徐钰淼, 单国轩. 面向多模态情感分析的增强型特征融合网络[J]. 智能计算机与应用,2026,16(5):197-203. DOI:10.20169/j.issn.2095-2163.25121603

面向多模态情感分析的增强型特征融合网络

徐钰淼[1], 单国轩[2]
(1 重庆移通学院 通信与信息工程学院, 重庆 401520; 2 重庆中车四方所智能装备技术有限公司,重庆 401133)

摘 要: 针对多模态情感分析模型在融合多模态表征时准确率低的问题,提出一种增强型特征融合网络(Enhanced Feature Fusion Network, EFFNet)。以文本为中心的跨模态交互模块(Text-Centered Cross-Modal Interaction Module,TCIM)基于改进 Transformer 结构以文本模态为主导,引导视觉、音频模态学习文本模态的关键语义信息,得到特征交互对,从而缓解模态间不一致性;同时引入轻量级多模态分割融合模块(Multi-Modal Segmentation Fusion Module, MSFM)提取增强型全局特征称为第四模态,实现跨模态的紧密语义耦合。通过该双融合机制最后得到多模态情感分析的增强型联合表示。在 CMU-MOSI 和 CMU-MOSEI 数据集上开展实验,回归和分类任务中的实验结果表明,未对齐设置下的 *F*1 指标分别达到了 86.43%和 86.32%。

关键词: 多模态情感分析; 跨模态交互; 多模态分割融合; 双融合机制
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2026)05-0197-07

Enhanced feature fusion network for multimodal sentiment analysis

XU Yumiao[1], SHAN Guoxuan[2]
(1 School of Communication and Information Engineering, Chongqing College of Mobile Communication, Chongqing 401520, China; 2 Chongqing CRRC Sifang Institute Intelligent Equipment Technology Company Limited, Chongqing 401133, China)

Abstract: To address the low accuracy of multimodal sentiment analysis models in fusing multimodal representations, an Enhanced Feature Fusion Network (EFFNet) is proposed. The Text-Centered Cross-Modal Interaction Module (TCIM), based on an improved Transformer architecture, establishes textual modality as the dominant modality, enabling visual and audio modalities to learn key information from text. This process generates feature interaction pairs, thereby mitigating inter-modal inconsistency. Additionally, a lightweight Multi-modal Segmentation Fusion Module (MSFM) is introduced to extract enhanced global features, termed the fourth modality, achieving tight semantic coupling across modalities. Through this dual-fusion mechanism, an enhanced joint representation for multimodal sentiment analysis is ultimately obtained. Experiments on the CMU-MOSI and CMU-MOSEI datasets demonstrate that the proposed model achieves *F*1 scores of 86.43% and 86.32%, respectively, under unaligned settings for regression and classification tasks.

Key words: multimodal sentiment analysis; cross-modal interaction; multimodal segmentation fusion; dual-fusion mechanism

0 引 言

自媒体时代下,用户依托视频载体自由表达观点与态度,此类内容往往蕴含着用户的丰富情感倾向。在此背景下,情绪分析的研究价值日益凸显,相关领域的关注度亦持续提升——其研究成果不仅为推荐系统的优化提供了重要理论支撑,对提升人机交互的自然度与效率更具有关键指导意义[1-2]。然而,人类情绪的复杂性决定了传统单模态分析方法难以全面、精准地捕获情感信息[3-5],因此多模态情感分析(MSA)方法应运而生。该方法需融合视觉、音频、文本三大核心模态,通过挖掘模态间的互补与互斥关系,实现对情感类别的精准划分与情绪强度的有效预测[6-8]。

鉴于多模态情感分析数据集包含时变信号特征,注意力机制[9]因能够有效捕获不同模态间的远程依赖关系,在该领域的应用愈发广泛[10-14]。Tsai 等[15]提出多模态变换器(MulT),通过执行多模态序列的软对齐操作,使不同时间步长的模态序列实现信息交互,任一模态均可借助其他两种模态的信息完成自身序列的动态更新。Rahman 等[16]则将多模态适应门(MAG)模块嵌入 BERT 与 XLNet 模型,以实现微调

