，来验证 本文所提出模型的有效性。

CMU-MOSI[27]是一个多模态数据集，包括文本、视觉和声学模态。它来源于93个YouTube电影评论视频。这些视频被剪辑成2 199个片段。每个片段都用[–3,3]范围内的情感强度进行注释。其中，

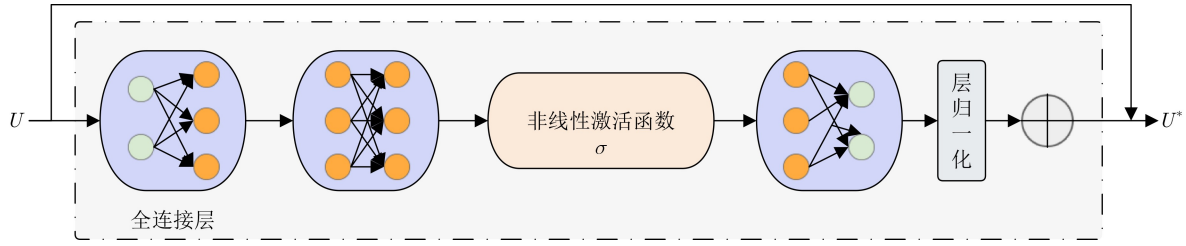


图 4 MLP融合单元

情感标签的正负极性代表了人物情感的积极倾向和人物情感的消极倾向，数值的大小反映了人物情感的剧烈程度。数据集分为三个部分，即训练集(1284段)、验证集(229段)和测试集(686段)。

CMU-MOSEI[23]具有和CMU-MOSI相同的注释，是目前规模最大的视频多模态情感分析数据集。该数据集有16 315个话语用于训练，1817个话语用于验证，4654个话语用于测试。

4.2 实验设置和评价标准

在经过数据预处理之后，模型得到了以下特征维度：文本特征维度d_t为768，音频特征维度d_a为74，CMU-MOSI和CMU-MOSEI数据集的视频特征维度d_v分别为47，35。

本文使用 PyTorch来创建模型。经过多次实验，本文将门控机制的阈值设置为0.75，超参数α设置为0.1时，模型能够获得最佳效果。在训练过程中，本文使用了Adam优化器来优化模型参数，学习率在CMU-MOSI和CMU-MOSEI数据集上分别设置为0.004和0.001，每迭代50轮，将学习率乘以0.1。模型使用128作为batch_size大小，进行120次迭代，并且设置了提前终止训练以防止过拟合。所有的实验都在两张NVIDIA RTX 3 090 GPU上进行。

根据先前的研究[28–30]将多模态情感分析任务分为分类任务和回归任务。本文使用以下指标评估模型的性能：2分类准确率(Acc-2)、F1-Score、 7分类准确率(Acc-7)、平均绝对误差(MAE)和皮尔逊相关性(Pearson correlation, Corr)，用于测量模型输出与人工标注的实际值之间的相关性。

4.3 模型对比

为了证明本文提出的模型的有效性，该模型将与现有的一些模型比较。

TFN[13]：利用笛卡尔积计算各模态的张量，获取单模态，双模态，三模态的交互信息。

LMF[14]：是对张量融合网络(TFN)的改进，利用低秩张量融合网络降低了计算负担。

MULT[28]：采用6个跨模态注意力机制，两两模态间进行模态交互，充分提取3模态间的交互信息。

ICCN[31]：使用文本作为主要模态，得到包含语言信息的非语言特征。然后对得到的特征使用相关性分析，以促进信息交互。

MISA[18]：将模态映射入两个不同的特征子空间，学习模态内情感不变特征和模态间的共享特征。

MAG-BERT[32]：基于BERT和XLNet模型创建多模态适应门结构，使得模型可以接受非语言信息。

MGHF[33]：利用门控循环网络，实现不同模态间信息的充分交互，并通过门控机制有效的剔除了在模态交互过程中产生的冗余信息和不相关信息。

MHMF-BERT[34]：设计BERT外部分层融合结构，将中间层信息与非语言模态分多阶段进行融合。

CENet[21]：通过特征转换缩小非语言信和文本信息的差距，然后利用非语言信息去增强文本信息。

4.4 实验结果与分析

本文提出模型与现有模型[13,14,18,21,28,31–34]进行比较，为了保证实验的公平性，本文将在相同的实验环境下复现所要对比的模型。对比实验结果如表1所示。

