计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(8):62−69 [doi: 10.15888/j.cnki.csa.009923] [CSTR: 32024.14.csa.009923]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

大模型优化的 BERT 图文多模态情感分析①

杨宇飞[1,2], 钱育蓉[1,2,3,4], 公维军[1,2], 侯树祥[2,4], 路　燚[1,2], 陈嘉颖[1,2,3]

[1](新疆大学 软件学院 新疆大数据与智能软件工程研究中心, 乌鲁木齐 830091)
[2](新疆大学 软件工程重点实验室, 乌鲁木齐 830091)
[3](新疆大学 丝路多语言认知计算国际合作联合实验室, 乌鲁木齐 830046)
[4](新疆大学 计算机科学与技术学院, 乌鲁木齐 830046)
通信作者: 钱育蓉, E-mail: qyr@xju.edu.cn

摘　要: 方面级多模态情感分析属于情感分析以及观点挖掘方向的一个子领域, 旨在方面或属性级别开展情感和观点的分析. 在过去的图文多模态情感分析研究里, 研究者针对如何从图像和文本中提取并融合特征提出了各种方法, 由于图文信息初始所在的语义空间不一致, 最常用的方法是先从单模态中通过模块提取相应的深层信息, 将图像和文本特征映射到一个相同的深层语义空间中, 随后使用特征融合模块进行图文特征融合. 然而, 这种方法势必会引入多个模块用于处理图像和文本的特征并进行融合, 这不可避免增加了模型的参数量和复杂度. 随着如今大模型的发展, 在浅层空间将图像和文本的语义空间进行对齐已成为可能. 本研究利用通义千问开源大模型, 在预处理阶段通过提示词生成图像的文本描述, 让多模态情感分析回归到单模态情感分析任务, 仅通过文本处理模块就能得到最终的图文情感分析结果. 实验表明, 与先前的大多数模型相比, 该方法显著降低了参数量且取得了性能上的提升. 与同样轻量级的 TISRI 模型相比, 该模型在训练速度和资源占用上均取得了显著优势. 具体代码实现参考 https://github.com/triangleXIV/ITFFT.

关键词: 情感分析; 多模态; 大模型; 图像处理; 特征融合; Transformer

引用格式: 杨宇飞,钱育蓉,公维军,侯树祥,路燚,陈嘉颖.大模型优化的 BERT 图文多模态情感分析.计算机系统应用,2025,34(8):62–69. http://www.c-s-a.org.cn/1003-3254/9923.html

LLM-optimized BERT for Image-text Multimodal Sentiment Analysis

YANG Yu-Fei[1,2], QIAN Yu-Rong[1,2,3,4], GONG Wei-Jun[1,2], HOU Shu-Xiang[2,4], LU Yi[1,2], CHEN Jia-Ying[1,2,3]

[1](Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of Software, Xinjiang University, Urumqi 830091, China)
[2](Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China)
[3](Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Xinjiang University, Urumqi 830046, China)
[4](School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China)

Abstract: Aspect-level multimodal sentiment analysis involves analyzing sentiment and opinions at the aspect or attribute level. Previous studies on image-text multimodal sentiment analysis have explored various methods for extracting and fusing features from images and text. Since the initial semantic spaces of images and text are not aligned, common approaches extract deep features from each modality, map them to a shared deep semantic space, and then apply a fusion module. However, this inevitably increases model complexity. With advancements in large language models, aligning the semantic spaces of images and text at a shallow level has become feasible. This study leverages Qwen to generate textual descriptions of images through prompt engineering during preprocessing, transforming multimodal sentiment analysis into a unimodal sentiment analysis task. This enables sentiment analysis results to be obtained using only a text processing

① 基金项目: 自治区重点研发专项 (2023B01029-1, 2023B01029-2); 国家自然科学基金 (62266043); 自治区杰出青年科学基金 (2023D01E01); 自治区青年拔尖人才项目 (2023TSYCCX0043); 天山创新团队计划 (2023D14012)
收稿时间: 2024-12-31; 修改时间: 2025-02-12; 采用时间: 2025-02-24; csa 在线出版时间: 2025-06-27
CNKI 网络首发时间: 2025-06-30

module. Experimental results show that, compared to most previous models, the proposed method significantly reduces the number of parameters while achieving performance improvements. Compared to the similarly lightweight TISRI model, it also demonstrates notable advantages in training speed and resource utilization. The code is available at: https://github.com/triangleXIV/ITFFT.