作者简介: 徐钰淼(2000—),女,硕士研究生,助教。主要研究方向:多模态情感分析。Email:1530228388@qq.com。

收稿日期: 2025-12-16

阶段对多模态非语言数据的有效处理。值得注意的是,三大模态的语义表达能力存在显著不平衡性,这一特性易引发模态间的信息不协调问题[17-20]。例如,高昂的语音语调与含消极词汇“terrible”的文本内容相结合时,其整体传递的情感倾向仍为负面,而该情感极性的核心标识词“terrible”恰恰来源于文本模态。由此可见,文本模态在多模态情感分析中具有相较于非文本模态更为关键的地位[21-23]。Wang 等[24]的研究亦证实,文本模态包含更直观的情感极性词汇,通过构建文本多头注意力机制,可有效学习跨模态情感映射关系,兼顾模态间的一致性与差异性特征。Lei 等[25]则提出以文本模态为核心的融合框架,通过跨模态交互机制实现文本与非文本模态的定向信息融合及权重分配,并依托文本语义先验知识增强视听模态特征的内部表征,从而同步完成模态间关联与模态内上下文关系的建模。此外,由于非文本模态含有大量噪声,含有空间冗余信息的特征块会影响整体情感分析效果,因此需突出具有贡献度的特征块才能进行高效的融合[26]。Joze 等[27]借鉴 Hu 等[28]提出的 SENet 模型中的压缩与激励机制,通过引入跨模态连接模块,将其扩展应用于双模态特征交互,实现了多模态特征的重标定。Cheng 等[29]通过注意力时序卷积网络提取层次化特征块,并结合相关性感知的多层融合机制,有效解决了长时序依赖建模不足与跨模态融合粗糙两大问题,显著提升了多模态情感分析性能。

据此,本文提出 EFFNet,在保障效率的同时降低信息冗余,旨在同步探索模态间的关联关系。针对模态间不协调与语义不平衡问题,改进自注意力机制设计文本中心跨模态交互模块(TCIM):以文本模态为引导融入先验知识,构建文本-视觉、文本-音频特征交互对;结合分裂注意力机制[30]设计多模态分割融合模块(MSFM),将特征图分割为多组,通过全局上下文注意力突出高贡献特征块,其输出的增强型全局特征作为第四模态;后续通过门控注意力融合模块(GAFM)动态调整各模态维度权重,有效提升情绪强度预测精度。

1 模型方法

本文任务中涉及文本(T)、视觉(V)和音频(A)3 种模态,从数据集中定义每个给定视频片段的输入序列为 $I=(I_1^m,I_2^m,\cdots,I_i^m)$,代表共有 i 个输入序列的集合,其中 $m\in\{T,V,A\}$。

由图 1 所示,模型主要包括以下几个部分:模态编码部分,特征融合部分和预测部分。其中特征融合部分包括 3 个模块:跨模态交互模块、多模态分割融合模块和门控注意力融合模块。其中,TCIM、MSFM、GAFM 为特征融合部分。f_{vt} 为文本视觉特征交互对,f_t 为文本表征,f_{at} 为文本音频特征交互对,f'_{avt} 包含中间生成特征,f_{avt} 为 f'_{avt} 整合后的特征。