由表1可以看出，本文提出模型相比对比模型在各个性能指标上都取得了更为显著的提升。与一些具有复杂融合策略的模型(如TFN和LMF)相比，本文提出的模型在各项指标上都拥有更高的准确率，说明模型在使用MLP融合降低了计算复杂度的同时并未损失太多的性能。与MULT相比，本文

表 1 模型对比实验结果CMU-MOSI和CMU-MOSEI

模型	CMU-MOSI					CMU-MOSEI				
	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1
TFN	0.901	0.698	34.90	80.80	80.70	0.593	0.700	50.20	82.50	82.10
LMF	0.917	0.695	33.20	82.50	82.40	0.623	0.677	48.00	82.00	82.10
MULT	0.918	0.680	36.47	79.30	79.34	0.580	0.703	51.80	82.50	82.30
ICCN	0.860	0.710	39.00	83.00	83.00	0.565	0.713	51.60	84.20	84.20
MISA	0.783	0.761	42.30	83.40	83.60	0.555	0.756	52.20	85.50	85.30
MAG-BERT	0.713	0.789	—	84.30	84.30	0.539	0.753	—	85.23	85.08
MGHF	0.709	0.802	45.19	85.21	85.21	0.528	0.767	**53.70**	85.30	84.86
MHMF-BERT	0.701	0.787	—	85.30	85.30	0.519	0.761	—	85.60	85.60
CENet(B)	0.698	0.806	—	86.74	86.66	0.515	0.816	—	86.24	86.16
Ours	**0.633**	**0.843**	**48.10**	**88.73**	**88.80**	**0.482**	**0.830**	52.80	**87.30**	**87.30**

提出的模型在提取模态信息交互部分增加了自注意力机制和门控机制的使用。该模型在各项指标上均有提升，说明增加注意力机制可以提取模态的显著特征，门控机制可以去除因模态交互产生的信息冗余。另外，相比于ICCN和MISA，模型不但采用文本信息去增强非语言信息，还在模态融合前提取了模态间的交互信息，从而在降低计算复杂度的同时考虑了模态在情感分析中的贡献比。对比学习通过对正负样本的学习，使得模型能够理解不同模态间信息表达的差异，有助于模型对模态交互信息的提取。相比于MAG-BERT和MGHF，模型不仅可以有效去除模态交互产生的信息冗余，还在序列、通道和模态3个轴上进行融合，说明MLP融合使模型能更好地理解融合特征以及分析出情感对象的情感特征之间的关系。

与最新的模型比较，本文所提出的模型在CMU-MOSI数据集上皮尔逊相关系数(Corr)提高了3.7%，在F1-Score和准确率(Acc)上也有着超过2%的提升。模型在CMU-MOSI数据集上实现了0.633的MAE，在CMU-MOSEI数据集上实现了0.482的MAE。实验结果均表明模型在捕捉人物在长时间序列中情感变化和处理具有情感歧义问题时具有显著的优势。本文对评价指标进行了可视化，有助于展示模型的执行情况，如图5所示。

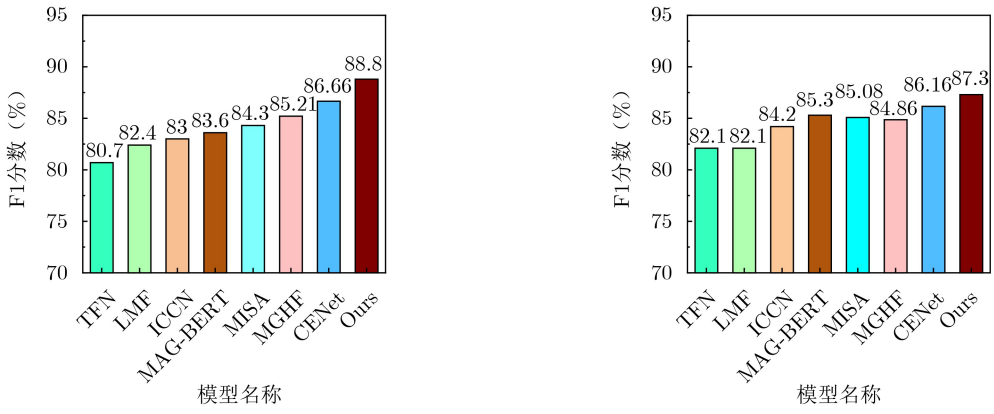
模型在Acc-7上的表现差于 最新的模型，可能是由于MLP融合模块降低了模型的计算负担，导致模型无法获得含有更加细粒度情感交互线索的语义向量，这使得模型难以准确分析人物情感变化的强度，进而影响了整体效果。

