



基于 Transformer 的多子空间多模态情感分析

田昌宁¹, 贺昱政¹, 王 笛¹, 万 波¹, 郭栩彤²

(1. 西安电子科技大学 计算机科学与技术学院, 陕西 西安 710071;

2. 中国电子科技集团公司第五十四研究所, 河北 石家庄 050081)

摘要 多模态情感分析是指通过文本、视觉和声学信息识别视频中人物表达出的情感。现有方法大多通过设计复杂的融合方案学习多模态一致性信息, 而忽略了模态间和模态内的差异化信息, 导致缺少对多模态融合表示的信息补充。为此提出了一种基于 Transformer 的多子空间多模态情感分析(multi-subspace Transformer fusion network for multimodal sentiment analysis, MSTFN)方法。该方法将不同模态映射到私有和共享子空间, 获得不同模态的私有表示和共享表示, 学习每种模态的差异化信息和统一信息。首先, 将每种模态的初始特征表示分别映射到各自的私有和共享子空间, 学习每种模态中包含独特信息的私有表示与包含统一信息的共享表示。其次, 在加强文本模态和音频模态作用的前提下, 设计二元协同注意力跨模态 Transformer 模块, 得到基于文本和音频的三模态表示。然后, 使用模态私有表示和共享表示生成每种模态的最终表示, 并两两融合得到双模态表示, 以进一步补充多模态融合表示的信息。最后, 将单模态表示、双模态表示和三模态表示拼接作为最终的多模态特征进行情感预测。在 2 个基准多模态情感分析数据集上的实验结果表明, 该方法与最好的基准方法相比, 在二分类准确率指标上分别提升了 0.025 6/0.014 3 和 0.000 7/0.002 3。

关键词 多模态情感分析; Transformer 结构; 多子空间; 多头注意力机制

中图分类号: TP391.1 **DOI**: 10.16152/j.cnki.xdxbr.2024-02-002

Multi-subspace multimodal sentiment analysis method based on Transformer

TIAN Changning¹, HE Yuzheng¹, WANG Di¹, WAN Bo¹, GUO Xutong²

(1. School of Computer Science and Technology, Xidian University, Xi'an 710071, China;

2. China Electronics Technology Group Corporation 54th Research Institute, Shijiazhuang 050081, China)

Abstract Multimodal sentiment analysis refers to recognizing the emotions expressed by characters in a video through textual, visual and acoustic information. Most of the existing methods learn multimodal coherence information by designing complex fusion schemes, while ignoring inter-and intra-modal differentiation information, resulting in a lack of information complementary to multimodal fusion representations. To this end, we

收稿日期: 2023-12-09

基金项目: 国家科技创新 2030 - “新一代人工智能”重大项目(2022ZD0117103); 中央高校基本科研业务费项目(QTZX23084); 国家自然科学基金面上项目(62072354)。

第一作者: 田昌宁, 男, 从事多模态情感计算研究, cntian@stu.xidian.edu.cn。

通信作者: 王笛, 女, 副教授, 博士生导师, 从事情感计算、多模态机器学习研究, wangdi@xidian.edu.cn。

propose a multi-subspace Transformer fusion network for multimodal sentiment analysis (MSTFN) method. The method maps different modalities to private and shared subspaces to obtain private and shared representations of different modalities, learning differentiated and unified information for each modality. Specifically, the initial feature representations of each modality are first mapped to their respective private and shared subspaces to learn the private representation containing unique information and the shared representation containing unified information in each modality. Second, under the premise of strengthening the roles of textual and audio modalities, a binary collaborative attention cross-modal Transformer module is designed to obtain textual and audio-based tri-modal representations. Then, the final representation of each modality is generated using modal private and shared representations and fused two by two to obtain a bimodal representation to further complement the information of the multimodal fusion representation. Finally, the unimodal representation, bimodal representation, and trimodal representation are stitched together as the final multimodal feature for sentiment prediction. Experimental results on two benchmark multimodal sentiment analysis datasets show that the present method improves on the binary classification accuracy metrics by 0.025 6/0.014 3 and 0.000 7/0.002 3, respectively, compared to the best benchmark method.

Keywords multimodal sentiment analysis; Transformer structure; multiple subspaces; multi-head attention mechanism

在互联网发展初期,用户大多只用文本这一单一模态来表达观点态度,但随着互联网不断向着多模态信息方向发展,仅从文本中获得的信息不足以挖掘人们的观点态度^[1]。现有的多模态情感分析方法致力于探索一种复杂且有效的多模态融合方法以学习融合表示,从而获得多模态一致性信息,但由于说话者的特殊风格(如反讽),单一模态中可能包含与多模态一致性信息不同的情感信息。这些方法大多将这些单一模态信息当作噪声处理,因此,丢失了模态内和模态间的差异性信息,导致模型学习到的情感信息并不全面,从而限制了模型的性能。

尽管不同模态之间存在异质性,但均具有相同的动机和情感倾向,与视频片段整体的情感走向一致。因此,将不同模态映射到同一个特征子空间中,可以学习到包含统一信息的多模态表示。

同时,将不同模态映射到不同的特征子空间中,能够学习特定于某个模态的信息,从而更全面地学习每种模态所表达的情感信息。现有的多模态情感分析方法在设计跨模态 Transformer 进行模态融合时使用三元对称的结构,如图 1 (a)所示,这种方式将每种模态分别与另外 2 种模态进行融合,导致多模态融合表示存在冗余信息,从而影响了模型对视频中情感的判断。因此,如何设计一种能够充分融合多模态表示并去除冗余信息的二元跨模态 Transformer 融合方案成为目前多模态情感分析中的一个热点研究问题。图 1(b)为二元融合结构示意图,这种方式只选取其中 2 个模态与另外 2 个模态进行融合,因此,能够在保证模态融合的前提下,减少冗余信息的生成,提升模型的性能。

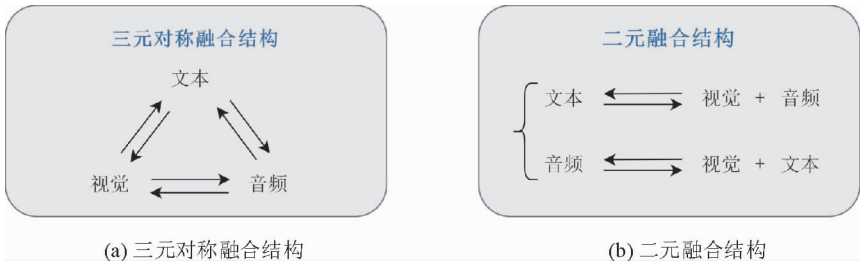


图 1 三元对称融合方案与二元融合方案示意图

Fig. 1 Schematic diagram of ternary symmetric fusion scheme and binary fusion scheme