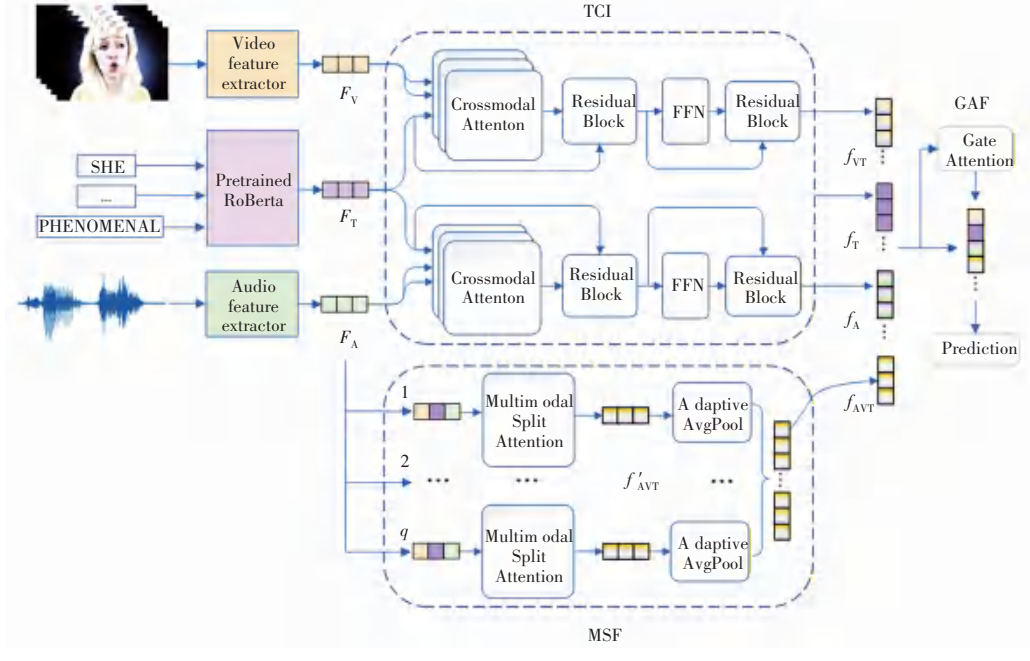


图 1 增强型特征融合网络结构

Fig. 1 Architecture diagram of the enhanced feature fusion network

1.1　模态编码

首先，在特征提取阶段，分别从 3 个特定模态中提取初始的特征信息。对于文本模态，使用预训练的 RoBERTa 模型[31]处理输入单词序列，得到句子表示 F_t：

$$F_t = \mathrm{RoBERTa}(I_i^t, \theta^{\mathrm{roberta}}) \tag{1}$$

式中：$F_t \in R^{L_t \times d_t}$，$\theta^{\mathrm{roberta}}$ 表示 RoBERTa 模型的网络参数。

对于视觉与音频模态，则采用堆叠的 LSTM 结合全连接层来捕获其时间特征序列：

$$F_m = FC(\mathrm{sLSTM}(I_i^m, \theta_m^{\mathrm{lstm}})) \tag{2}$$

式中：$F_m \in R^{L_m \times d_m}$，$m \in \{V, A\}$，$L_m$ 表示包含的各模态分别的序列长度，d_m 表示包含的各模态分别的特征维度，θ_m^{STM} 表示 LSTM 模型的网络参数。

最后，对 F_t 进行平均池化得到统一维度的文本情感表征 f_t，以便后续融合。

1.2　跨模态交互模块

在获得 3 个单模态表征后，通过文本模态的引导，以 F_m 和 F_t 作为输入，$m \in \{V, A\}$，利用跨模态注意力机制学习文本模态与非文本模态之间的关联特征，即 f_{VT} 和 f_{VT}，跨模态交互模块的核心部分是 Crossmodal Transformers，公式如下：

$$\mathrm{CMA}(F_m, F_t) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_t \boldsymbol{K}_m^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}_m = \mathrm{Softmax}\left(\frac{[F_t W_{Q_t}][W_{K_m}^{\mathrm{T}} F_m^{\mathrm{T}}]}{\sqrt{d}}\right)[F_m W_{V_m}] \tag{3}$$

式中：CMA(·，·)为跨模态注意力计算函数，F_m 为模态特征(视觉/音频)，F_{T} 为文本模态特征，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 为注意力机制的查询(Query)、键(Key)、值(Value)矩阵。

模块采用多头注意力，以捕获不同特征子空间的关联信息，并通过残差连接确保训练稳定性。在完成 X 个注意力头的计算后，将各头的输出结果进行拼接，再通过映射操作还原至原始表示空间，进而得到最终的注意力值，具体计算过程如下：

$$\mathrm{MultiHead}(F_t, F_m) = [\mathrm{head}_1 \oplus \mathrm{head}_2 \oplus \mathrm{head}_X]\boldsymbol{W}^O \tag{4}$$