4.5 消融实验

为了评估本文所提出的模型，本文设置了消融实验。以研究每种模态和每个模块对的表现。

4.5.1 非语言信息增强模块有效性实验

如表2所式，本文在第二组实验中，加入了非语言信息增强模块(NE)，并分别将音频、视频和文本作为辅助模态增强其他两个模态信息，旨在比较不同模态的组合方式在MSA任务中的性能表现。如表2所示，选择文本作为辅助模态用于增强音频和视频信息的模型，在CMU-MOSI和CMU-MOSEI数据集上的每个性能指标均达到了最佳水平。这是因为文本模态增强了声学和视觉信息。在此基础上，模型删除了NE模块的门控结构，观察到在两个数据集上每个性能指标都出现了明显的下降，由此可见使用文本特征增强音频和视频模态信息中产



(a) CMU-MOSI上F1分数　　(b) CMU-MOSEI上F1分数

图 5 各基准模型性能比较

表 2　CMU-MOSI和CMU-MOSEI数据集上消融实验研究结果(-G代表移除了NE模块中的门控机制)

模型	CMU-MOSI			CMU-MOSEI		
	MAE(↓)	Corr(↑)	F1(↑)	MAE(↓)	Corr(↑)	F1(↑)
Base	0.702	0.810	85.32	0.550	0.767	84.92
Base + $NE^{(A)}$	0.688	0.812	85.52	0.552	0.772	85.01
Base + $NE^{(V)}$	0.674	0.823	85.61	0.542	0.776	85.53
Base + $NE^{(T)}$	0.654	0.833	87.06	0.525	0.793	86.03
Base + $NE^{(T)}$-G	0.665	0.825	86.60	0.535	0.787	85.60
Base + contrast	0.673	0.824	86.63	0.539	0.790	85.66
Ours	**0.633**	**0.843**	**88.88**	**0.482**	**0.830**	**87.30**

生了信息冗余或情感歧义，而门控机制则有助于去除这种信息冗余，消除情感歧义并保留听觉模态和视觉模态的独立性。

本文从CMU-MOSI测试集中随机选择这部分实验的10个样本，并对部分实验指标得分进行可视化，如图6所示。

4.5.2 对比学习有效性实验

为了全面探索本文提出的对比学习方法在多模态情感分析任务中的有效性。本文分别考虑了两种情境：一是没有采用对比学习策略，另一种则是集成了对比学习策略。

在这些评价指标中，除了均方误差(MAE)之外，得分越高意味着模型表现越好。如表2所示，与未采用对比学习策略的模型相比，集成了对比学习的模型展现出了明显优越的性能。这一实验结果证明了对比学习确实可以有效地促进多模态数据间的深度交互和整合，进一步增强模型的鲁棒性和泛化能力。