Key words: sentiment analysis; multimodal; large language model (LLM); image processing; feature fusion; Transformer

根据 2024 年的《中国互联网络发展状况统计报告》[1]显示, 中国有超过 10 亿网民, 所使用的社交平台呈现的数据为多模态形式, 如文字、图片等. 方面级多模态情感分析 (图 1) 旨在利用此类多模态数据开展方面或属性级情感和观点的分析, 达成舆情监控[2]、商业分析[3]、突发事件响应[4]等目的.

文本信息	Eddie and the faithful Pearl Jam fans in Buenos Aires. Photo by @ epozzoni # PJSA2013
分析方面/目标	Pearl Jam fans
分析结果	Positive

图 1 给定一段文本与图像, 分析某一目标情感表达

在过去研究中, 针对多模态情感分析的研究集中在如何精确提取各自模态的信息并利用模态间的信息互补提升情感分析结果的准确性. 例如在对 Twitter 数据集进行情感分析时, 多数研究通常使用 BERT 或其变体模型对文本特征进行提取[5], 使用 vision Transformer (ViT) 或更先进的图像特征提取模型进行图像特征提取, 随后将这些深层特征传入融合模块, 采用诸如交叉注意力、自注意力、门控网络和对比学习等方法, 学习各个模态间的交互信息[6], 输出最终的情感分析结果[7].

然而, 先前的方法也存在一定的局限性: 1) 过多模块的引入使得训练时调整的参数不可避免的增大, 这增加了训练成本. 2) 特征提取过程中无法保证将图像和文本通过不同的模块映射到相同的语义空间, 这增加了图文特征信息融合的难度. 3) 图文特征融合过程可能存在另一模态的冗余信息, 如何去除冗余信息识别另一模态的关键要素也是一大挑战.

大模型的出现让复杂问题简单化成为可能[8], 凭借强大的特征提取和文本生成能力, 依靠提示词将图像转换为相应的文本描述, 在初期就能对齐图像和文本的语义空间, 这不仅降低了模型的复杂性, 还减少了计算开销.

综上, 本文的贡献有以下 3 点.

1) 利用大模型的提示词工程对图像进行数据预处理, 提出一种轻量化的多模态情感分析模型, 图像到文本描述的特征融合 Transformer (image-to-text feature fusion Transformer, ITFFT), 相较于先前模型显著降低了模型的复杂度和计算开销.

2) 借助大模型对图文的 Twitter-15 和 Twitter-17 数据集进行处理, 提供了一套人工校对后的纯文本方面级情感分析数据集供以后的文本分析研究使用.

3) 在最终决策时采用门控网络结合激活函数控制信息的输出, 这有助于过滤掉文本中的冗余信息, 保留最相关的部分, 同时提升模型的非线性表达能力.

1 相关工作

方面级多模态情感分析旨在结合多种模态信息提升情感分析的准确性. 例如用户发帖的文本内容为: “今天的天气真好啊!”, 可能无法从中判别用户真实的情感表达, 而倘若这句话配上一个倾盆大雨的图片, 则可以知道用户是在说反话, 整个帖子的内容传递出了负面情绪. 在过去的研究中, UNITER[9]通过在输入初期将图像和文本拼接输入编码-解码器中, 输出最终的图文分析结果, 展示了拼接特征的可行性[10]. MMBT[11]在此基础上, 使用双流式 Transformer 架构, 进一步增强了模型在多模态任务中的表现. TomBERT[12]使用交叉注意力机制, 建立图文特征的联系, 通过自注意力去除图文之间的噪声干扰. HIMT[13]更进一步, 通过两种方式融合不同特征并通过交叉注意力建立两种融合方式的联系. 然而, 这些方式依然存在一些问题: 1) 图像和文本的原始语义空间并不相同, 因此需要通过特征提取模块, 将图像和文本信息映射到相同的语义空间; 2) 虽然文本分析目前统一采用 Transformer 进行特征提取, 然而对于图像分析的方法却各不相同, 从早期的卷积到现在的 Transformer, 从早期单纯的特征提取到图像中相关元素的识别, 由于图像和文本初始的空间

维度不同, 如何让二维的图像对齐一维的文本空间一直存在挑战. 3) 当前特征的融合方式从简单走向复杂, 以获得更好的融合表现, 这增加了模型的复杂度, 过多模块的耦合也降低了模型的可扩展性.