式中：$\boldsymbol{W}^O$ 为线性变换权重矩阵，⊕表示拼接操作。

最终，对输出在时间尺度上进行平均池化，得到文本视觉特征交互对 f_{VT}、文本音频特征交互对 f_{AT}。

1.3　多模态分割融合模块

如图 2 所示，本模块对文本、视觉、音频 3 种模态进行拆解、连接等操作。

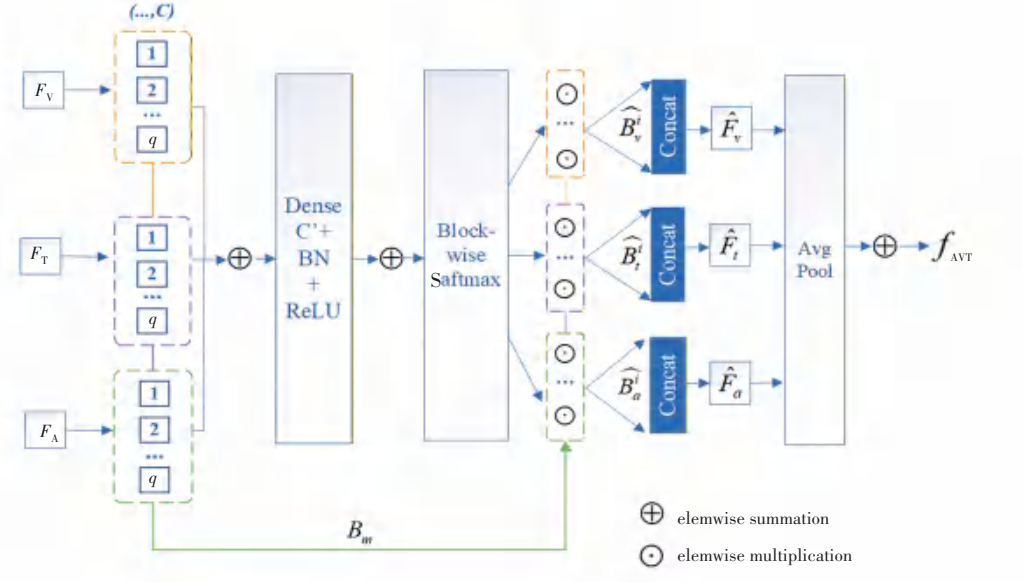


图 2　多模态分割融合模块结构图

Fig. 2　Structure of the multimodal segmentation and fusion module

拆解操作各模态特征按其通道数均匀分割为若干特征，各特征块通道数均为 C。对于其中任意一种模态，其对应的特征块集合可记为 B_m，若该模态总通道 C_m 无法被 C 整除时，在最后一个特征块的通道维度上补 0，即 $|B_m| = \lceil C_m / C \rceil$。

连接操作是学习多模态上下文的关键步骤，将同一模态的块连接成一个共享表示 D_m：

$$D_m(c) = \frac{1}{\prod_{i=1}^{K} N_i^{(n_1, s_m, n_K)}} \sum S_m(n_1, \cdots, n_K, c) \tag{5}$$

式中：$D_m(c)$ 为第 c 个通道的共享表示，N 为特征块的数量 。

对 D_m 进行行元素求和，得到多模态表示 G，经

过全连接层等变换后,得到捕获通道依赖关系的联合表示 Z,计算过程如下:

$$Z = \boldsymbol{W}_Z G + b_Z \tag{6}$$

式中:Z 为通道依赖关系的联合表示,$\boldsymbol{W}_Z$ 为全连接层权重矩阵,$Z \in R^{C'}$,$\boldsymbol{W}_Z \in R^{C' \times C}$,$b_Z \in R^{C'}$。

