文章编号:1671-6833(2024)02-0042-09

基于 CLIP 和交叉注意力的多模态情感分析模型

陈 燕[1,2], 赖宇斌[1], 肖 澳[1], 廖宇翔[1], 陈宁江[1]

(1.广西大学 计算机与电子信息学院,广西 南宁 530000;2.广西大学 广西多媒体通信与网络技术重点实验室,广西 南宁 530000)

摘 要:针对多模态情感分析中存在的标注数据量少、模态间融合不充分以及信息冗余等问题,提出了一种基于对比语言-图片训练(CLIP)和交叉注意力(CA)的多模态情感分析(MSA)模型 CLIP-CA-MSA。首先,该模型使用 CLIP 预训练的 BERT 模型、PIFT 模型来提取视频特征向量与文本特征;其次,使用交叉注意力机制将图像特征向量和文本特征向量进行交互,以加强不同模态之间的信息传递;最后,利用不确定性损失特征融合后计算输出最终的情感分类结果。实验结果表明:该模型比其他多模态模型准确率提高 5 百分点至 14 百分点,*F*1 值提高 3 百分点至 12 百分点,验证了该模型的优越性,并使用消融实验验证该模型各模块的有效性。该模型能够有效地利用多模态数据的互补性和相关性,同时利用不确定性损失来提高模型的鲁棒性和泛化能力。

关键词:情感分析; 多模态学习; 交叉注意力; CLIP 模型; Transformer; 特征融合

中图分类号:TP391 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2024.02.003

随着科技的发展和短视频平台的流行,人们在社交媒体和各种网站平台上的情感表达方式也越来越多样化,不仅有文本,还有图片、音频、视频等多模态信息。相比于单一模态信息,多模态数据可以从不同视角表达语义信息,包含更多情感内容。图 1 为一个多模态表达的例子,一段视频片段配上字幕“你那点财务还需要助理?”,如果只看文字,可能会感觉说话者是在轻视和嘲讽对方,让观众认为此处表现出消极的情感极性。但是结合视频内容,就可以发现说话者是在开玩笑地和对方交流,想表达的是积极情感。因此,利用多模态信息的互补和增强,可以更全面和准确地理解人们的情感状态。

你那点财务还需要助理?

(a) 字幕内容

(b) 视频片断

图 1 多模态数据示例

Figure 1 Examples of multimodal data

在文本数据缺乏情感信息的情况下,可以利用图片、视频或其他模态信息来加强和补充。但文本和图像包含的情感信息属于不同层次和不同程度的信息,因此存在相关性的同时也会包含冗余信息和噪声信息。此外,目前许多特征融合方法依赖预设的规则或权重,不能自适应地调整模态之间的关系和重要性。因此,多模态情感分析任务面临着一些挑战。

情感分析任务最早由 Pang 等[1]提出,通过词袋框架和有监督的机器学习方法对电影文本评论进行情感分类。随着数据语料库和人工智能技术的发展,情感分析任务得到了越来越多人的重视,并得到了广泛的应用[2]。目前,情感分析研究不局限于单一模态的文本数据,还包括图片、动图、视频等多种模态数据相融合的情感分析。

在文本情感分析方面,李勇等[3]基于双向长短时记忆网络(Bi-LSTM)与位置注意力机制提取语义特征,使用 CNN 对食品评论进行分类,得到比较好的分类效果。Munikar 等[4]通过 BERT 预训练模型对 10 000 余条电影评论数据进行细粒度情感分析,提高了多分类情感任务的效果。在视觉情感分析方

收稿日期:2023-08-20;**修订日期**:2023-10-18

基金项目:广西壮族自治区科学研究与技术开发计划资助项目(桂科 AA20302002-3);广西壮族自治区自然科学基金资助项目(2020GXNSFAA159090)

作者简介:陈燕(1975—),女,广西玉林人,广西大学教授,博士,主要从事人工智能、优化算法等研究,E-mail:cy@gxu.edu.cn。

引用本文:陈燕,赖宇斌,肖澳,等. 基于 CLIP 和交叉注意力的多模态情感分析模型[J]. 郑州大学学报(工学版),2024,45(2):42-50. (CHEN Y,LAI Y B,XIAO A,et al. Multimodal sentiment analysis model based on CLIP and cross-attention model[J]. Journal of Zhengzhou University(Engineering Science), 2024,45 (2):42-50.)