本文提出了一种能够同时学习模态统一性信息和模态间与模态内差异化信息,并减少冗余信

息的多模态情感分析方法,即基于 Transformer 的多子空间多模态情感分析方法。该方法设计了多

个子空间,将不同模态的表示映射后得到包含统一信息和特定于某种模态的差异化信息的特征表示。设计了基于文本、音频、视频模态的协同注意力二元跨模态 Transformer 模块,使得一种模态能够同时与另外 2 种模态相互映射,简化复杂的融合结构、减少冗余信息的同时更充分地建模模态间的交互作用。此外,将每种模态的 2 种表示融合后生成最终包含多方面情感信息的单模态表示,并两两结合生成双模态表示,使模型更进一步学习到与情感相关的信息。最后,通过在 2 个基准数据集上设计的一系列实验验证了该方法的优越性和有效性。

1 相关工作

1.1 单模态情感分析

情感分析作为当前的热门研究领域,其发展初期是以文本为主的单模态情感分析,许多研究者在文本情感分析领域做了深入研究,为情感分析的进一步发展奠定了基础。早期的文本情感分析工作首先使用词袋模型 Bagofwords^[2]或带有频率的 N-gram^[3]从文本中提取特征,将文本转换成向量。然后使用支持向量机^[4]、朴素贝叶斯^[5]等传统的机器学习方法对情感极性进行分类。随着深度学习技术的广泛应用,卷积神经网络^[6]和循环神经网络^[7]被应用到文本情感分析中,由于这些网络能够捕捉上下文之间的关系,因此取得了良好的效果。近年来,随着基于 Transformer 结构的预训练模型的兴起,文本情感分析的性能也取得了突破性的发展。除了文本数据中包含丰富的情感信息之外,视频和音频数据中也包含了大量的情感信息。视频中人物的面部表情和肢体动作均包含了人物的情感信息,早期的方法采用局部二值模式^[8]、局部相位量化特征^[9]以及 Gabor 特征^[10]等手工特征对图像的情感进行分类。近年来,大多数方法均采用深度神经网络对图像中的情感极性进行分析。音频中的语速、语调、声音强度以及 Mel 频率倒谱系数^[11]等声学特征均与说话者的情感息息相关,对这些音频特征进行分析也可以得到说话者的情感极性。

尽管各领域的单模态情感分析方法在不断发展进步,但由于单一模态所包含的与情感相关的信息是有限的,在缺乏与其他模态信息进行交互的情况下难以全面理解人的真实情感,而多模态

情感分析可以很好地解决该问题。

1.2 多模态情感分析

随着新媒体产业的高速发展以及智能手机的普及,包含多种模态数据的短视频数量激增,越来越多的人通过短视频来表达对某一产品或事物的看法。多模态情感分析任务利用短视频中文本、图像以及音频多种模态的数据分析人们的情感极性。针对这一任务,研究者提出了许多方法来提高情感分析的准确率。Poria 等人使用各个模态的特征提取器提取特征,然后将 3 个模态的特征输入到浅层模型中进行拼接,最后将拼接的特征输入到分类模块中得到情感极性^[12]。由于不同模态对情感极性判断的贡献度不同,Kampman 等人对不同的模态特征进行加权融合^[13]。罗渊貽等人提出一种自适应权重融合策略获取不同模态对情感分析的贡献度^[14]。Zadeh 等人提出的 TFN 模型通过笛卡尔积和张量融合的方式对模态内和模态间进行建模^[15]。随着注意力机制在各个领域表现出的显著成果,许多研究者利用注意力机制将不同模态的信息进行融合。张涛等人和陈宏松等人均使用交叉注意力对不同模态的信息进行融合^[16-17]。周柏男等人使用模态内自注意力和模态间的交叉注意力使各模态间信息共享并减少噪声信息^[18]。卢婵等人提出文本指导的多模态层级自适应融合方法,利用注意力机制将文本模态与其他模态进行融合^[19]。不同的是,Han 等人提出了一种多模态融合的新思路,摒弃了设计复杂的融合方式的想法,设计一种层次化框架,最大化单模态输入对和多模态融合结果与单模态输入之间的互信息,通过多模态融合维护情感分析任务相关信息^[20]。

2 多模态情感分析方法

多模态情感分析的目标是利用视频中的多模态信号检测说话者所传达的情感。视频被切割成话语级的视频片段并作为模型的输入,对于每个话语片段,输入包括来自文本(t)、视觉(v)和音频(a)模态的特征序列。使用语言预训练模型 BERT 提取文本高级特征序列 $I_t \in \mathbf{R}^{T_t \times d_t}$,采用视觉预训练模型 ViT 提取视觉模态的高级特征序列 $I_v \in \mathbf{R}^{T_v \times d_v}$,采用音频预训练模型 wav2vec 提取高级声学特征 $I_a \in \mathbf{R}^{T_a \times d_a}$ 。其中, T_m 代表各自特征序列的长度, d_m 代表各自的特征维度, $m \in \{t, v, a\}$ 。

图2为本文提出的基于Transformer^[21]的多子空间多模态情感分析方法的整体网络框架图,该网络主要由4个模块组成,分别是模态私有与共享表示学习模块(private and shared representations learning module, PSRLM)、协同注意力跨模态Transformer模块(co-attention cross-modal Transformer, CACT)、双模态表示生成模块(bi-modal representation generation module, BRGM)以及情感预测模块。模态私有与共享表示学习模块通过将不同模态的特征序列映射到不同子空间来学习模态的私有表示和共享表示。协同注意力跨模态

Transformer模块同时建模1种模态和其余2种模态之间的交互作用,获得基于文本和基于音频模态的三模态表示。在双模态表示生成模块中,首先使用模态的私有表示和共享表示生成该模态的最终表示,每2种模态相融合后再生成双模态表示。在情感预测模块中,将单模态表示、双模态表示与协同注意力跨模态Transformer模块中获得的多模态表示进行拼接后,经过Transformer编码器编码得到最终的融合表示,并通过全连接层进行情感预测。