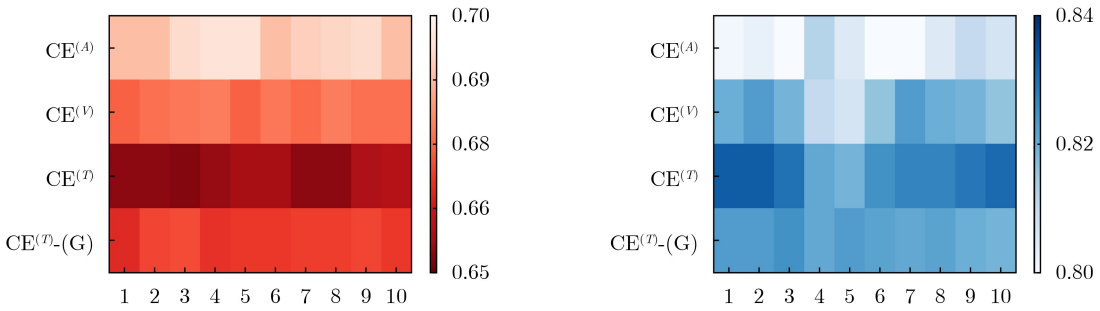
4.6 定量分析

为了阐明模型的工作原理，本文展示了在有无NE模块和对比学习策略嵌入的情境下，模型对情绪强度的分析情况。表3展示本文模型通过整合非语言信息来调整情绪强度的一些例子。

在前两个示例中，模型未使用NE模块和对比学习策略的模态信息进行情感极性的预测，模型预测的情感强度不足。然而，当加入NE模块和对比学习策略后，预测值接近真实情感强度。对于第三个例子，未使用NE模块增强和对比学习策略的模型得出的情感极性是有歧义的。在这种情境下，非语言信息可以帮助模型确定情绪的极性。这些观察结果表明，模型可以有效的结合非语言信息和语言信息来提高情绪预测的准确性。

5 结束语

本文提出了一种非语言信息增强和对比学习的多模态网络模型。该模型利用预训练模型BERT和LSTM来提取模态特征，使用NE模块通过长程文本信息来增强非语言信息的表征能力，并引入了对比学习策略。最后，将特征在时序、通道、模态3个轴上进行融合。与其他算法相比，本文提出的NE模块考虑到了非语言特征长依赖性，并采用门控机制删除了因模态交互产生的信息冗余，提供了更好的语义向量。本文在模型融合前采用对比学习，考虑了每个输入特征之间的相似程度，提高了模型对模态特征的情感线索的理解，并提升了模型的泛化性，从而进一步解决了模态间情感倾向不匹配等问题。同时，MLP模块减少了模态融合的计



(a) MAE值 (b) Corr值

图 6 各辅助模态性能可视化

表 3 (A代表本文提出的模型，B代表去除非语言信息增强模块和对比学习策略的模型)

	示例	标签	A	B
T	It's been a great day.			
A	正常、平和的语气	2.53	2.52	2.05
V	微笑			
T	He was the only character that slightly interesting.			
A	迟疑的语气	–0.80	–0.82	–0.10
V	摇头、眉头紧皱			
T	I give Shrek Forever After directed by Mike Mitchell a grade of B minus.			
A	正常的语气	1.00	0.95	–0.98
V	面无表情			

算负担。实验结果表明，该模型在基准数据集CMU-MOSI和CMU-MOSEI上的性能优于现有技术。

由于实际应用中可能存在数据缺失以及不同模态情感线索的重要程度不一致等问题，在未来工作中，我们希望设计一个模态自适应机制。该机制能够根据不同模态中情感线索的重要程度动态选择合适的辅助模态，以便通过获得更精准的模态交互表示来更好地适应目标任务。同时，我们也将重点关注对情感强度的分析，使识别的情绪状态更加明确。

参 考 文 献

[1] 李霞, 卢官明, 闫静杰, 等. 多模态维度情感预测综述[J]. 自动化学报, 2018, 44(12): 2142–2159. doi: 10.16383/j.aas.2018.c170644.
LI Xia, LU Guanming, YAN Jingjie, *et al*. A survey of dimensional emotion prediction by multimodal cues[J]. *Acta Automatica Sinica*, 2018, 44(12): 2142–2159. doi: 10.16383/j.aas.2018.c170644.

[2] 丁永刚, 李石君, 付星, 等. 面向时序感知的多类别商品方面情感分析推荐模型[J]. 电子与信息学报, 2018, 40(6): 1453–1460. doi: 10.11999/JEIT170938.
DING Yonggang, LI Shijun, FU Xing, *et al*. Temporal-aware multi-category products recommendation model based on aspect-level sentiment analysis[J]. *Journal of Electronics & Information Technology*, 2018, 40(6): 1453–1460. doi: 10.11999/JEIT170938.