面,Zhu 等[5]提出了一种统一的 CNN-RNN 模型,通过不同层次的特征融合和依赖关系,有效地实现了视觉情感识别。You 等[6]提出了一个基于注意力机制的视觉情感分析模型,能够自动发现和加权图像中与情感相关的局部区域。在多模态情感分析方面,针对多模态情感分析中存在的标注数据量少、模态间融合不充分,以及信息冗余等问题,Wang 等[7]使用选择加法学习方法将不同模态的特征进行加权平均,得到一个多模态的特征表示,可以提高神经网络的泛化能力;吴思思等[8]使用后端融合的方法,提出了一种基于感知融合的多任务多模态情感分析模型,有效地整合了文本、语音和图像 3 种模态信息,并利用多任务学习来提高模型的泛化能力。但上述多模态情感分析模型在特征融合上有一定缺陷,需要使用自注意力机制加强模态之间信息交互和融合。

针对多模态情感分析数据集数据缺乏、模型特征融合不足等问题,本文提出了一种基于对比语言-图片训练(contrastive language-image pretraining, CLIP)[9]和交叉注意力(cross-attention, CA)的多模态情感分析(multimodal sentiment analysis, MSA)模型 CLIP-CA-MSA。本文使用了根据自然语言指示从图像中预测最相关的文本片段的 CLIP 多模态预训练模型和利用提示学习,在少量数据下得到较好的文本情感分类效果的 PIFT[10]模型,并进行特征提取,同时引入了交叉注意力机制来实现不同模态之间的信息传递。对于视觉情感分析,借助 CLIP 预训练模型的丰富先验信息,使用标签文本作为提示信息,并采用预训练的对比学习方法进行相似度计算,得到相似度分数最高的类别作为视觉情感分析结果。为了减少冗余和噪声信息的影响,使用了不确定性损失函数来自动分配视觉和文本的重要性占比,以增强模型的泛化能力和鲁棒性。

1　基于 CLIP 和交叉注意力的多模态情感分析模型

本文提出的 CLIP-CA-MSA 模型结构如图 2 所示。

首先将视频按照一定的帧率分割成若干张图片,然后使用 CLIP 预训练的 BERT 模型和 ViT 模型来提取标签特征和每张图片的图像特征,并使用 Transformer 编码器将图像特征构建成一个视频特征向量。接着使用 PIFT 模型来提取文本数据的文本特征。随后,使用交叉注意力机制将图像特征向量和文本特征向量进行交互。最后,再利用标签特征计算视频和标签之间的相似度,得到一个视频分类特征向量。该向量和文本分类特征向量一起被输入到同方差不确定性损失中进行计算,并输出最终的情感分类结果。

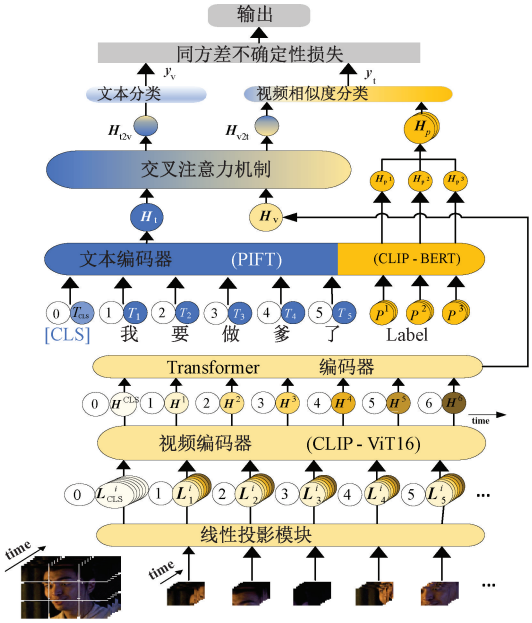


图 2　CLIP-CA-MSA 模型结构

Figure 2　CLIP-CA-MSA model structure

CLIP-CA-MSA 模型算法如下。

输入:视频数据集 D_v 和文本数据集 D_t,数据集数量大小 M,最大迭代次数 N;

输出:模型 f。

① for $t=1,2,\cdots,N$ do
② 　for $m=1,2,\cdots,M$ do
③ 　　//将视频数据输入到视频编码器
④ 　　video_encorder←$D_v(m)$;
⑤ 　　Transformer_encorder←video_encorder;
⑥ 　　//将文本数据输入到文本编码器
⑦ 　　text_encorder←$D_t(m)$;
⑧ 　　//特征提取
⑨ 　　$\boldsymbol{H}_v$←Transformer_encorder;
⑩ 　　$\boldsymbol{H}_t$←text_encorder;
⑪ 　　//交叉注意力机制
⑫ 　　cross_attention←$\boldsymbol{H}_v,\boldsymbol{H}_t$;
⑬ 　　$\boldsymbol{H}_{v2t},\boldsymbol{H}_{t2v}$←cross_attention;
⑭ 　　//损失函数
⑮ 　　$Loss_{all}$←$\boldsymbol{H}_{v2t},\boldsymbol{H}_{t2v}$;
⑯ 　end for
⑰ 　更新权重参数;
⑱ end for。