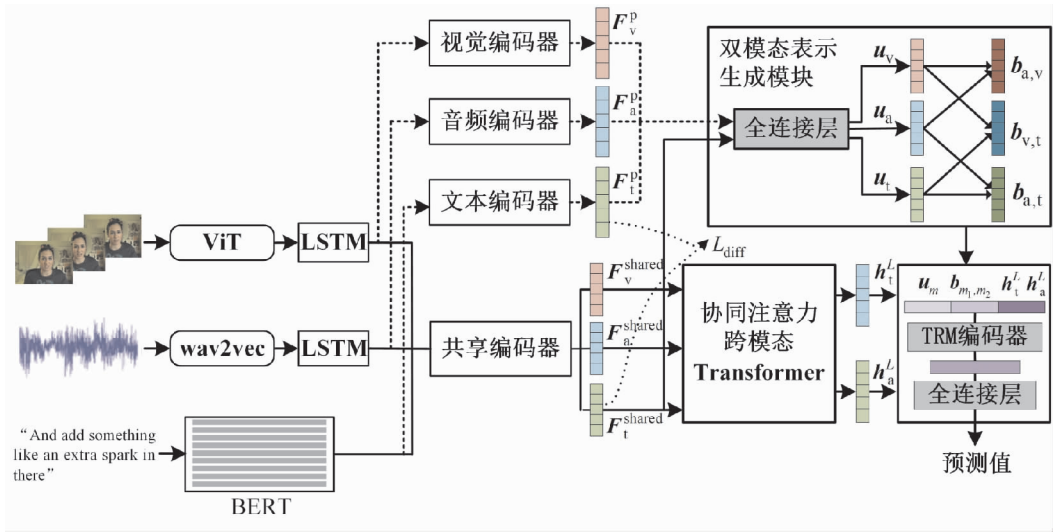


图2 基于Transformer的多子空间多模态情感分析网络框架图

Fig. 2 Transformer-based multi-subspace multimodal sentiment analysis network framework diagram

2.1 模态私有与共享表示学习模块

在分别使用语言预训练模型BERT^[22]、视觉预训练模型ViT^[23]以及音频预训练模型wav2vec从原始数据中提取初始的特征序列 I_m 之后,为了使视觉和音频模态的特征序列获得时序信息,使用单层单向长短期记忆网络为这2种模态的特征序列注入上下文相关信息和长期依赖,并使用全连接层将3种模态的特征表示映射到同一维度,以便输入到后续的网络模型中进行操作,该过程如式(1)~(3)所示。

$$F_a = \text{FC}(\text{sLSTM}(I_a \theta_a^{\text{lstm}})) \quad (1)$$

$$F_v = \text{FC}(\text{sLSTM}(I_v \theta_v^{\text{lstm}})) \quad (2)$$

$$F_t = \text{FC}(I_t) \quad (3)$$

式中: $F_m \in \mathbf{R}^{T_m \times d}$,代表将各模态的初始特征表示经过长短期记忆网络编码并且统一特征维度之后的投影表示; d 代表统一后的特征维度。

为了学习同一模态内不同方面的情感信息和不同模态间的差异化信息,设计了模态私有与共

享表示学习模块。首先,同一个视频片段中的每种模态的数据都具有相同的情感倾向,这也是多模态情感分析能够实现的原因。因此,为了获取每种模态所包含的统一性信息,首先定义一个共享编码器 $\text{Enc}^{\text{shared}}$,将每种模态的初始特征序列表示经过 $\text{Enc}^{\text{shared}}$ 编码后映射到同一个子空间中,获得具有统一情感倾向和共性情感情信息的共享表示。同时,特征序列在经过共享编码器编码后有助于缩小不同模态之间的异质鸿沟,如式(4)所示。

$$F_m^{\text{shared}} = \text{Enc}^{\text{shared}}(F_m, \theta^{\text{shared}}) \quad (4)$$

式中: $F_m^{\text{shared}} \in \mathbf{R}^{T_m \times d_{\text{map}}}$ 代表每种模态的共享表示; d_{map} 是共享表示的特征维度。采用简单的全连接层对每种模态的初始特征序列进行映射,3种模态的共享参数为 θ^{shared} 。

同一模态的数据在包含统一情感倾向信息的同时,也具有特定于模态和包含说话者敏感风格的特殊信息,比如文本模态时常具有的讽刺倾向,一部分人群习惯于采用夸张的面部表情表达与所

述文字相反的情感。因此,学习特定模态与情感相关的信息,能够实现对情感分析进一步的信息补充,从而提升模型的性能。在将每种模态的初始特征序列通过 $\text{Enc}^{\text{shared}}$ 映射到共享子空间以学习模态统一表示的同时,分别为 3 种模态定义各自的私有编码器 Enc_m^p , 经过 Enc_m^p 编码后将每种模态的初始特征序列映射到各自的私有子空间中,以捕获特定于不同模态的差异化信息,并且学习模态内和模态间的不同特征,该过程如式(5)所示。

$$\mathbf{F}_m^p = \text{Enc}_m^p(\mathbf{F}_m, \theta_m^p) \quad (5)$$

式中: $\mathbf{F}_m^p \in \mathbf{R}^{T_m \times d_{\text{map}}}$ 代表各模态的私有表示。同样地,采用全连接层将每种模态的初始特征序列映射到各自的私有子空间中,每种模态各自的编码器具有不同的网络参数 θ_m^p 。

2.2 协同注意力跨模态 Transformer 模块

在将每种模态的初始特征序列映射到共享子空间并获得共享表示之后,每种模态的共享表示都获得了一致的情感倾向和统一性信息,为了更进一步挖掘每种模态与情感相关的信息,同时建模不同模态之间的交互作用,并减小不同模态间的异质性,从而学习多模态融合表示,本文设计了一种协同注意力跨模态 Transformer,其结构如图 3 所示。

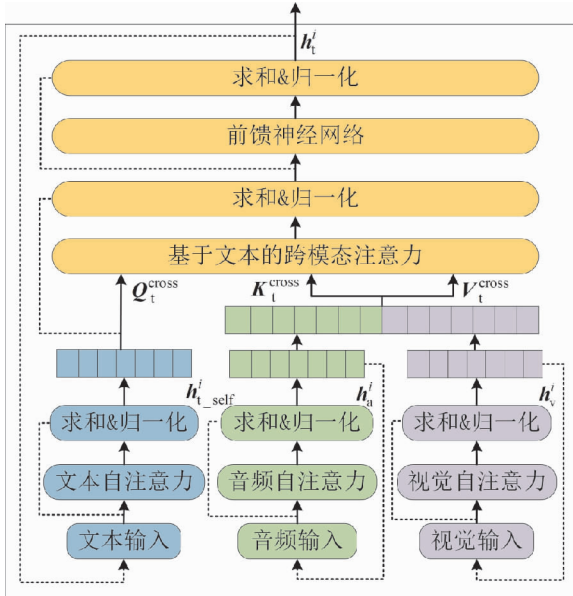


图 3 基于文本的协同注意力跨模态 Transformer 层示意图

Fig. 3 Text-based schematic diagram of the cross-modal Transformer layer of collaborative attention