[3] 李紫荆, 陈宁. 基于图神经网络多模态融合的语音情感识别模型[J]. 计算机应用研究, 2023, 40(8): 2286–2291,2310. doi: 10.19734/j.issn.1001-3695.2023.01.0002.
LI Zijing and CHEN Ning. Speech emotion recognition based on multi-modal fusion of graph neural network[J]. *Application Research of Computers*, 2023, 40(8): 2286–2291,2310. doi: 10.19734/j.issn.1001-3695.2023.01.0002.

[4] ZHENG Jiahao, ZHANG Sen, WANG Zilu, *et al*. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition[J]. *IEEE Transactions on Multimedia*, 2023, 25: 2213–2225. doi: 10.1109/TMM.2022.3144885.

[5] FU Yahui, OKADA S, WANG Longbiao, *et al*. Context- and knowledge-aware graph convolutional network for multimodal emotion recognition[J]. *IEEE MultiMedia*, 2022, 29(3): 91–100. doi: 10.1109/MMUL.2022.3173430.

[6] NGUYEN D, NGUYEN D T, ZENG Rui, *et al*. Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition[J]. *IEEE Transactions on Multimedia*, 2022, 24: 1313–1324. doi: 10.1109/TMM.2021.3063612.

[7] 吕卫, 韩镓泽, 褚晶辉, 等. 基于多模态自注意力网络的视频记忆度预测[J]. 吉林大学学报:工学版, 2023, 53(4): 1211–1219. doi: 10.13229/j.cnki.jdxbgxb.20210842.
LYU Wei, HAN Jiaze, CHU Jinghui, *et al*. Multi-modal self-attention network for video memorability prediction[J]. *Journal of Jilin University:Engineering and Technology Edition*, 2023, 53(4): 1211–1219. doi: 10.13229/j.cnki.jdxbgxb.20210842.

[8] 陈杰, 马静, 李晓峰, 等. 基于DR-Transformer模型的多模态情感识别研究[J]. 情报科学, 2022, 40(3): 117–125. doi: 10.13833/j.issn.1007-7634.2022.03.015.
CHEN Jie, MA Jing, LI Xiaofeng, *et al*. Multi-modal emotion recognition based on DR-Transformer model[J]. *Information Science*, 2022, 40(3): 117–125. doi: 10.13833/j.issn.1007-7634.2022.03.015.

[9] MA Hui, WANG Jian, LIN Hongfei, *et al*. A transformer-based model with self-distillation for multimodal emotion recognition in conversations[J]. *IEEE Transactions on Multimedia*, 2023: 1–13. doi: 10.1109/TMM.2023.3271019.

[10] WU Yujin, DAOUDI M, and AMAD A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition[J]. *IEEE Transactions on Affective Computing*, 2024, 15(1): 157–172. doi: 10.1109/TAFFC.2023.3263907.

[11] YANG Kailai, ZHANG Tianlin, ALHUZALI H, *et al*. Cluster-level contrastive learning for emotion recognition in conversations[J]. *IEEE Transactions on Affective Computing*, 2023, 14(4): 3269–3280. doi: 10.1109/TAFFC.2023.3243463.

[12] WANG Min, CAO Donglin, LI Lingxiao, *et al*. Microblog sentiment analysis based on cross-media bag-of-words model[C]. Proceedings of International Conference on Internet Multimedia Computing and Service, Xiamen, China, 2014: 76–80. doi: 10.1145/2632856.2632912.

[13] ZADEH A, CHEN Minghai, PORIA S, *et al*. Tensor fusion network for multimodal sentiment analysis[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017: 1103–1114. doi: 10.18653/v1/D17-1115.

[14] LIU Zhun, SHEN Ying, LAKSHMINARASIMHAN V B, *et al*. Efficient low-rank multimodal fusion with modality-specific factors[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018: 2247–2256. doi: 10.18653/v1/P18-1209.

[15] TSAI Y H H, BAI Shaojie, 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. doi: 10.18653/v1/P19-1656.

[16] 韩虎, 吴渊航, 秦晓雅. 面向方面级情感分析的交互图注意力网络模型[J]. 电子与信息学报, 2021, 43(11): 3282–3290. doi: 10.11999/JEIT210036.
HAN Hu, WU Yuanhang, and QIN Xiaoya. An interactive graph attention networks model for aspect-level sentiment analysis[J]. *Journal of Electronics & Information*

Technology, 2021, 43(11): 3282–3290. doi: 10.11999/JEIT210036.