1.1　特征提取

本文使用的多模态情感分析数据集包含文本、视频 2 个模态。

(1)文本特征提取。文本模态由文本信息(视频对应的字幕信息)和标签信息(加入提示的标签文本)构成,如式(1)所示。

$$Text = \{T, P^1, P^2, P^3\}。 \tag{1}$$

式中:$Text$ 表示文本模态;T 表示文本信息;P^1、P^2、P^3 表示加入提示的分类标签(如中性、积极和消极)信息。

将文本信息 T 和标签信息 P^1、P^2、P^3 按字粒度划分,如式(2)、(3)所示。

$$T = \{T_{\mathrm{CLS}}, T_1, \cdots, T_n\}; \tag{2}$$

$$P^i = \{P^i_{\mathrm{CLS}}, prompt, P^i_1, P^i_2\}。 \tag{3}$$

式中:T_{CLS} 表示文本分类字符;T_1、T_n 分别表示文本信息中的第 1 个和第 n 个字符,n 不大于最大句子长度;$i \in [1,3]$;$prompt$ 表示提示文本;P^i_1、P^i_2 代表分类标签。

为了避免模型规模过大和训练难度大的问题,采用了基于提示嵌入和焦点损失函数的 PIFT 模型来提取文本特征,具体提取过程如式(4)所示。为保证模型的情感分析精度,利用经过 CLIP 预训练的 BERT 模型来提取标签信息,提取过程如式(5)所示。

$$\boldsymbol{H}_{\mathrm{t}} = \mathrm{PIFT}(T_{\mathrm{CLS}}, T_1, T_2, \cdots, T_n); \tag{4}$$

$$\boldsymbol{H}_P = [\boldsymbol{H}_{P^1}, \boldsymbol{H}_{P^2}, \boldsymbol{H}_{P^3}] = \mathrm{BERT}(P^1, P^2, P^3)。 \tag{5}$$

式中:$\boldsymbol{H}_{\mathrm{t}}$ 表示文本特征向量;$\boldsymbol{H}_p$ 表示所有类别的标签特征向量。

(2)视频特征提取与融合。为了获取视频表示,首先从视频片段中按帧提取出一组图像,即 $V = (V^1, V^2, \cdots, V^m)$,其中 m 表示每组图片最大数量(本文实验中 $m=6$)。然后通过视觉编码器对其进行编码,得到视频特征序列。

本文使用 CLIP 预训练的 ViT-B-16 视觉模型提取视频特征,如图 2 线性投影模块所示。ViT 视觉编码器将图像分割成不重叠的块,并添加位置信息,在开头插入 1 个特殊标记 $\boldsymbol{V}^i_{\mathrm{CLS}}$,以表示整个图像的全局特征,即 $\boldsymbol{V}^i = (\boldsymbol{V}^i_{\mathrm{CLS}}, \boldsymbol{V}^i_1, \boldsymbol{V}^i_2, \cdots, \boldsymbol{V}^i_k)$,其中 $i \in [1,m]$,k 为图像块的数量。接着使用线性投影将二维的图像块映射为一维序列作为 ViT 模型的输入,过程如式(6)所示。

$$\boldsymbol{L}^i = \{\boldsymbol{L}^i_{\mathrm{CLS}}, \boldsymbol{L}^i_1, \boldsymbol{L}^i_2, \cdots, \boldsymbol{L}^i_k\} = \mathrm{Linear}(\boldsymbol{V}^i_{\mathrm{CLS}}, \boldsymbol{V}^i_1, \boldsymbol{V}^i_2, \cdots, \boldsymbol{V}^i_k)。 \tag{6}$$

式中:Linear 表示线性投影操作,其在保留图像中的信息的同时可以减少计算量;$\boldsymbol{L}^i_1$ 表示第 i 张图像的第 1 块经过线性投影后得到的一维序列。

如图 2 中视频编码器模块所示,CLIP-CA-MSA 利用 ViT 编码器对输入图像中每个块之间的相互关系进行建模以获取图像特征,如式(7)所示。

$$\boldsymbol{H} = \{\boldsymbol{H}^1, \boldsymbol{H}^2, \cdots, \boldsymbol{H}^m\} = \mathrm{ViT}(\boldsymbol{L}^1, \boldsymbol{L}^2, \cdots, \boldsymbol{L}^m)。 \tag{7}$$