为每种模态的共享表示添加位置编码以使模型能够捕获序列的顺序信息,如式(6)所示,

$$\mathbf{X}_m = \mathbf{F}_m^{\text{shared}} + \mathbf{P}_m \quad (6)$$

式中: $\mathbf{P}_m \in \mathbf{R}^{T_m \times d_{\text{map}}}$ 为每种模态的位置编码; $\mathbf{X}_m \in \mathbf{R}^{T_m \times d_{\text{map}}}$ 为每种模态获得位置编码后的表示,用于输入到后续的网络模型中进行多模态融合。

多头注意力机制 output(式中简记 O_{output} 作为跨模态注意力的核心组成部分,定义多头注意力机制 $O_{\text{output}} = \text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, 如式(7) ~ (9) 所示,

$$O_{\text{output}} = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_{h_n}) \mathbf{W}^O \quad (7)$$

$$\text{head}_i = \text{Attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \quad (8)$$

$$\text{Attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax}(\mathbf{Q}_i \mathbf{K}_i^T \lambda) \mathbf{V}_i \quad (9)$$

式中: $\mathbf{Q}_i = \mathbf{h}_\alpha \mathbf{W}_i^{Q_\alpha} \in \mathbf{R}^{T_m \times d_f}$, $\mathbf{K}_i = \mathbf{h}_\beta \mathbf{W}_i^{K_\beta} \in \mathbf{R}^{T_m \times d_f}$, $\mathbf{V}_i = \mathbf{h}_\beta \mathbf{W}_i^{V_\beta} \in \mathbf{R}^{T_m \times d_g}$ 分别代表多头注意力的输入经过不同权重映射后学习到的投影表示, $\alpha, \beta \in \{\text{a}, \text{t}, \text{v}\}$ 。其中, $\mathbf{W}_i^{Q_\alpha} \in \mathbf{R}^{d \times d_f}$; $\mathbf{W}_i^{K_\beta} \in \mathbf{R}^{d \times d_f}$; $\mathbf{W}_i^{V_\beta} \in \mathbf{R}^{d \times d_g}$; $\mathbf{W}^O \in \mathbf{R}^{h_n d_g \times d}$; h_n 代表多头注意力的“头”数; 令 $d = d_{\text{map}}$ 。

在协同注意力跨模态 Transformer 模块中,以基于文本的协同注意力跨模态 Transformer($\text{t} \rightarrow \text{a} + \text{v}$) 为例,共包含 L 层协同注意力跨模态 Transformer 层(简称为 CACT 层),对于第 i 层 CACT 层($i = 1, 2, \dots, L$),针对每种模态的输入首先采用自注意力机制探索模态内的交互作用,如式(10) ~ (12) 所示。

$$\mathbf{h}_m^0 = \mathbf{X}_m \quad (10)$$

$$\mathbf{h}_{\text{t_self}}^i = \text{Attn}_t(\text{LN}(\mathbf{Q}_t), \text{LN}(\mathbf{K}_t), \text{LN}(\mathbf{V}_t)) + \text{LN}(\mathbf{h}_{\text{t_self}}^{i-1}) \quad (11)$$

$$\mathbf{h}_{\text{a/v}}^i = \text{Attn}_{\text{a/v}}(\text{LN}(\mathbf{Q}_{\text{a/v}}), \text{LN}(\mathbf{K}_{\text{a/v}}), \text{LN}(\mathbf{V}_{\text{a/v}})) + \text{LN}(\mathbf{h}_{\text{a/v}}^{i-1}) \quad (12)$$

式中: $\mathbf{h}_{\text{t_self}}^i \in \mathbf{R}^{T_m \times d}$; $\mathbf{h}_{\text{a/v}}^i \in \mathbf{R}^{T_m \times d}$; 查询 $\mathbf{Q}_m = \mathbf{h}_m^{i-1} \mathbf{W}_m^Q$, 键 $\mathbf{K}_m = \mathbf{h}_m^{i-1} \mathbf{W}_m^K$, 值 $\mathbf{V}_m = \mathbf{h}_m^{i-1} \mathbf{W}_m^V$ 分别是特定于模态的注意力矩阵; $\text{LN}(\cdot)$ 代表进行层归一化处理; $\mathbf{h}_{\text{a/v}}^i$ 作为下一层协同注意力跨模态 Transformer 层中音频模态和视觉模态的输入。

在每种模态进行模态内的交互作用之后,建模文本模态与音频、视觉模态的交互作用以及音频模态与文本、视觉模态的交互作用。以基于文本模态的协同注意力跨模态 Transformer 为例,将文本模态的表示投影后作为查询,音频模态和视觉模态的表示拼接后作为中间值,将中间值经过投影后获得键和值。此时,能够获得文本模态句子中的每个单词与每一帧音频和视觉特征的相似性关

系,其次,将跨模态注意力的输出送入前馈神经网络中获得协同注意力跨模态 Transformer 层的输出,如式(13)、(14)所示。

$$\begin{aligned} h_{t_cross}^i &= \text{Cross-Attn}_i(\text{LN}(Q_t^{\text{cross}}), \\ &\quad \text{LN}(K_t^{\text{cross}}), \text{LN}(V_t^{\text{cross}})) + \text{LN}(h_{t_self}^i) = \\ &\quad \text{MHA}(\text{LN}(Q_t^{\text{cross}}), \text{LN}(K_t^{\text{cross}}), \text{LN}(V_t^{\text{cross}})) + \\ &\quad \text{LN}(h_{t_self}^i) \end{aligned} \quad (13)$$

$$h_t^i = \text{FFN}(\text{LN}(h_{t_cross}^i)) + \text{LN}(h_{t_cross}^i) \quad (14)$$

式中: $h_{t_cross}^i \in \mathbf{R}^{T_m \times d}$; $h_t^i \in \mathbf{R}^{T_m \times d}$; 查询 $Q_t^{\text{cross}} = h_{t_self}^i W_{t_cross}^Q$, 键 $K_t^{\text{cross}} = [h_a^i, h_v^i] W_{t_cross}^K$, 值 $V_t^{\text{cross}} = [h_a^i, h_v^i] W_{t_cross}^V$ 分别是跨模态注意力投影矩阵。 h_t^i 是基于文本的协同注意力跨模态 Transformer 第 i 层的输出,取 h_t^L 最后一个时间步的向量表示作为基于文本的多模态融合表示。对于基于音频模态的协同注意力跨模态 Transformer,同理可得将 h_a^L 的最后一个时间步的向量表示作为基于音频的协同注意力跨模态 Transformer 的最终输出。