GPT 在众多 NLP 任务中无需微调便取得了良好成绩[14], 显示了大模型的巨大潜力. 借助大模型强大的理解能力, 利用提示词为图像生成相应的文本描述也愈发准确, 本文思路便是借助多模态大模型, 在初期就将图像和文本的语义空间对齐, 让多模态情感分析回归为一个单模态任务, 这会显著降低模型的复杂度.

此外, 大模型近几年也贡献出许多实用的技术. 例如, 旋转位置编码[15]用以表示文本的相对位置, 增强模型在长距离依赖任务中的表现, 在 LLaMA[16]、Qwen[17]、PaLM[18]等大模型中被广泛使用, 本模型也将引入相对位置信息[19]用于增强模型在情感分析中的表现; 极端量化[20]的出现降低了大模型的资源消耗, 让大模型在资源受限的下游任务中进行部署成为可能. 本文借助先前研究的技术, 构建了一个在初始阶段将图像和文本对齐, 进行情感分析的模型框架.

2 本文方法

本文模型整体架构如图 2 所示, Qwen2-VL 将文本编码, 将图像通过 ViT 进行特征提取后, 传入内部的 Qwen2 大模型. 在预处理模块读取图像, 通过提示词将图像转换为文本的描述, 将生成的文本与原始的文本信息拼接, 将方面描述的文本通过[SEP]分隔, 表示两段相关联的话. 随后传入一个 12 层的 Transformer 结构进行文本特征提取, 最后将提取得到的特征通过 SiLU 激活函数激活, 通过门控结构筛选信息, 通过线性层输出最终的情感类别.

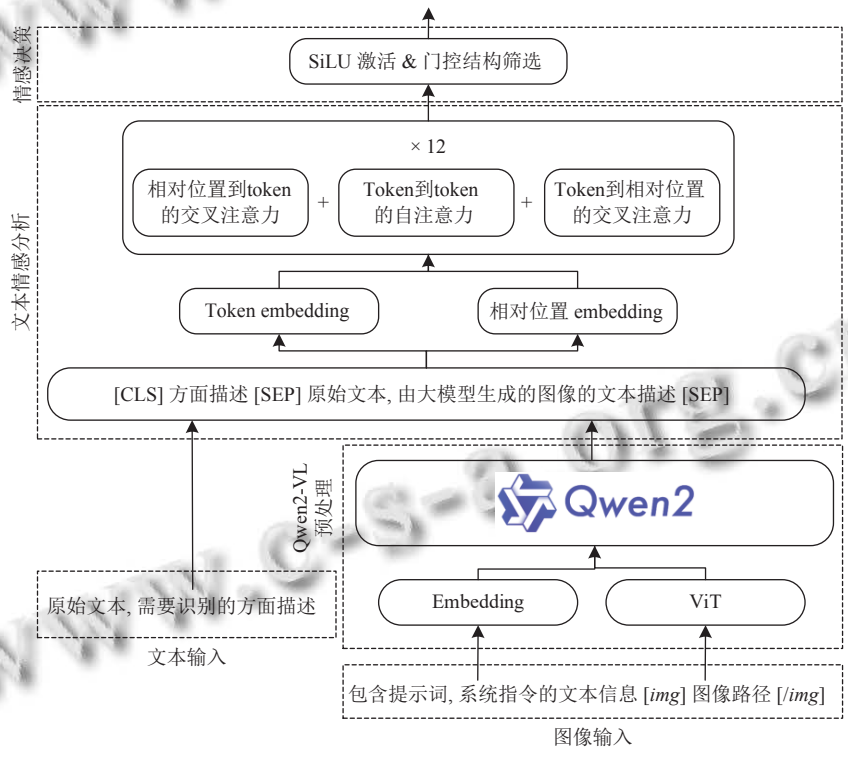

图 2 模型整体结构

2.1 预处理

预处理部分采用的模型为 Qwen2-VL[21], 它使用 ViT 进行图像特征提取, ViT 将图像划分为 14×14 的图像块, 将每个图像块视为一个 token, 传入 Transformer 结构进行特征提取.