随后,通过对 Z 进行线性变换并应用 Softmax,得到每个特征块的注意力权重 A_m^i,公式如下:

$$A_m^i = \frac{\exp(U_m^i)}{\sum_k^M \sum_j^{|Bk|} (U_k^j)} \tag{7}$$

式中:A_m^i 为第 m 个模态中第 k 个特征块的注意力权重,U_m^i 为第 m 个模态中第 k 个特征块的未归一化权重。

使用超参数 $\lambda \in [0,1]$ 抑制能力,计算优化后的特征块 $\hat{B}_m^i$:

$$\hat{B}_m^i = [\lambda + (1-\lambda) \times A_m^i] \odot B_m^i] \tag{8}$$

式中:$\hat{B}_m^i$ 为优化后的特征块,λ 为超参数,控制注意力权重的影响程度。

最后,将同一模态所有优化后的特征块在通道维度合并生成 $\hat{F}_m$,即图 2 中的 f_{AVT}^t,再经过自适应平均池化操作,得到最终的 f_{AVT}。

1.4 门控注意力融合模块与预测

为融合这些多源特征,将其串联形成多模态中间表示 f^t,再通过一个门控注意力网络计算权重矩阵 $\boldsymbol{W}_{\text{Att}}$,用于动态调整不同特征的重要性,并将 f^t 与 $\boldsymbol{W}_{\text{Att}}$ 相乘得到最终的联合表示 f。$\boldsymbol{W}_{\text{Att}}$ 计算过程如下:

$$\boldsymbol{W}_{\text{Att}} = \boldsymbol{W}_2(\text{ReLU}(\boldsymbol{W}_1 \cdot f^t)) \tag{9}$$

式中:$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 为可学习参数,$\{\boldsymbol{W}_1, \boldsymbol{W}_2\} \epsilon R^{4\times d}$。

最后,对其进行层归一化与 Dropout 处理,并通过一个全连接层和 Softmax 函数生成最终的情感预测结果。

2 实验与结果分析

2.1 数据集

为验证模型的有效性,本文选用 CMU-MOSI 与 CMU-MOSEI 两个广泛使用的基准数据集进行实验。其中,CMU-MOSI 包含 2 199 个 YouTube 影评视频片段;CMU-MOSEI 作为其升级版,含 3 228 个视频片段(共 22 856 个语句)。两类数据集的样本均标注了情感强度分数($y \in [-3, 3]$)及 7 种离散的情感标签(涵盖强烈、消极至强烈、积极),分数越小负面情绪越显著,越大则正面情绪越突出,0 代表中性。

2.2 模型训练

实验在 NVIDIA GeForce RTX 4090 GPU 上完成。模型使用 Adam 优化器进行训练,参数设置如下:文本学习率 tlr 均设为 5e-5,视觉学习率 vlr 分别设为 7e-4 与 2e-4,音频学习率 alr 分别为 2e-3 与 6e-3。跨模态交互模块的多头注意力头数为 12;多模态分割融合模块参数设置中,缩减因子(reduction_factor)固定为 4,各模态均划分为 5 个有序段($q = 5$);分割融合模块的丢弃率(block_dropout)取值为 0.3 和 0.2,衰减因子则设为 4 和 3。

2.3 实验结果分析

本文与 TFN[32]、MFM[33]、MulT[15]、Mag-Bert[16]、MISA[34]、Self-MM[35]、TMRN[25]、TETFN[24] 8 种现有方法开展对比实验。上述方法的实验数据集分为对齐与未对齐两类,通常对齐数据集更利于模型性能提升,而本文实验采用未对齐数据集进行验证,结果见表 1 和表 2。

表 1 不同模型在 CMU-MOSI 上的实验结果

Table 1 Experimental results of different models on CMU-MOSI

Model	MAE	Corr	Acc-7	Acc-2	*F*1
TFN	0.901	0.698	34.90	-/80.80	-/80.70
MFM	0.877	0.706	35.40	-/81.70	-/81.60
MulT	0.861	0.711		-/81.10	-/81.00
MAG-Bert	0.727	0.781	43.62	82.37/84.43	82.50/84.61
MISA	0.783	0.761	42.30	81.80/83.40	81.70/83.60
Self-MM	0.737	0.779	44.17	82.80/84.60	82.67/84.54
TMRN	0.704	0.784	48.68	83.67/85.67	83.45/85.52
TETFN	0.717	0.800		84.05/86.10	83.83/86.07
EFFNet	0.669	0.818	46.65	84.40/86.43	84.46/86.43