式中:$\boldsymbol{H}^1$ 表示 ViT 从第 1 张图片提取出图像特征;$\boldsymbol{H}$ 表示视频特征序列。

最终,需要融合图像特征序列得到代表整组图像特征的视频特征向量 $\boldsymbol{H}_{\mathrm{v}}$。本文使用 Transformer 编码器来融合视频特征序列。首先,插入标记 $\boldsymbol{H}^{\mathrm{CLS}}$ 作为视频全局特征表示,并为图像加入时序信息;其次,使用自注意力机制获取视频中的时空关系,以有效地帮助识别视频情感极性。具体融合过程如式(8)所示。

$$\boldsymbol{H}_{\mathrm{v}} = \mathrm{Transformer}(\boldsymbol{H}^{\mathrm{CLS}}, \boldsymbol{H}^1, \boldsymbol{H}^2, \cdots, \boldsymbol{H}^m)。 \tag{8}$$

式中:$\boldsymbol{H}_{\mathrm{v}}$ 为视频的特征向量,蕴含视频的重要信息。

1.2 交叉注意力机制

为了减少单一模态情感信息不足或噪声污染的问题,本文使用交叉注意力机制进行模态交互。交叉注意力机制是一种在多模态情感分析中用于融合不同模态信息的注意力机制,它可以在图像、文本等模态之间交叉计算注意力分数,以提取共享的情感特征,并增强每个模态的表示能力。本文采用的交叉注意力机制的基本原理如图 3 所示。

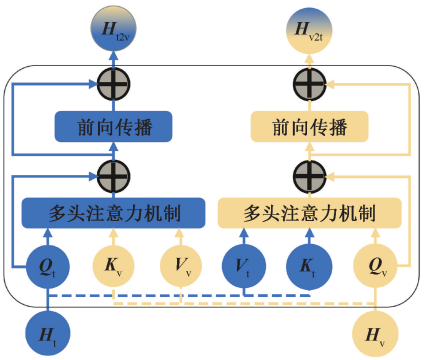


图 3 交叉注意力机制

Figure 3 Cross-attention mechanism

首先,使用一个输入作为查询($\boldsymbol{Q}$),另一个输入作为键($\boldsymbol{K}$)和值($\boldsymbol{V}$),用注意力机制来计算 2 个输入每个元素之间的相关性;其次,将注意力权重与值($\boldsymbol{V}$)相乘并求和,得到模态间的交互特征;最后,将交互特征与原始输入连接起来,形成新的融合了多模态信息的特征表示。通过这种方式,可以有效地减少单个模态在情感分析中的局限,提高模型的泛化性能和鲁棒性。

多头注意力机制是交叉注意力机制进行模态交互的重点,其计算过程如式(9)、(10)所示。

$$\boldsymbol{S}_{\mathrm{t}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{t}} \cdot \boldsymbol{K}_{\mathrm{v}}^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot \boldsymbol{V}_{\mathrm{v}}; \tag{9}$$

$$\boldsymbol{S}_{\mathrm{v}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{v}} \cdot \boldsymbol{K}_{\mathrm{t}}^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot \boldsymbol{V}_{\mathrm{t}}。\tag{10}$$

式中:· 为点乘操作;Softmax 代表归一化函数;d_k 表示键向量的维度,此处的作用为对点积的结果进行缩放,避免结果过大或过小影响 Softmax 的梯度。

以式(9)为例,计算文本查询向量 $\boldsymbol{Q}_{\mathrm{t}}$ 与视频键向量 $\boldsymbol{K}_{\mathrm{v}}^{\mathrm{T}}$ 的点积,并进行缩放与归一化,得到模态间的注意力分数。再将注意力分数与视频值向量 $\boldsymbol{V}_{\mathrm{v}}$ 点积,获得文本-视频的注意力向量 $\boldsymbol{S}_{\mathrm{t}}$,其中包含了文本与视频的相似度信息。式(10)同理。

残差连接与正则化计算过程如式(11)、(12)所示。

$$\boldsymbol{S}'_{\mathrm{t}} = \mathrm{LayerNorm}(\boldsymbol{S}_{\mathrm{t}} + \boldsymbol{Q}_{\mathrm{t}});\tag{11}$$

$$\boldsymbol{S}'_{\mathrm{v}} = \mathrm{LayerNorm}(\boldsymbol{S}_{\mathrm{v}} + \boldsymbol{Q}_{\mathrm{v}})。\tag{12}$$

以式(11)为例,将 $\boldsymbol{Q}_{\mathrm{t}}$ 与 $\boldsymbol{S}_{\mathrm{t}}$ 相加能够帮助特征向深层网络传递,再进行正则化以提高模型的稳定性和收敛速度,得到 $\boldsymbol{S}'_{\mathrm{t}}$。然后将其进行前向传播为特征增加非线性变化,增强其表达能力。最后经过一次求和与归一化得到文本-视频特征向量 $\boldsymbol{H}_{\mathrm{t2v}}$。式(12)同理,最后得到视频-文本特征向量 $\boldsymbol{H}_{\mathrm{v2t}}$。