Qwen2-VL 具体流程为: 1) 传入图像路径和相应的提示词; 2) 读取图像路径并通过 ViT 进行编码得到 V_I; 3) 对于提示词和系统指令, 则直接传入 embedding 层进行编码得到 V_T; 4) 随后将经过初步特征提取的图像特征与文本经过 embedding 后的特征进行拼接得到 $V_{I,T}$; 5) 传入 Qwen2-VL 内部的 Qwen2 中, 完成图像生成文本描述的任务输出 $D_{I,T}$ (输入输出示例如图 3 所

示). 公式表达如下:

$$Input = (prompt, [img]img\_path[/img]) \tag{1}$$

$$V_I = ViT(img\_path) \tag{2}$$

$$V_T = Embedding(prompt) \tag{3}$$

$$V_{I,T} = Concat(V_I, V_T) \tag{4}$$

$$D_{I,T} = Qwen2(V_{I,T}) \tag{5}$$

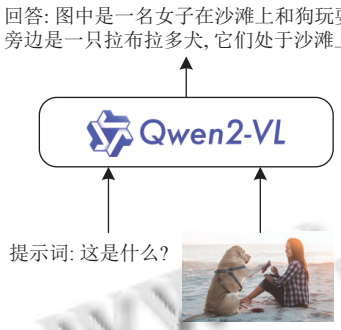

图 3 Qwen2-VL 输入输出示例

2.2 情感分析模块

传统的 BERT 模型为保存每个词之间的信息, 通常会引入一个绝对位置编码添加到经过 embedding 后的句子中. 然而相对位置信息在文本中同样重要, 例如“深度”和“学习”两个词, 如果在文章中距离很远, 则这两个词表示两个不同的意思; 倘若这两个词相邻, 则表示一个专有名词. 本模型使用预训练后的相对位置编码进行微调以适应下游任务, 例如第 1 个 token 到第 0 个 token 的距离用–1 表示; 到第 2 个 token 的距离用 1 表示. 若传入了 K 个 token, 则最终可以用 $K\times K$ 大小的矩阵表示每个 token 间的相对位置, 将该矩阵经过 embedding 层编码后, 进行注意力的计算以提取文本特征. 注意力计算主要分为 3 个部分: 文本与文本的自注意力分数A_{t2t}、文本与相对位置的交叉注意力分数A_{t2p}以及相对位置、文本的交叉注意力分数A_{p2t}. E_{input}表示文本 token 经过 embedding 层后的特征, P_{input}表示相对位置矩阵经过 embedding 层后特征, W_1, W_2, W_3表示 3 种不同的线性变换. Q, K, V表示输出的E_{input}和P_{input}经过线性变换的结果, 下标 t2t、t2p、p2t 分别表示这个结果和文本的自注意力、文本与相对位置的交叉注意力、相对位置与文本的交叉注意力有关. d_k表示缩放系数, 通过注意力机制提取特征的过程如下:

$$Q_{t2t} = E_{input}W_1, K_{t2t} = E_{input}W_2, V_{t2t} = E_{input}W_3 \tag{6}$$

$$A_{t2t} = Softmax\left(\frac{Q_{t2t}K_{t2t}^{T}}{\sqrt{d_k}}\right)V_{t2t} \tag{7}$$

$$Q_{t2p} = E_{input}W_1, K_{t2p} = P_{input}W_2, V_{t2p} = P_{input}W_3 \tag{8}$$

$$A_{t2p} = Softmax\left(\frac{Q_{t2p}K_{t2p}^{T}}{\sqrt{d_k}}\right)V_{t2p} \tag{9}$$

$$Q_{p2t} = P_{input}W_1, K_{p2t} = E_{input}W_2, V_{p2t} = E_{input}W_3 \tag{10}$$

$$A_{p2t} = Softmax\left(\frac{Q_{p2t}K_{p2t}^{T}}{\sqrt{d_k}}\right)V_{p2t} \tag{11}$$

r_p表示一个相对位置映射矩阵, 用以将二维的相对位置注意力结果A_{p2t}和A_{t2p}, 映射到一维的向量空间得到A'_{p2t}和A'_{t2p}, 和文本的自注意力结果A_{t2t}对齐, 得到最终的注意力结果A_{final}.

$$A'_{p2t} = r_p \cdot A_{p2t} \tag{12}$$

$$A'_{t2p} = r_p \cdot A_{t2p} \tag{13}$$

$$A_{final} = A_{t2t} + A'_{t2p} + A'_{p2t} \tag{14}$$

2.3 决策模块

为增强模型的非线性表达能力并进行特征筛选, 该模型在决策层采用门控结构结合激活函数的方法. 将经过注意力与前馈神经网络进行特征提取后的特征表示为 X, W 和 b 表示线性变换的权重和偏置项, 输出为 y, 则决策层可以表达为:

$$y = Dropout(SwiGLU(X \cdot W_1 + b_1) \cdot W_2 + b_2) \tag{15}$$

其中, $SwiGLU$ 是一个激活函数结合门控网络的组合结构. 其过程为, 将 X 经过第 1 次线性变换得到的 X_1 在最后一个维度平等拆分为$X_{1,1}$和$X_{1,2}$, 通过 $SiLU$ 函数激活, 将激活后的结果通过哈达玛积运算实现门控结构, 得到最终的输出$output$. 将输出结果传入一个线性层实现最终的情感分类.