表 2　不同模型在 CMU-MOSEI 上的实验结果

Table 2　Experimental results of different models on CMU-MOSEI

Model	MAE	Corr	Acc-7	Acc-2	*F*1
TFN	0.593	0.700	50.20	-/82.50	-/82.10
MFM	0.568	0.717	51.30	-/84.40	/84.30
MulT	0.580	0.703		-/82.50	-/82.30
MAG-Bert	0.543	0.755	52.67	82.51/84.82	82.77/84.71
MISA	0.555	0.756	52.20	83.60/85.50	83.80/85.30
Self-MM	0.535	0.763	53.48	79.46/84.16	80.18/84.24
TMRN	0.535	0.762	53.65	83.39/86.19	83.67/86.08
TETFN	0.551	0.748		84.25/85.27	84.18/85.27
EFFNet	0.525	0.785	53.90	83.88/86.30	84.22/86.32

表中 Acc-2 与 *F*1 指标分别列出“负/非负”与“负/正”两类分类结果。可见 EFFNet 的整体性能最优；与传统方法 TFN 和 MFM 相比较，基于 Transformer 的方法表现更优；而在基于 Transformer 的 MulT、MAG-Bert、MISA、Self-MM、TETFN 方法中，EFFNet 表现更优。具体说来，相较于 TETFN[23]，EFFNet 通过构建以文本为中心并融合 4 种模态的策略，在 CMU-MOSI 和 CMU-MOSEI 数据集上 *F*1 值分别提升了 0.36%、1.05%，二分类 Acc 指标分别提升了 0.33%和 1.03%，MAE 分别降低了 4.8%和 2.6%，验证了所提融合策略的有效性。

2.4　消融实验

2.4.1　模态消融

为系统评估各模态在多模态情感分析中的贡献，本文在 CMU-MOSI 数据集上开展了 4 组模态消融实验（表 3）。实验通过“w/o”标记分别剔除视觉、音频、文本及第四模态的输入，以观察模型性能变化。结果显示，移除视觉或音频模态均导致模型性能显著下滑，而文本模态的缺失对模型影响最为严重，其关键性作用得到充分验证。此外，第四模态的移除亦引起各指标回落，表明该模块能够有效增强关键特征块的贡献，并抑制冗余信息干扰。实验从模态互补性与重要性维度验证了本文方法的合理性与有效性。

表 3　各模态的消融实验结果

Table 3　Ablation experimental results of each modality

Model	MAE ↓	Acc ↑	*F*1 ↑
EFFNet	0.669	84.40/86.43	84.46/86.43
w/o text	1.464	54.69/53.35	54.98/54.65
w/o acoustic	0.612	83.86/86.30	83.58/85.46
w/o visual	0.674	82.46/84.88	83.02/85.39
w/o forth	0.670	83.42/85.21	84.08/85.78

2.4.2　模块消融

为验证特征融合模块的必要性，本文在 CMU-MOSI 数据集上进行了 3 组模块消融实验（表 4）。实验以“w/o”标记表示移除对应模块，结果显示：移除 TCIM 模块后 Acc 下降 1.98%，*F*1 下降 2.01%；移除 MSFM 模块后 Acc 下降 1.98%，*F*1 下降 1.99%；而 GAFM 模块的缺失对模型性能影响最为显著。实验表明，3 个模块通过协同作用有效提升了模态整合效率，验证了其组合设计的必要性。

表 4　各模块的消融实验结果

Table 4　Ablation experimental results of each module

Model	MAE ↓	Acc ↑	*F*1 ↑
EFFNet	0.669	84.40/86.43	84.46/86.43
w/o TCIM	0.683	82.49/84.45	82.87/84.42
w/o MSFM	0.704	82.44/82.45	82.25/84.44
w/o GAFM	0.694	81.04/84.49	82.14/84.46