2.3 双模态表示生成模块

在学习到每种模态的私有表示和共享表示之后,为了综合表示单模态表示,将同一模态的 2 种不同表示融合后生成最终的单模态表示,每种单模态表示的输出由式(15)定义,

$$u_m = \text{Sigmoid}(\text{FC}(F_m^{\text{shared}}, F_m^{\text{p}})) \quad (15)$$

式中: $u_m \in \mathbf{R}^d$ 。此时,最终获得的单模态表示融合了同一模态不同方面与情感相关的信息,既包含统一性信息,又同时具备特定于某种模态独特的信息。

为了实现补充基于文本和基于音频的多模态融合表示的信息,设计了双模态表示生成模块,将包含不同方面信息的单模态表示双双进行融合,以获得双模态融合表示,弥补了多模态融合表示只注重学习多模态统一性信息的缺陷,如式(16)所示,

$$\begin{aligned} b_{m_1 m_2} &= \text{MLP}((u_{m_1}, u_{m_2}) \theta^{\text{MLP}}) \\ m_1, m_2 &\in \{a, v, t\}, m_1 \neq m_2 \end{aligned} \quad (16)$$

式中: $b_{m_1 m_2} \in \mathbf{R}^d$; $\text{MLP}(\cdot)$ 表示多层神经网络,由 2 层全连接层分别后接 Leaky ReLU 激活函数和 tanh 激活函数组成; θ^{MLP} 代表其网络参数。

2.4 情感预测模块

将单模态表示、双模态表示和多模态融合表示进行拼接后,得到最终的多模态融合表示,该多模态融合表示同时具备统一性信息以及模态内和模态间的差异化信息,使模型能够学习到更全面

的与情感相关的信息。将拼接后的多模态融合表示输入到两层标准 Transformer 编码器中进行编码以建模不同表示之间的自注意力,再将具有自注意力的融合表示送入多层预测网络进行情感预测,如式(17)、(18)所示。

$$f = \text{TRM}(u_a, \dots, b_{av}, \dots, h_t^L, h_a^L) \quad (17)$$

$$\hat{y} = \text{MLPN}(f, \theta^{\text{MLPN}}) \quad (18)$$

式中: $f \in \mathbf{R}^{8d}$; $\text{MLPN}(\cdot)$ 表示多层预测网络,由 2 层全连接层和一个 ReLU 激活函数组成; θ^{MLPN} 代表其网络参数。

2.5 损失函数

为了确保每种模态的私有表示和共享表示捕获到同一种模态不同方面的信息,同时确保能够学到不同模态之间的差异化信息,采用软正交约束计算同一模态间不同表示的相似性和不同模态表示之间的相似性之和作为差异损失,其定义如式(19)所示,

$$\begin{aligned} L_{\text{diff}} &= \sum_{i=1}^N \left[\sum_{m \in \{t, v, a\}} \| (F_m^{\text{p}})^T F_m^{\text{shared}} \|_F^2 + \right. \\ &\quad \left. \sum_{(m_1, m_2)} \| (F_{m_1}^{\text{p}})^T F_{m_2}^{\text{p}} \|_F^2 \right] \\ m_1, m_2 &\in \{a, v, t\}, m_1 \neq m_2 \end{aligned} \quad (19)$$

式中: $\| \cdot \|_F^2$ 表示弗罗贝尼乌斯范数的平方操作; N 表示训练样本的个数。

对于情感分析中的回归任务,采用均方误差损失作为损失函数,将差异损失加权后作为回归任务中的总体损失函数来衡量情感强度预测的准确性,如式(20)所示,

$$L_1 = \frac{1}{N} \sum_i (y_i - \hat{y})^2 + \gamma L_{\text{diff}} \quad (20)$$

对于分类任务,采用交叉熵损失作为损失函数,将差异损失加权后作为分类任务中的总体损失函数来衡量情感分类的准确性,如式(21)所示,

$$L_2 = -\frac{1}{N} \sum_i y_i \cdot \log(\hat{y}_i) + \gamma L_{\text{diff}} \quad (21)$$

式中: γ 作为超参数,用于调整差异化损失 L_{diff} 在总体损失函数中的权重值。

3 实验及结果分析

本节将对本文所提出的基于 Transformer 的多子空间多模态情感分析方法的有效性进行验证,在 2 个基准数据集上设计并完成多项实验。

3.1 实验设置

1) 基准数据集。为评估本方法的有效性,本文使用由卡耐基梅隆大学公开的多模态意见级情感强度数据集(Multimodal Opinion-level Sentiment Intensity, CMU-MOSI)和多模态意见情感与情绪强度数据集(Multimodal Opinion Sentiment and Emotion Intensity, CMU-MOSEI),数据集中分别包含文本、视觉和音频 3 种模态的数据,2 个数据集的组成与划分如表 1 所示。CMU-MOSI 数据集是一个评论性视频的集合,每条评论数据的情感强度被标注为区间 $[-3,3]$ 内的实数,从-3 到 3 分表代表:强消极、消极、弱消极、中立、弱积极、积极和强积极。CMU-MOSEI 比 CMU-MOSI 的视频数量更多,涵盖的话题范围更广,其情感标签包含二分类、五分类以及七分类的标注,常用的是区间为 $[-3,3]$ 的七分类情感强度标签。除此之外,该

表 1 CMU-MOSI 和 CMU-MOSEI 数据集的组成和划分

Tab. 1 Composition and division of the CMU-MOSI and CMU-MOSEI datasets

数据集	训练集	验证集	测试集	总数
CMU-MOSI	1 284	229	686	2 199
CMU-MOSEI	16 326	1 871	4 659	22 856