$$X_1 = [X_{1,1}, X_{1,2}] \tag{16}$$

$$X'_{1,1} = SiLU(X_{1,1}) \tag{17}$$

$$output = X'_{1,1} \odot X_{1,2} \tag{18}$$

3 实验

3.1 数据集

为验证本文模型的性能, 在两个多模态情感分析

基准数据集上进行了实验, 数据集介绍如下.

Twitter-15 和 Twitter-17 数据集收集自社交媒体 Twitter 平台的英文推文, 包含了 4 个内容: 文字内容、图片内容、实体/方面和标签 (3 分类).

Twitter-15 包含 5 338 条带图片推文, 数据集划分及数量为训练集 (3 179)、验证集 (1 122)、测试集 (1 037).

Twitter-17 包含 5 972 条带图片推文, 数据集划分及数量为训练集 (3 562)、验证集 (1 176)、测试集 (1 234).

3.2 参数设置与评价指标

如表 1 所示, 本研究 Batch_size 大小设置为 48, 预处理采用的模型为 Qwen2-VL 的 INT8 量化版, 参数量为 7B, Data_type 为 FP16 以加速训练, Prompt 用以识别图像中的关键要素并分析图像中的表情、动作等潜在情感要素, 同时限制输出的文本描述长度, 让输出的结果更精炼. Epoch 统一设置为 6 个周期. Max_token 为 192, 多余的 token 直接截断. Learning_rate 为 3E–5, 前 10% 的 steps 学习率从 0 匀速增加到 3E–5, 后 90% 的 steps 学习率从 3E–5 余弦退火到 0.

表 1 参数设置介绍

参数名称	值
LLM	Qwen2-VL
Batch_size	48
Max_token	192
Data_type	FP16
Epoch	6
Learning_rate	3E–5
Prompt	Identify the key elements in the image and describe each element's specific emotions, actions, or expressions in a single paragraph of no more than 50 words.

评价指标为准确率 (Acc) 和 $F1$ 值, Acc 是预测正确的样本在总样本中的比例, $F1$ 是通过准确率 (*Precision*) 和召回率 (*Recall*) 两个指标共同决定的, 具体公式如下:

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{21}$$

其中, TP 表示模型将实际为正类的样本正确预测为正类的数量; FP 表示模型将实际为负类的样本错误预测为正类的数量; FN 表示模型未能识别出的正类样本.

3.3 基线模型对比

BERT[22]: 2019 年提出, 基于 Transformer 架构的双向编码器. 通过自注意力机制和前馈神经网络层, 双向编码上下文信息, 并使用[CLS]标记的表示进行情感分类.

RoBERTa[23]: 2021 年提出, BERT 的改进版本, 通过去除 NSP 任务, 优化了预训练策略, 专注于单句的掩蔽语言模型训练.

DeBERTa[19,24]: 2020 年提出, 至今已迭代 3 个版本, BERT 的进一步改进版本, 通过结合相对和绝对位置嵌入增强了预训练时解码器的能力, 从而提升了模型的表现力.

MIMN[25]: 2019 年提出, 将不同模态的向量表示存储在独立的记忆矩阵中, 使用了交互式记忆模块. 经过多轮记忆更新后, 得到一个融合多模态信息的全局表示.

TomBERT[12]: 2019 年提出, 通过 BERT 模型处理文本, 使用 ResNet 提取图像特征. BERT 的文本表示和 ResNet 的图像表示通过自注意力和交叉注意力机制进行融合.

HIMT[13]: 2023 年提出, 使用交互式 Transformer 层捕捉不同模态表示之间的关系. 最后通过全局 Transformer 层处理, 得到全局情感特征表示.

TISRI[26]: 2024 年提出, 模型通过权值共享并引入多模态特征相关性识别、图像门控机制和文本-图像交互中的注意力机制, 有效过滤无关图像信息, 提升了面向方面的多模态情感分析性能.