1.3　图像相似度分类

为了在少量数据下提高模型效果,根据视频特征向量与每个情感分类标签之间的相似程度来判断其所属类别。具体相似度计算过程如式(13)所示。

$$\boldsymbol{y}_{\mathrm{v}} = logit\_scale = \boldsymbol{H}_{\mathrm{v2t}} \cdot \boldsymbol{H}_{P}^{\mathrm{T}}。\tag{13}$$

式中:*logit_scale* 为一个可调节视频和标签之间相似度分数的学习参数,它决定了相似度分数的范围和敏感度,*logit_scale* 越小,相似度分数越平滑,越大则区分能力越强;“·”为点乘操作,能够计算视频特征向量 $\boldsymbol{H}_{\mathrm{v2t}}$ 和标签特征向量 $\boldsymbol{H}_{P}^{\mathrm{T}}$ 的相似度。

1.4　同方差不确定性损失

多模态任务的重点之一在于如何平衡不同任务损失之间的权重,目前大多数方法采用对多个模态的损失函数进行加权,其损失函数如式(14)所示。

$$Loss_{\mathrm{all}} = \mu_{\mathrm{t}} Loss_{\mathrm{t}} + \mu_{\mathrm{v}} Loss_{\mathrm{v}}。\tag{14}$$

式中:μ_{t} 与 μ_{v} 分别表示文本和视频模态损失函数的权重;$Loss_{\mathrm{t}}$ 与 $Loss_{\mathrm{v}}$ 表示文本和视频模态损失函数,$Loss_{\mathrm{t}}$ 使用焦点损失函数,$Loss_{\mathrm{v}}$ 使用相似度计算损失函数;$Loss_{\mathrm{all}}$ 表示总体网络的损失函数,即多模态任务的优化目标。

然而,简单的线性加权求和方法需要人为设定每个模态的权重,这不符合实际数据的分布和特性,会导致某些模态被过分强调或忽略,也限制了模型的泛化能力。

因此,本文使用同方差不确定性损失来自动平衡不同模态之间的损失函数权重,同时避免信息的丢失或者冗余。假设 $\boldsymbol{x}$ 表示模型的输入,$\boldsymbol{W}$ 为参数矩阵,其概率似然估计如式(15)所示。

$$P(\boldsymbol{y} \mid f^{\boldsymbol{W}}(\boldsymbol{x})) = \mathrm{Softmax}(f^{\boldsymbol{W}}(\boldsymbol{x}))。\tag{15}$$

式中:Softmax 函数用来从产生的概率向量中抽取样本。假设文本与视频模态的输出向量为 $\boldsymbol{y}_{\mathrm{t}}$ 与 $\boldsymbol{y}_{\mathrm{v}}$,并都服从高斯分布,则模型的最大似然函数如式(16)所示。

$$\begin{aligned} P(\boldsymbol{y}_{\mathrm{t}}, \boldsymbol{y}_{\mathrm{v}} \mid f^{\boldsymbol{W}}(\boldsymbol{x})) &= P(\boldsymbol{y}_{\mathrm{t}} \mid f^{\boldsymbol{W}}(\boldsymbol{x})) \cdot P(\boldsymbol{y}_{\mathrm{v}} \mid f^{\boldsymbol{W}}(\boldsymbol{x})) = \\ & N(\boldsymbol{y}_{\mathrm{t}}; f^{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\mathrm{t}}^{2}) \cdot N(\boldsymbol{y}_{\mathrm{v}}; f^{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\mathrm{v}}^{2})。 \end{aligned}\tag{16}$$

为了最大化似然参数,需要最小化其负对数似然函数,过程如式(17)所示。

$$\begin{aligned} L(\boldsymbol{W}, \sigma_{\mathrm{t}}, \sigma_{\mathrm{v}}) &= -\log P(\boldsymbol{y}_{\mathrm{t}}, \boldsymbol{y}_{\mathrm{v}} \mid f^{\boldsymbol{W}}(\boldsymbol{x})) \propto \\ & \frac{1}{2\sigma_{\mathrm{t}}^{2}} \| \boldsymbol{y}_{\mathrm{t}} - f^{\boldsymbol{W}}(\boldsymbol{x}) \|^2 + \frac{1}{2\sigma_{\mathrm{v}}^{2}} \| \boldsymbol{y}_{\mathrm{v}} - f^{\boldsymbol{W}}(\boldsymbol{x}) \|^2 + \\ & \log \sigma_{\mathrm{t}} \sigma_{\mathrm{v}} = \frac{1}{2\sigma_{\mathrm{t}}^{2}} L_{\mathrm{t}}(\boldsymbol{W}) + \frac{1}{2\sigma_{\mathrm{v}}^{2}} L_{\mathrm{v}}(\boldsymbol{W}) + \\ & \log \sigma_{\mathrm{t}} + \log \sigma_{\mathrm{v}}。 \end{aligned}\tag{17}$$