3.2 基准方法

LMF^[15]:利用低秩张量进行多模态融合,以提高效率。

MFN^[24]:明确地解释了神经架构中的 2 种交互,并随着时间对其进行建模。

RAVEN^[25]:对非语言子词序列的细粒度结构进行建模,并基于非语言线索动态地调整单词表示。

MuT^[26]:使用基于跨模态注意力的跨模态 Transformer 进行模态翻译。

ICCN^[27]:通过深度典型相关性分析来学习模态之间的相关性。

MAG-BERT^[28]:采用多模态适应门将对齐的非语言信息与文本表示相结合。

MISA^[29]:将各模态表示投影到模态特定和模态不变空间中,并学习分布相似性、正交损失、重构损失以及任务预测损失。

Self-MM^[30]:设计了一个多模态和单模态任务来学习模态间的一致性和模态内的特异性。

MMIM^[20]:分层最大化单模态输入对和多模态融合结果与单模态输入之间的互信息。

Hycon-B^[31]:通过不同种类的对比学习模态

数据集也包含 6 种情绪标签,分别为:开心、悲伤、生气、厌恶、惊喜与恐惧,每种情绪的强度为取值在 $[0,3]$ 的实数。

2) 实现细节。在模态私有与共享表示学习模块中,长短期记忆网络的隐藏状态的维度设置为 768,输出维度为 128。在协同注意力跨模态 Transformer 模块中,每个协同注意力跨模态 Transformer 块包含 4 层协同注意力跨模态 Transformer 层,自注意力的注意力头数设置为 4,基于文本和基于音频的跨模态注意力中的注意力头数都设置为 8,最终输出的多模态表示的维度设置为 128 维。在训练过程中,网络训练采用的是 Adam 优化器, batch_size 的调整范围是 $\{16,32,64\}$,迭代次数 epoch 设置为 50。本文提出的方法使用深度学习框架 Pytorch 实现,在 RTX 3090GPU 上进行模型训练。

间和模态内的交互关系以及样本间和类间关系,从而减小模态差距。

ICDN^[32]:结合跨模态 Transformer 方法和自监督获取单模态情感标签方法,同时学习一致性和差异化信息。

PS-Mixer^[33]:采用基于 MLP-Mixer 的极性向量和强度向量的混合器模型实现不同模态数据之间的通信。

3.3 实验结果分析

本小节将本文所提出的基于 Transformer 的多子空间多模态情感分析方法与基准方法在 CMU-MOSI 和 CMU-MOSEI 这 2 个基准数据集上进行性能比较,并对实验结果进行了多方面的详细分析,实验结果如表 2 和表 3 所示。

对表 2 中的实验结果分析可得,在 CMU-MOSI 数据集上,本文所提方法的实验结果均要高于基线方法。其中,在回归任务中,本文方法在 MAE、Corr 评价指标上得出了 0.705、0.800 的实验结果,相比最好的基准方法,MAE 下降了 0.008,Corr 提升了 0.008;在分类任务上,二分类准确率 Acc-2 和 F1 得分分别为 85.71%/86.63%、85.64%/86.63%,相比最好的基准方法分别提升

了 0.025 6/0.014 3、0.025 2/0.015 3。本方法在各项评价指标上的结果均优于最好的基准模型,验证了本文提出的模型的有效性。其中,“/”左

边是消极与非消极时的二分类结果,“/”右边是消极与积极时的二分类结果。

表2 本方法与基准方法在 CMU-MOSI 数据集上的实验结果

Tab. 2 Experimental results of this method and benchmark method on CMU-MOSI dataset

模型	MAE	Corr	Acc-2/%	F1-score/%	数据状态
LMF	0.917	0.695	- /82.5	- /82.4	非对齐
MFM	0.877	0.706	- /81.7	- /81.6	对齐
RAVEN	0.915	0.691	78.0/ -	76.6/ -	对齐
MuT	0.861	0.711	81.5/84.1	80.6/83.9	对齐
ICCN	0.862	0.714	- /83.0	- /83.0	非对齐
MISA	0.783	0.761	81.8/83.4	81.7/83.6	对齐
MAG-BERT	0.734	0.789	82.42/84.15	82.45/84.13	对齐
Self-MM	0.720	0.792	83.15/84.82	83.12/84.84	非对齐
MMIM	0.719	0.782	82.36/84.99	82.38/84.96	非对齐
Hycon-B	0.713	0.790	- /85.2	- /85.1	非对齐
ICDN	0.886	0.679	- /81.5	- /81.6	非对齐
PS-Mixer	0.794	0.748	80.3/82.1	80.3/82.1	非对齐
MSTFN	0.705	0.800	85.71/86.63	85.64/86.63	非对齐

表3 本方法与基准方法在 CMU-MOSEI 数据集上的实验结果

Tab. 3 Experimental results of this method and benchmark method on CMU-MOSEI dataset

模型	MAE	Corr	Acc-2/%	F1-score/%	数据状态
LMF	0.623	0.677	- /82.0	- /82.1	非对齐
MFM	0.568	0.717	- /84.4	- /84.3	对齐
RAVEN	0.614	0.662	79.1/ -	79.5/ -	对齐
MuT	0.58	0.703	- /82.5	- /82.3	对齐
ICCN	0.565	0.713	- /84.2	- /84.2	非对齐
MISA	0.568	0.724	82.59/84.23	82.67/83.97	对齐
MAG-BERT	0.555	0.758	82.38/85.16	82.07/85.24	对齐
Self-MM	0.533	0.761	81.33/84.63	81.77/84.59	非对齐
MMIM	0.546	0.759	81.04/83.92	81.04/84.22	非对齐
Hycon-B	0.601	0.776	- /85.4	- /85.6	非对齐
ICDN	0.590	0.705	- /81.6	- /81.6	非对齐
PS-Mixer	0.537	0.760	83.1/85.76	83.1/85.77	非对齐
MSTFN	0.537	0.762	83.17/85.99	83.31/85.92	非对齐

表3展示了本方法在 CMU-MOSEI 数据集上的多模态情感分析实验结果。从表3中的数据分析得知,本方法在回归任务的评价指标 MAE 和 Corr 上的结果与相对应的最好的基准方法 SELF-MM 和 Hycon-B 分别相差 0.004 和 0.014;在分类任务中,本方法在二分类准确率 Acc-2 和 F1 得分上均取得了最优的结果:83.17%/85.99% 和 83.31%/85.92%,与最好的基准方法 PS-Mixer 相比分别提

升了 0.000 7/0.002 3 和 0.002 1/0.001 5,进一步证明去除冗余信息后对提升多模态情感分析精度有着重要作用。

综合表 2 和表 3 的实验结果可以得到,本方法在 2 个数据集上均实现了优于其他基线方法的性能,强基线模型 PS-Mixer 仅在较小的数据集 CMU-MOSI 上的实验结果较为突出,证明该模型的鲁棒性较差。本文提出的方法性能不局限于输入数据的类型,具有较高的鲁棒性。与应用传统的三元 Transformer 结构方法如 MuT 和 ICDN 相比,MSTFN 在各项评价指标上的实验结果均远高于这 2 种方法的实验结果,由此可以分析得出,采用二元 Transformer 融合方案的模型能够更好地实现多模态数据的融合,缩小模态间的异质性,提升模型情感分析能力。综合以上分析可以得出,本方法是具有较高性能的多模态情感分析方法。