如表 2 和表 3 所示, 引入相对位置的 DeBERTa 显著优于传统的 BERT 模型, 此外多数多模态模型效果也优于单模态模型, 说明了相对位置和图像在辅助文本进行决策时的重要性. 最后, 本文模型 ITFFT 以最低的参数量, 实现了最优的结果, 说明了借助大模型优化多模态情感分析任务的可行性.

表 2 Twitter-15 数据集效果对比

模态	模型名称	Acc (%)	$F1$ (%)	参数量 (M)
文本	BERT	74.15	68.86	109
	RoBERTa	76.54	71.31	124
	DeBERTa	77.81	73.54	179
文本+图像	MIMN	71.84	65.96	259
	TomBERT	77.03	72.85	235
	HIMT	78.14	73.68	303
	TISRI	78.50	74.42	198
	ITFFT	**79.17**	**75.08**	**190**

3.4 参数调整

参数量的调整主要涉及句子长度、模型参数量, 我们通过在 Twitter-17 数据集上实验以观察模型在不

同规模和不同长度输入信息下的表现.

通过表 4 可以发现, 随着模型层数的增加, 模型的 Acc 从 72.17% 增加到 72.9% 附近. 若是将输入的 token 长度从 192 增加到 256, 虽然模型的准确率有了小幅提升, 但 $F1$ 却出现大幅下降. $F1$ 较低说明模型在某一类别的表现不够好, 也就是模型的判断变得极端. 综合考虑下, 使用 12 层 Transformer 并将 token 长度限制为 192 是一个兼顾性能和效率的方案.

表 3 Twitter-17 数据集效果对比

模态	模型名称	Acc (%)	$F1$ (%)	参数量 (M)
文本	BERT	68.15	65.23	109
	RoBERTa	70.11	68.55	124
	DeBERTa	71.51	69.54	179
文本+图像	MIMN	65.88	62.99	259
	TomBERT	69.77	67.59	235
	HIMT	71.14	69.16	303
	TISRI	71.98	71.20	198
	ITFFT	**72.97**	**71.57**	**190**

表 4 参数调整

层数	token输入长度	Acc (%)	$F1$ (%)	对应参数量 (M)
6	192	72.17	70.79	148
12	192	72.97	71.57	190
24	192	72.93	72.24	448
12	128	72.34	71.17	190
12	256	73.07	70.50	190

3.5 训练效率对比

从表 2、表 3 可知, 本文模型相较于先前大多数模型在参数量上有明显下降, 持平于 2024 年的 TISRI, 且 Acc 和 $F1$ 两个指标均有一定的提升, 本节将对比同样轻量的 TISRI 模型用于说明本文模型在训练效率和资源占用上的优势.

虽然 TISRI 相较于先前模型在参数量和性能指标上均有一定的优势, 但其仍没有摆脱传统多模态情感分析框架的束缚. TISRI 在特征处理上, 不同于先前模型实例化 3 个模块分别用于处理图像、文本和方面信息, 而是实例化 2 个模块处理这 3 种的信息, 其中一个模块调用一次用于处理图像信息, 另一个模块调用两次分别处理文本和方面信息, 最后将处理的特征通过其提出的特征融合模块加以融合得出情感分析结果, 实例化模块的减少使得 TISRI 参数的下降.

基于此, 我们在训练批次大小设置为 16 的情况下开展实验用于对比本文模型与 TISRI 在训练过程中的差异. 值得注意的是, TISRI 原生采用了 FP32 进行训练, 而本文模型采用 FP16 进行训练, 为了控制变量, 我们修改了 TISRI 代码, 添加 AMP (automatic mixed precision, 自动混合精度) 库的调用使其支持 FP16 训练方式, 得到了表 5 所示实验结果.

表 5 训练效率对比

模型	精度	模型显存 (GB)	训练显存 (GB)	每秒训练样本数
TISRI	FP32	1.2	10.4	56
ITFFT	FP32	1.1	8.5	103
TISRI	FP16	1.2	8	106
ITFFT	FP16	1.1	6.6	184

如表 5 所示, 本文模型在相同精度训练时的显存占用和训练速度均显著优于 TISRI 模型, 我们认为这种差异主要源于模型层数的不同. 将数据集通过大模型进行预处理并保存后, 本文模型只需要在训练时使用 12 层的 Transformer 用于提取文本信息, 再通过一层决策层即可输出结果. 而 TISRI 模型则需要使用 12 层的 Transformer 分别提取文本和方面信息, 通过 152 层的 ResNet 提取图像信息, 通过 8 层 Transformer 融合各个特征, 最后通过一个决策层输出结果. 过多的层数需要更多的显存来存储层与层之间的中间激活值和梯度信息, 这显著增加了计算量.