[17] SUN Hao, WANG Hongyi, LIU Jiaqing, *et al*. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation[C]. Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 3722–3729. doi: 10.1145/3503161.3548025.

[18] . HAZARIKA D, ZIMMERMANN R, and PORIA S. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis[C]. Proceedings of the 28th ACM International Conference on Multimedia, Seattle, USA, 2020: 1122–1131. doi: 10.1145/3394171.3413678.

[19] 蔡宇扬, 蒙祖强. 基于模态信息交互的多模态情感分析[J]. 计算机应用研究, 2023, 40(9): 2603–2608. doi: 10.19734/j.issn.1001-3695.2023.02.0050.
CAI Yuyang and MENG Zuqiang. Multimodal sentiment analysis based on modal information interaction[J]. *Application Research of Computers*, 2023, 40(9): 2603–2608. doi: 10.19734/j.issn.1001-3695.2023.02.0050.

[20] DEVLIN J, CHANG Mingwei, LEE K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. 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, USA, 2019: 4171–4186. doi: 10.18653/v1/N19-1423.

[21] WANG Di, LIU Shuai, WANG Quan, *et al*. Cross-modal enhancement network for multimodal sentiment analysis[J]. *IEEE Transactions on Multimedia*, 2023, 25: 4909–4921. doi: 10.1109/TMM.2022.3183830.

[22] HOCHREITER S and SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

[23] ZADEH A B, 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. doi: 10.18653/v1/P18-1208.

[24] WANG Yaoting, LI Yuanchao, LIANG P P, *et al*. Cross-attention is not enough: Incongruity-aware dynamic hierarchical fusion for multimodal affect recognition[EB/OL]. https://doi.org/10.48550/arXiv.2305.13583, 2023.

[25] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[26] MAI Sijie, ZENG Ying, ZHENG Shuangjia, *et al*. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis[J]. *IEEE Transactions on Affective Computing*, 2023, 14(3): 2276–2289. doi: 10.1109/TAFFC.2022.3172360.

[27] 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. doi: 10.1109/MIS.2016.94.

[28] TSAI Y H H, BAI Shaojie, 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. doi: 10.18653/v1/P19-1656.**(查阅网上资料,本条文献与第15条文献重复,请确认)**.

[29] QI Qingfu, LIN Liyuan, ZHANG Rui, *et al*. MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis[J]. *IEEE Access*, 2022, 10: 28750–28759. doi: 10.1109/ACCESS.2022.3157712.

[30] 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. doi: 10.1016/j.inffus.2022.09.025.

[31] SUN Zhongkai, SARMA P, SETHARES W, *et al*. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]. Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, USA, 2020: 8992–8999. doi: 10.1609/aaai.v34i05.6431.

[32] 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. doi: 10.18653/v1/2020.acl-main.214. **(查阅网上资料,未找到本条文献出版地信息,请确认并补充)**.

[33] QUAN Zhibang, SUN Tao, SU Mengli, *et al*. Multimodal sentiment analysis based on cross-modal attention and gated cyclic hierarchical fusion networks[J]. *Computational Intelligence and Neuroscience*, 2022, 2022: 4767437. doi: 10.1155/2022/4767437.

[34] ZOU Wenwen, DING Jundi, and WANG Chao. Utilizing BERT intermediate layers for multimodal sentiment analysis[C]. IEEE International Conference on Multimedia and Expo (ICME), Taipei, China, 2022: 1–6. doi: 10.1109/ICME52920.2022.9860014.

刘 佳：女，教授，研究方向为多模态情感分析、计算机视觉与图像处理、虚拟/增强现实.

宋 泓：男，硕士生，研究方向为多模态情感分析、计算机视觉与图像处理.

陈大鹏：男，副教授，研究方向为计算机视觉与图像处理、虚拟/增强现实、人机交互.

王 斌：男，硕士生，研究方向为计算机视觉与图像处理、虚拟/增强现实.

张增伟：男，硕士生，研究方向为增强现实、人机交互.

责任编辑：马秀强