3.4 消融实验

本文提出的基于 Transformer 的多子空间多模态情感分析方法由 4 个模块组成,其中的主要模块为模态私有与共享表示学习模块、协同注意力跨模态 Transformer 模块以及双模态表示生成模块。为了验证 3 个主要模块的有效性,在非对齐的 CMU-MOSI 数据集上设计了多项消融实验。在完整模型 MSTFN 的基础上,逐步去除各个主要模块生成不同的模型版本,对本模型的各个版本的定义和解释如下。MSTFN w/o PSRLM:去除模态私有与共享表示学习模块的模型,相对应地,在训练阶段只保留基本的任务损失函数,去除差异

化损失函数。此时,模型只能学习到多模态一致性信息而忽略了模态内和模态间的差异化信息。MSTFN w/o CACT:去除协同注意力跨模态 Transformer 模块的模型。3 种模态的共享表示被简单地拼接起来,再经过情感预测模块进行情感分析。这样模态共享表示之间缺乏交互作用,导致不同模态之间仍存在较大的异质鸿沟。MSTFN w/o BRGM:去除双模态表示生成模块的模型。这会导致缺少同一模态内的交互作用,并且缺少对包含不同方面情感信息的单模态表示之间关系的建模,缺少对多模态融合表示的信息补充。

表 4 显示了使用本模型的几种变体所进行的消融实验的结果。由表 4 中的数据可以得出,模态私有与共享表示学习模块以及协同注意力跨模态 Transformer 模块是使本方法达到最优性能的必要条件,协同注意力跨模态 Transformer 模块对本方法的贡献度最高。模态私有与共享表示在建模多模态统一性信息的同时,学习了模态内和模态间的差异化信息,对多模态融合信息进行了补充。协同注意力跨模态 Transformer 模块同时建模文本和视觉、音频模态以及音频和视觉、文本模态之间的交互作用,保证了多模态融合表示包含足够的与情感相关的信息,促进了多模态数据进一步融合。同时,相比于传统的三元对称的跨模态 Transformer 结构,二元跨模态 Transformer 结构能够在保证减小模态间异质性的前提下减少冗余信息。

表 4 在 CMU-MOSI 数据集上进行的消融实验
Tab. 4 Ablation experiments on CMU-MOSI datasets

	MAE	Corr	Acc-2/ %	F1-score/ %
MSTFN	0. 705	0. 800	85. 71/86. 74	85. 64/86. 63
MSTFN w/o PSRLM	0. 752	0. 785	82. 94/85. 06	82. 85/85. 04
MSTFN w/o CACT	0. 749	0. 780	81. 53/83. 66	81. 63/83. 69
MSTFN w/o BRGM	0. 752	0. 791	82. 65/85. 21	82. 40/85. 07

在协同注意力跨模态 Transformer 模块中,本方法采用基于文本和基于音频模态的多模态融合表示,为了验证基于不同模态的协同注意力跨模态 Transformer 模块对整个模型的影响,除了使用文本和音频模态作为目标模态的完整模型上的实验之外,分别在 CMU-MOSI 数据集上进行了 1 种目标模态、2 种目标模态以及 3 种目标模态的实验,实验结果如表 5 所示。只有 1 种目标模态的

情况下,分别以 a、v、t 为目标模态,利用 1 个单一的协同注意力跨模态 Transformer 获得音频与视觉和文本的交互[$a \rightarrow t + v$]、视觉与音频和文本的交互[$v \rightarrow t + a$]以及文本与音频和视觉的交互[$t \rightarrow a + v$]。其次,对于 2 种目标模态的情况,同时采用 2 个协同注意力跨模态 Transformer,获得基于 2 种模态的多模态表示:[$a \rightarrow t + v$]和[$v \rightarrow t + a$]、[$a \rightarrow t + v$]和[$t \rightarrow a + v$]、[$v \rightarrow t + a$]和[$t \rightarrow a +$

v]。最后,在有3种目标模态时,构建3个协同注意力跨模态 Transformer 模块以同时建模基于文本、音频和视觉模态的多模态融合表示,即 $[t \rightarrow a + v]$ 、 $[a \rightarrow t + v]$ 和 $[v \rightarrow t + a]$ 。由表5中的实验结果可以得出,在只有1种目标模态时,基于音频的协同注意力跨模态 Transformer 的性能最优,文本仅次于音频模态,二者之间的差异很小且结果均高于单独的视觉模态,说明基于具有语义信息

的音频、文本模态更能够提升模态间的交互作用。同理,在2种目标模态的情况下,本方法中使用基于文本和音频的协同注意力跨模态 Transformer 的结果最高。在3种目标模态的情况下,虽然对多种模态之间进行了复杂且全面的交互作用的建模,但是由于存在过多冗余信息,导致模型的性能降低,分类准确率以及预测精度甚至低于单独的文本和音频模态。

表5 不同数量的目标模态对模型性能的影响
Tab.5 The effect of different numbers of target modes on model performance

目标模态数量	模态交互	Acc-2/%	F1-score/%	MAE	Corr
1种目标模态 (本方法)	$[t \rightarrow a + v]$	82.80/84.91	82.75/84.81	0.726	0.788
	$[a \rightarrow t + v]$	83.38/85.06	83.27/85.00	0.723	0.789
	$[v \rightarrow t + a]$	82.51/84.45	82.35/84.36	0.737	0.779
2种目标模态 (本方法)	$[t \rightarrow a + v]$	85.71/86.74	85.64/86.63	0.705	0.800
	$[a \rightarrow t + v]$				
	$[a \rightarrow t + v]$	83.24/85.37	83.08/85.28	0.728	0.790
	$[v \rightarrow t + a]$				
	$[v \rightarrow t + a]$	83.24/85.52	82.95/85.32	0.712	0.798
	$[t \rightarrow a + v]$				
3种目标模态 (本方法)	$[t \rightarrow a + v]$	82.36/84.91	82.12/84.77	0.7406	0.787
	$[v \rightarrow t + a]$				
	$[a \rightarrow t + v]$				

4 结语

本文提出了一种通过将不同模态的初始特征表示映射到不同子空间中以获得多模态统一信息并学习模态内和模态间差异化信息的方法,称为基于 Transformer 的多子空间多模态情感分析方法。该方法通过构建多模态的共享子空间和每种模态的私有子空间,获得具有统一信息的共享表示和包含模态内与模态间的差异化信息的私有表示,摒弃传统的三元对称 Transformer 融合方案,设计了二元的协同注意力跨模态 Transformer 模块对共享表示进行融合,同时建模1种模态与其余2种模态之间的交互作用,生成双模态表示,对多模态融合表示进行信息补充。与以往的工作相比,本方法在2个多模态情感分析数据集中取得了具有竞争力的结果,这也有力地证明了模态内和模态间的差异化信息对多模态数据中包含的统一性信息在情感分析中能够起到信息补充的作用,同时,验证了二元 Transformer 融合方案能够

去除冗余信息,以此提升情感分析的准确率。更进一步地,设计多个消融实验证明每个模块的有效性,同时验证了具有不同目标模态的协同注意力跨模态 Transformer 模块对本方法的性能影响。最后,对超参数对整体损失函数的贡献度进行分析,并从模型复杂度的角度进一步评估和分析本方法。

参考文献

[1] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: From unimodal analysis to multimodal fusion [J]. Information Fusion, 2017, 37 (C): 98-125.