3.6 消融实验与局限性

表 2 和表 3 展示了模型在纯文本和文本+图文输入的表现. 本节我们将在 Twitter-17 数据集进行多次实验以计算模型每种修改方案的 t 值和 p 值, 用以分析每一处模块修改是否会对效果产生影响.

通常 p 值在 0.05 以下说明假设成立, t 值越大越能说明差异的显著, 具体公式如下所示:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{22}$$

$$p = P(T > |t|) \tag{23}$$

其中, $\bar{x}$是样本均值, μ_0是零假设中的总体均值, s 是样本标准差, n 是样本容量, T 是 t 的分布.

如表 6 和表 7 所示, 有无门控模块和有无图像信息对最终的情感决策都有一定影响. 除此以外, 在数据量只有 3 500 的小数据集上, 我们观察到模型存在比较严重的过拟合现象, 因此我们尝试使用 AEDA[27]方法为数据集引入文本噪声 (在随机位置插入随机标点符号), 使用翻转、裁切、反色等方式增强图像数据, 将训练集数据量翻倍至 7 000, 然而数据增强所带来的效

果并未达到预期, 甚至情感分析的效果出现轻微下降, 不能很好地缓解过拟合问题是模型的一个局限性.

表 6 消融实验 (%)

模块设置	Acc	*F*1
原始模型	72.97	71.57
去除门控模块	72.34	70.97
去除图像输入	71.79	70.54
添加数据增强	72.87	71.34

表 7 *t*/*p* 值计算

模块设置	*t*值	*p*值
是否有门控模块	2.77	0.021
是否有图像	6.72	0.008
是否进行数据增强	0.92	0.87

另一个局限性则来源于大模型, 虽然 Qwen2-VL 大模型展示出极强的图像描述能力, 但模型偶尔也会生成错误文本, 如图 4 所示, 包含重复描述、语言错误等问题, 需要人工核查图像的文本描述也是一项不足.

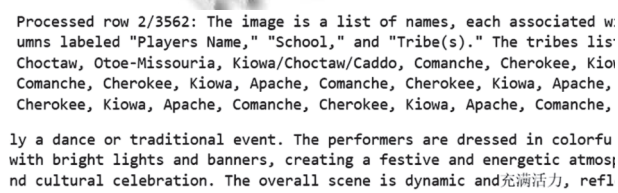

图 4 图像的错误描述

此外, 我们也获取到模型的分析结果, 如图 5 所示, 第 1 组图文对里, 文本介绍了照片的拍摄地点, 叙述偏中性, 而图像中每个人的笑容则传达出了喜悦的气氛, 内容是积极的. 此时的分析目标是这群人的情感表达, 捕获到图像的积极情绪, 过滤掉文本的中性叙述至关重要. 所以在没有门控模块过滤信息, 或者没有图像输入的情况下都得出了错误的分析结果. 第 2 组图文对里, 文本的“refugee”, 即难民一词展示了极强的消极情绪, 而图像中的人面色平静, 情感表达偏中性. 此时分析目标是难民中的一个人名, 难民所传递的情感是消极的. 此时过滤掉图像中的中性情感至关重要, 而这也是去掉门控模块的模型所不具备的. 这两组结果也说明了图像信息的输入和门控模块在情感分析过程中的重要性.

4 结语

本文所提出的模型 ITFFT, 借助大模型对数据进行预处理, 将图文多模态情感分析任务中图像的关键元素转换为一段文本描述, 使多模态情感分析回归为一项单模态情感分析任务. 此外, 对文本的处理中, 引入了相对位置编码用以弥补传统 BERT 模型仅有绝对位置编码的不足. 最后, 在决策层使用门控模块对特征进行筛选输出最终的分析结果, 本文模型相较于先前模型以最少的参数实现了最优的结果.

图 5 分析结果展示

此项研究仍有以下问题需要解决.

1) 在复现了多个模型后, 发现各模型在 Twitter-15/17 这种小规模数据集上训练存在严重的过拟合问题, 如何寻找到一种方案缓解模型在该数据集上的过拟合问题是值得研究的.