式中:$L_{\mathrm{t}}(\boldsymbol{W}) = -\log \mathrm{Softmax}(\boldsymbol{y}_{\mathrm{t}}, f^{\boldsymbol{W}}(\boldsymbol{x}))$ 表示文本模态的损失;σ 表示模态的噪声大小,反映了模型对某个模态的难度和置信度;$\frac{1}{\sigma^2}$ 表示各模态损失函数的权重,但由于模型会最小化损失函数,会使 σ 变得很大,因此加入了正则化项 $\log \sigma$,可以有效防止噪声增加过多。模型训练时会自动寻找各模态的最优权重。

2　实验结果及分析

本文将详细介绍所采用的多模态数据集、实验评价指标和实验参数设置,将 CLIP-CA-MSA 模型与其他多模态模型进行对比实验并进行分析。

2.1　多模态数据集与评价指标

为验证 CLIP-CA-MSA 模型的情感分析性能,本文采用公开数据集 CH-SIMS(chinese single and multimodal sentiment)[11] 进行实验。数据集分布情况如图 4 所示。

CH-SIMS 数据集是一个中文多模态情感分析数据集,视频来源于中文电影、电视剧和演出节目,根据说话者的话语将视频帧划分为多个片段,每个片段对应一个说话者的一句话,长度在 1~10 s 之间,对每个视频片段的文本和视觉模态分别进行消极、中性和积极的情感极性标注。

2.2　多模态模型对比实验

本文选取了几种常用的多模态情感分析模型作为基准模型,并与 CLIP-CA-MSA 模型进行实验对比

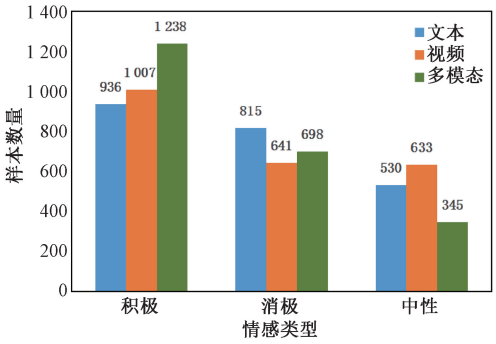


图 4 CH-SIMS 数据分布图

Figure 4 CH-SIMS data distribution diagram

和分析。这些基准模型包括以下几种。

TFN[12]:使用张量外积来显式地聚合单模态、双模态和三模态之间的交互关系。

LMF[13]:使用 LSTM 编码文本,CNN 编码图像,用低秩张量外积来聚合模态间的交互关系。

MulT[14]:利用方向性成对交叉模态注意力,可以在不同时间步中关注多模态序列之间的交互,并隐式地适应数据的对齐方式。

Self-MM[15]:利用自监督任务来增强多模态情感分析的方法,可以同时学习不同模态的特征表示和情感分类。

MMAF[8]:利用多任务学习和感知融合层对多模态数据进行情感分析。

MMAF+T+A+V:将 MMAF 提取的多特征向量与单模态特征向量融合。

CLIP-CA-MSA 模型与上述基准模型在 CH-SIMS 数据集的对比实验结果如表 1 所示。

表 1 多模态模型对比实验结果

Table 1 Multimodal model comparison experiment results %

模型	准确率 *Acc*	*F*1
TFN	64.83	65.14
LMF	65.41	65.49
MulT	66.28	66.72
Self-MM	67.94	68.28
MMAF	69.53	69.95
MMAF+T+A+V	73.61	74.13
CLIP-CA-MSA	**78.07**	**77.39**

由表 1 可以看出,TFN 和 LMF 的效果相对较差,说明张量外积聚合交互关系并不足以捕捉多模态数据之间复杂的关联性。MulT 能够捕捉不同时间步中多模态序列之间的交互关系,但仍没有考虑到多模态数据之间的语义一致性和情感相关性。Self-MM 让模型同时学习到多模态数据之间的内在联系,提高情感分类效果。MMAF 通过引入多任务学习和感知融合层来增强多特征向量的表达能力,而 MMAF+T+A+V 能更好地理解每个单独模态以及整体模态对于情感分类任务的贡献程度,并且避免了信息冗余或丢失,两者的 *Acc* 和 *F*1 值均有明显的提升。