2.4.3　模态权重可视化分析

图 3 展示了首轮训练过程中各模态权重的变化趋势。结果表明：文本模态（T）与第四模态（AVT）的权重随训练进程逐步上升，在约 20 个 batch 后趋于稳定，且最终显著高于 VT 和 AT 模态。该现象说明模型能够自适应地区分模态重要性，并验证了所构建的第四模态在特征融合中的有效贡献。

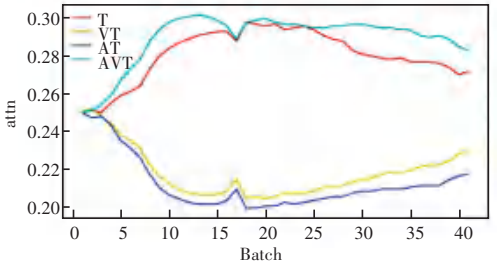


图 3　4 种模态权重比例变化曲线图

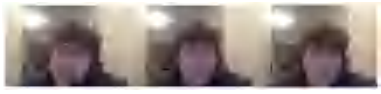
Fig. 3　Chart of the variation curve of four modality weight proportions

2.4.4 实例验证与分析

表 5 呈现了 3 个典型样例的预测结果与真实标签对比情况。其中样例 1 中,文本内容倾向积极,且视觉、音频模态的情感倾向与文本保持一致,本文模型据此准确判断出情感极性及情绪强度,预测误差较小。样例 2 的情况则有所不同:文本内容呈中性倾向,但视觉和音频信息偏积极,模态间的情感倾向存在差异。即便如此,模型仍重点捕捉到“pretty average”这类关键表述,充分发挥了文本模态在模型中的主导地位,最终该样本预测为中性情感,情绪强度绝对误差仅为 0.050,既印证了文本模态在情感判断中的主导地位与高权重特性,也体现出第四模态对冗余信息的有效抑制作用。综合实验结果可见,本文方法在多模态信息存在冲突时,第四模态作为增强型全局特征可有效缓解模态间歧义对分析结果的干扰。

表 5 在 CMU-MOSI 测试集上的案例研究

Table 5 Case study on the CMU-MOSI test set

样例		多模态信息	真实值	预测值	情感标签
	文本	and actually is better than scream 2			
1	视觉	 情微微激动	1.6	1.588	积极
	音频	语气坚定			
	文本	and all i can say is he′s a pretty average guy			
2	视觉	 嘴巴张大,神情愉悦	0	0.050	中性
	音频	语气欢快			
	文本	and it′s truly heartbreaking to see that contrasted with the state of things			
3	视觉	 低头皱眉,目光漂移	-1.5	-1.532	消极
	音频	语气稍有停顿			

3 结束语

本文提出了一种用于多模态情感分析的增强型特征融合网络(EFFNet),以解决非对齐多模态序列的交互与融合问题。与传统跨模态方法不同,本文充分考虑视频片段中不同模态的贡献差异,通过文本模态序列引导视觉与音频模态,有效缓解模态间的不协调性问题。同时,通过将 3 种模态切片重组,突出具有更高贡献度的特征块,增强语义表达能力。消融实验结果表明,各模态及融合模块均对模型性能具有重要贡献,文本模态在多模态情感分析中起主导作用。可视化分析显示,模型能够自适应区分不同模态的重要性,验证了所提融合策略的有效性。

参考文献

[1] 张国防, 袁国强, 赵胜利. 面向虚假信息的多模态在线评论情感分析[J]. 河北大学学报(自然科学版),2025,45(2): 216-224.

[2] 张宇轩. 基于多模态情感分析的人机交互关键技术研究[D]. 南京:南京信息工程大学,2025. DOI:10.27248/d.cnki.gnjqc.2025.001834.

[3] 李玥琪, 王晰巍, 王秋月. 基于多模态大语言模型的社交媒体多模态情感分析模型研究[J]. 情报理论与实践,2025,48(11): 188-197. DOI:10.16353/j.cnki.1000-7490.2025.11.022.

[4] DAS R, SINGH T D. Multimodal sentiment analysis: A survey of methods, trends, and challenges[J]. ACM Computing Surveys, 2023, 55(13s): 1-38.