2) 本文模型仅在一张图像对应一段文本的数据集上开展了实验, 倘若用户用一段文本对应多张图片, 图像的文本描述将远大于原始的文本信息, 如何处理多张图片的权重也是一项需要研究的问题.

3) 大模型通过图像生成的文本描述可能存在内容错误的问题, 如何设计一个方案来减轻甚至避免人工复查同样是一个需要解决的问题.

参考文献

1 第 54 次《中国互联网络发展状况统计报告》发布. 传媒论坛, 2024, 7(17): 121.
2 郭小宇, 马静. 基于 SEFusion-MPOR 的多模态特征融合舆情表征算法. 情报理论与实践, 2024, 47(7): 181–189.
3 韩滕跃, 牛少彰, 张文. 基于对比学习的多模态序列推荐算法. 计算机应用, 2022, 42(6): 1683–1688. [doi: 10.11772/j.issn.1001-9081.2021081417]

4 陈妍. 多模态通用科技赋能政府精准应急决策的互动逻辑与优化对策研究 [硕士学位论文]. 徐州: 中国矿业大学, 2023.
5 李振. 基于图文多模态的情感分析方法研究 [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2022.
6 Yin SK, Fu CY, Zhao SR, *et al*. A survey on multimodal large language models. National Science Review, 2024, 11(12): nwae403. [doi: 10.1093/nsr/nwae403]
7 Xu P, Zhu XT, Clifton DA. Multimodal learning with Transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113–12132. [doi: 10.1109/TPAMI.2023.3275156]
8 OpenAI. GPT-4 technical report. arXiv:2303.08774, 2024.
9 Chen YC, Li LJ, Yu LC, *et al*. UNITER: Universal image-text representation learning. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 104–120.
10 Cho J, Lei J, Tan H, *et al*. Unifying vision-and-language tasks via text generation. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 1931–1942.
11 Kiela D, Bhooshan S, Firooz H, *et al*. Supervised multimodal BiTransformers for classifying images and text. arXiv:1909.02950, 2019.
12 Yu JF, Jiang J. Adapting BERT for target-oriented multimodal sentiment classification. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019. 5408–5414. [doi: 10.24963/ijcai.2019/751]
13 Yu JF, Chen K, Xia R. Hierarchical interactive multimodal Transformer for aspect-based multimodal sentiment analysis. IEEE Transactions on Affective Computing, 2023, 14(3): 1966–1978. [doi: 10.1109/TAFFC.2022.3171091]
14 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 159.
15 Su JL, Ahmed M, Lu Y, *et al*. RoFormer: Enhanced Transformer with rotary position embedding. Neurocomputing, 2024, 568: 127063. [doi: 10.1016/j.neucom.2023.127063]
16 Touvron H, Lavril T, Izacard G, *et al*. LLaMA: Open and efficient foundation language models. arXiv:2302.13971, 2023.
17 Bai JZ, Bai S, Chu YF, *et al*. Qwen technical report. arXiv:2309.16609, 2023.
18 Chowdhery A, Narang S, Devlin J, *et al*. PaLM: Scaling language modeling with pathways. The Journal of Machine Learning Research, 2023, 24(1): 240.
19 He PC, Liu XD, Gao JF, *et al*. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv:2006.03654, 2020.
20 Xi HC, Li CH, Chen JF, *et al*. Training Transformers with 4-bit integers. Proceedings of the 37th International Conference on Neural Information Processing System. New Orleans: Curran Associates Inc., 2023. 2137.
21 Wang P, Bai S, Tan SN, *et al*. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191, 2024.
22 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]
23 Liu Z, Lin W, Shi Y, *et al*. A robustly optimized BERT pre-training approach with post-training. Proceedings of the 20th China National Conference on Computational Linguistics. Hohhot: Springer, 2021. 471–484. [doi: 10.1007/978-3-030-84186-7_31]
24 He PC, Gao JF, Chen WZ. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543, 2021.
25 Xu N, Mao WJ, Chen GD. Multi-interactive memory network for aspect based multimodal sentiment analysis. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 371–378. [doi: 10.1609/aaai.v33i01.3301371]
26 Zhang TZ, Zhou G, Lu JC, *et al*. Text-image semantic relevance identification for aspect-based multimodal sentiment analysis. PeerJ Computer Science, 2024, 10: e1904. [doi: 10.7717/peerj-cs.1904]
27 Karimi A, Rossi L, Prati A. AEDA: An easier data augmentation technique for text classification. arXiv:2108.13230, 2021.

(校对责编: 王欣欣)