[2] ZHANG Y, JIN R, ZHOU Z H. Understanding bag-of-words model: A statistical framework [J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1): 43-52.

[3] LI B F, LIU T, ZHAO Z, et al. Neural bag-of-ngrams [C] // Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: ACM, 2017: 3067-3074.

- [4] CHEN P H, LIN C J, SCHÖLKOPF B. A tutorial on ν -support vector machines [J]. *Applied Stochastic Models in Business and Industry*, 2005, 21(2): 111-136.
- [5] RISH I. An empirical study of the naive Bayes classifier[J]. *Journal of Universal Computer Science*, 2001, 1(2): 41-46.
- [6] ALBAWI S, ABED MOHAMMED T A, AL-ZAWI S. Understanding of a convolutional neural network[C]//2017 International Conference on Engineering and Technology (ICET). Antalya: IEEE, 2017: 1-6.
- [7] MALHOTRA P, VIG L, SHROFF G, et al. Long short term memory networks for anomaly detection in time series[C]//23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges: ESANN, 2015.
- [8] SONG K C, YAN Y H, CHEN W H, et al. Research and perspective on local binary pattern[J]. *Acta Automatica Sinica*, 2013, 39(6): 730-744.
- [9] WANG Z, YING Z L. Facial expression recognition based on local phase quantization and sparse representation[C]//2012 8th International Conference on Natural Computation. Chongqing: IEEE, 2012: 222-225.
- [10] KAMARAINEN J K. Gabor features in image analysis [C]//2012 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA). Istanbul: IEEE, 2012: 13-14.
- [11] HAN W, CHAN C F, CHOY C S, et al. An efficient MFCC extraction method in speech recognition[C]//2006 IEEE International Symposium on Circuits and Systems (ISCAS). Kos: IEEE, 2006: 4pp.
- [12] PORIA S, CHATURVEDI I, CAMBRIA E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). Barcelona: IEEE, 2016: 439-448.
- [13] KAMPMAN O, BAREZI E J, BERTERO D, et al. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne: Association for Computational Linguistics, 2018: 606-611.
- [14] 罗渊瀚, 吴锐, 刘家锋, 等. 基于自适应权值融合的多模态情感分析方法[J/OL]. *软件学报*. (2023-10-07) [2023-11-20]. <https://doi.org/10.13328/j.cnki.jos.006998>.
- LUO Y Y, WU R, LIU J F, et al. Multimodal sentiment analysis based on adaptive weight fusion[J/OL]. *Journal of Software*. (2023-10-07) [2023-11-20]. <https://doi.org/10.13328/j.cnki.jos.006998>.
- [15] ZADEH A, CHEN M H, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 1103-1114.
- [16] 张涛, 郭青冰, 李祖贺, 等. MC-CA: 基于模态时序耦合与交互式多头注意力的多模态情感分析[J]. *重庆邮电大学学报(自然科学版)*, 2023, 35(4): 680-687.
- ZHANG T, GUO Q B, LI Z H, et al. MC-CA: Multimodal sentiment analysis based on modal temporal coupling and interactive multi-head attention[J]. *Journal of Chongqing University of Posts & Telecommunications (Natural Science Edition)*, 2023, 35(4): 680-687.
- [17] 陈宏松, 安俊秀, 陶全桢, 等. 基于 BERT-VGG16 的多模态情感分析模型[J]. *成都信息工程大学学报*, 2022, 37(4): 379-385.
- CHEN H S, AN J X, TAO Q H, et al. Multi-modal sentiment analysis model based on BERT-VGG16[J]. *Journal of Chengdu University of Information Technology*, 2022, 37(4): 379-385.
- [18] 周柏男, 李旭, 范丰龙, 等. 基于交互注意力机制的多模态情感分析模型[J]. *大连工业大学学报*, 2023, 42(5): 378-384.
- ZHOU B N, LI X, FAN F L, et al. Multi-modal sentiment analysis model based on interactive attention mechanism[J]. *Journal of Dalian Polytechnic University*, 2023, 42(5): 378-384.
- [19] 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.
- LU C, GUO J J, TAN K W, et al. Multimodal sentiment analysis based on text-guided hierarchical adaptive fusion[J]. *Journal of Shandong University (Natural Science)*, 2023, 58(12): 31-40.
- [20] 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. Punta Cana: Association for Computational Linguistics, 2021: 9180-9192.
- [21] VASWANI A, SHAZEER N, PARMAR N, et al. At-

- tention is all you need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017: 6000-6010.
- [22] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2023-11-20]. <http://arxiv.org/abs/1810.04805>.
- [23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03)[2023-11-20]. <http://arxiv.org/abs/2010.11929>.
- [24] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[C] // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans: ACM, 2018: 5634-5641.
- [25] WANG Y S, SHEN Y, LIU Z, et al. Words can shift: Dynamically adjusting word representations using non-verbal behaviors[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 7216-7223.
- [26] TSAI Y H H, BAI S J, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[J]. Proceedings of the Conference Association for Computational Linguistics Meeting, 2019, 2019: 6558-6569.
- [27] SUN Z K, SARMA P, SETHARES W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8992-8999.
- [28] RAHMAN W, HASAN M K, LEE S W, et al. Integrating multimodal information in large pretrained transformers[J]. Proceedings of the Conference Association for Computational Linguistics Meeting, 2020, 2020: 2359-2369.
- [29] 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.
- [30] YU W M, XU H, YUAN Z Q, 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.
- [31] MAI S J, ZENG Y, ZHENG S J, et al. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing, 2023, 14(3): 2276-2289.
- [32] ZHANG Q A, SHI L, LIU P Y, et al. RETRACTED ARTICLE; ICDN: Integrating consistency and difference networks by transformer for multimodal sentiment analysis[J]. Applied Intelligence, 2023, 53(12): 16332-16345.
- [33] LIN H, ZHANG P L, LING J D, et al. PS-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis[J]. Information Processing & Management, 2023, 60(2): 103229.

(编辑 李静)