[5] 李萍, 王丽丽. 国内多模态技术的研究现状与发展趋势:基于 CiteSpace 的可视化分析[J]. 智能计算机与应用,2025,15(1): 194-202. DOI:10.20169/j.issn.2095-2163.24062803.

[6] LIN R, HU H. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1686-1702.

[7] SUN T, NI J, WANG W, et al. General debiasing for multimodal sentiment analysis[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York:ACM,2023: 5861-5869.

[8] HUANG C, ZHANG J, WU X, et al. TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis[J]. Knowledge-Based Systems, 2023, 269: 110502.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information

Processing Systems. NeurIPS，2017：5998-6008.
[10]张换香，刘璐瑶，张景，等. 基于信息增强多头注意力的多模态情感分析[J]. 计算机应用与软件，2025，42(9)：156-164.
[11]陈万志，侯悦. 融合多层次注意力和情感尺度向量的时序多模态情感分析模型[J]. 数据分析与知识发现，2025，9(7)：52-65.
[12]吕学强，田驰，张乐，等. 融合多特征和注意力机制的多模态情感分析模型[J]. 数据分析与知识发现，2024，8(5)：91-101.
[13]XUE X，ZHANG C，NIU Z，et al. Multi-level attention map network for multimodal sentiment analysis[J]. IEEE Transactions on Knowledge and Data Engineering，2022，35(5)：5105-5118.
[14]张海侠，张星，车雷，等. 基于稀疏注意力的多任务多模态情感识别模型[J]. 智能计算机与应用，2025，15(8)：18-24. DOI：10.20169/j.issn.2095-2163.250803.
[15]TSAI Y 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. ACL，2019：6558-6569.
[16]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. ACL，2020：2359-2369.
[17]LI M，ZHU Z，LI K，et al. Diversity and balance：Multimodal sentiment analysis using multimodal-prefixed and cross-modal attention[J]. IEEE Transactions on Affective Computing，2024，15(2)：570-581.
[18]ZHAO J，DONG W，SHI L，et al. Multimodal feature fusion method for unbalanced sample data in social network public opinion[J]. Sensors，2022，22(15)：5528.
[19]PAN Q，MENG Z. Hybrid uncertainty calibration for multimodal sentiment analysis[J]. Electronics，2024，13(3)：662.
[20]ZHU A，HU M，WANG X，et al. KEBR：Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York：ACM，2024：5732-5741.
[21]SUN Z，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 AAAI Conference on Artificial Intelligence. AAAI，2020，34(5)：8992-8999.
[22]LIN R，HU H. Multi-task momentum distillation for multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing，2023，15(2)：549-565.
[23]卢婵，郭军军，谭凯文，等. 基于文本指导的层级自适应融合的多模态情感分析[J]. 山东大学学报(理学版)，2023，58(12)：31-40.
[24]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.
[25]LEI Y，YANG D，LI M，et al. Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences[C]//Proceedings of CAAI International Conference on Artificial Intelligence. Cham：Springer，2023：189-200.
[26]ZHOU H，DU J，ZHANG Y，et al. Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2021，29：2617-2629.
[27]JOZE H R V，SHABAN A，IUZZOLINO M L，et al. MMTM：Multimodal transfer module for CNN fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2020：13289-13299.
[28]HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2018：7132-7141.
[29]CHENG H，YANG Z，ZHANG X，et al. Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion[J]. IEEE Transactions on Affective Computing，2023，14(4)：3149-3163.
[30]SU L，HU C，LI G，et al. Msaf：Multimodal split attention fusion[J]. arXiv preprint arXiv，2012.07175，2020.
[31]LIU Y，OTT M，GOYAL N，et al. Roberta：A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv，1907.11692，2019.
[32]ZADEH A，CHEN M，PORIA S，et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv，1707.07250，2017.
[33]TSAI Y H H，LIANG P P，ZADEH A，et al. Learning factorized multimodal representations[J]. arXiv preprint arXiv，1806.06176，2018.
[34]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：ACM，2020：1122-1131.
[35]YU W，XU H，YUAN Z，et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence. 2021，35(12)：10790-10797. DOI：10.48550/arXiv.2102.04830.