CLIP-CA-MSA 模型利用 CLIP 方法来提取多模态特征和标签特征,引入 PIFT 模型来提取文本特征,交叉注意力机制能保留模态内特征和关注模态间特征,通过同方差不确定性损失自动调整模态重点,使得模型在准确率 *Acc* 上达到了 78.07%,*F*1 值达到了 77.39%。

综上所述,CLIP-CA-MSA 模型最优,其成功的原因在于它引入了强大的多模态特征提取方法、多模态融合方式以及自动均衡模态权重,使得模型能够更好地利用多模态数据之间的交互关系。

2.3 多模态融合对比实验

为验证 CLIP-CA-MSA 模型多模态融合的效果,先将视频和文本模态用视觉模型和文本模型分别进行单模态实验,再进行模态融合实验。

2.3.1 视觉模型对比实验

本文对 CH-SIMS 数据集中的视频模态部分进行情感分类任务,采用了常用的 5 个深度学习视觉模型进行测试和比较。进行实验对比的模型相关信息如下。

VGG-16[16]:使用小卷积核和多卷积子层方法的深度神经网络,提高计算效率和网络性能。

ResNet[17]:由多个残差块组成深度神经网络,使用快捷连接的方法,解决了深层网络训练中的退化现象。

ConvNeXt[18]:基于 CNN 卷积网络,参考 Transformer 网络的思想,对 ResNet 网络的卷积层、池化层和注意力机制进行了改进。

OpenFace2.0[19]:一个面部行为分析工具,使用基于卷积神经网络的局部模型,可以从图片中检测出 68 个人脸关键点,并根据这些关键点估计头部姿态、眼睛注视方向和面部动作单元。

ViT[20]:通过将图片分成固定大小的块,然后通过线性变化作为 Transformer 的输入序列,从而进行特征提取和分类。ViT-B-16 使用 16×16 的块,ViT-B-32 使用 32×32 的块。

视觉模型的实验结果如表 2 所示,*P* 为精确率,*R* 为回收率。

由表 2 可知,VGG-16 模型的层数较浅,无法很好地提取视频特征,所以表现最差。而 ConvNeXt 网络的卷积层、池化层和注意力机制的改进能使视频分类效果有一定提升。OpenFace2.0 在面部行为分

表 2　视觉模型对比实验结果

Table 2　Visual model contrast experiment results　%

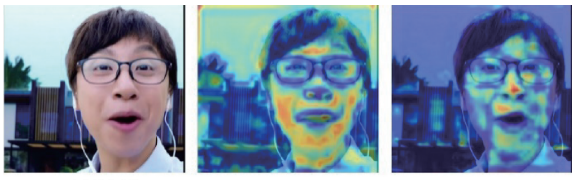
视觉模型	*Acc*	*F*1	*P*	*R*
VGG-16	48.25	35.95	31.08	42.63
ConvNeXt	52.52	55.85	57.60	54.21
OpenFace2.0	61.37	61.53	60.42	62.72
ResNet34	68.08	67.61	67.25	67.98
ResNet50	68.11	69.59	68.25	68.95
ViT-B-32	67.90	68.15	68.14	68.28
ViT-B-16	**69.90**	**69.14**	**68.98**	**69.50**

析上表现优异。ResNet34 具有良好的深度和残差连接结构,能够很好地提取视频特征。ResNet50 是 ResNet 系列中更深、更复杂的模型,具有更多的层和残差块,使网络能够更准确地进行视频分类,其效果略好于 ResNet34,这也证明了深层网络能够提高模型的表现。

ViT 模型中,相较于 ViT-B-32,ViT-B-16 的准确率和 *F*1 值分别提高了 2.00 百分点和 0.99 百分点,这是由于块的大小对模型性能的影响,更小的块可以捕捉到更细粒度的图像特征。相比于效果最差的 VGG-16,准确率和 *F*1 分别提高了 21.65 百分点和 33.19 百分点。相较于 ResNet50,其准确率提升了 1.79 百分点,但是 *F*1 降低了 0.45 百分点。

由于该数据集的规模不大、多样性不足,无法很好地判断 2 个模型的优劣。ViT-B-16 准确率较高,说明其在处理图像中的全局特征和细粒度特征方面表现更好,可以更好地识别视频中的物体和场景,但需要更多的计算资源和数据量。而 ResNet50 有较高的 *F*1 值,这说明该模型在处理视频中的空间信息方面表现更好,能够更准确地对视频进行分类,同时具有较好的稳健性。

本文使用基于消融分析的可视化方法 Ablation-CAM[21],为 2 个模型生成视觉解释并且定位图像中的相关区域,如图 5 所示。



(a) 原始图像　(b) ResNet50　(c) ViT-B-16

图 5　可视化分析

Figure 5　Visual analysis

这 2 张热力图显示出了模型对于人脸表情的关注点。ResNet 的热力图显示出了模型对于图像的整体区域都有关注,其中主要集中在人脸上,但分散的关注点可能会导致模型判断错误。而 ViT 的热力图则显示出了模型对于人脸表情的关注更加集中,这是因为 ViT 使用了自注意力机制,可以更好地捕捉到图像中的局部特征。

2.3.2　文本模型对比实验

采用文本分类模型 ALBERT[22]、BERT[23]、ERNIE[24]、MacBERT[25]和 RoBERTa[26]作为词嵌入工具,使用提示嵌入和焦点损失进行情感分类,得到的模型分别为 Al-PIFT、B-PIFT、E-PIFT、M-PIFT 和 PIFT。文本模型实验结果如表 3 所示。

表 3　文本模型对比实验结果

Table 3　Text model contrast experiment results　%

文本模型	*Acc*	*F*1	*P*	*R*
Al-PIFT	69.35	66.31	67.29	65.35
B-PIFT	71.48	71.75	71.77	71.88
E-PIFT	73.52	73.28	73.26	73.34
M-PIFT	75.00	74.61	74.94	74.38
PIFT	**75.88**	**75.33**	**75.26**	**75.41**

由表 3 可知,Al-PIFT 在所有指标上都表现最差,说明其在降低参数量和内存消耗的同时,也限制了模型容量和表征能力。B-PIFT 虽各项指标有了一定的提升,但表现不如其他模型。E-PIFT 的各项指标均有所提高,说明 ERNIE 模型能增强对中文语言特征的理解能力。M-PIFT 使用纠错型掩码语言模型等方法进一步提高模型性能。PIFT 模型在所有指标上都取得了最好的结果,这说明 RoBERTa 通过增加数据量和训练时间等方法进一步提高模型性能,让模型学习到更丰富的先验信息。

2.3.3　模型融合实验

为验证模态融合的有效性,文本模型均采用效果最好的 PIFT 进行文本特征提取,对视觉模型使用效果较好的 ResNet34、ResNet50、ViT-B-32 及 ViT-B-16 进行视觉特征提取,再使用本文方法进行模态融合,实验结果如表 4 所示。

表 4　模态融合对比实验

Table 4　Text-Visual model contrast experiment　%

文本-视觉模型	*Acc*	*F*1	*P*	*R*
CLIP-ResNet34	75.19	75.17	75.35	75.00
CLIP-ResNet50	77.06	76.85	76.65	77.06
CLIP-ViT32	76.04	75.49	75.70	75.29
CLIP-CA-MSA	**78.07**	**77.39**	**78.01**	**76.69**

通过实验评估,发现 CLIP-ResNet50 和 CLIP-CA-MSA 表现相近,但 CLIP-CA-MSA 在准确率和回收率 2 个指标上均优于 CLIP-ResNet50。

3　消融实验

为验证本文各模块对多模态情感分析的性能提

升效果,本文分别针对视频融合方法、特征融合方法、图像分类方法及损失函数,在 CH-SIMS 数据集上进行消融实验。

3.1 视频融合方法消融实验

视频融合的方法主要有 MeanP、LSTM、Transformer[27] 3 种。其中,MeanP 可以减少计算量和内存的消耗,但是也忽略了视频中的时序信息,无法捕捉视频的动态变化和关键帧。LSTM 可以学习视频中的长期依赖关系,捕捉视频的时序信息和动态变化,但计算量和内存消耗较大,容易出现梯度消失或爆炸的问题。Transformer 使用自注意力机制对多帧视频进行并行建模,实现全局交互和长范围依赖,捕捉视频中时空信息动态变化。视频融合方法消融实验结果如表 5 所示。

表 5 视频融合方法消融实验结果

Table 5 Video fusion ablation experiment results %

视频融合方法	*Acc*	*F*1	*P*	*R*
MeanP	75.44	75.59	75.74	75.68
LSTM	77.63	77.18	77.14	**77.32**
Transformer	**78.51**	**77.63**	**78.35**	77.07

从表 5 可知,MeanP 方法各项指标都较低,比文本单模态分类准确率低了 0.44 百分点,说明 MeanP 忽略了视频中情感的变换过程。LSTM 可以有效地考虑到视频特征之间的时序关系,在各项指标上都有提升。Transformer 方法能考虑到视频特征之间的空间关系与交互信息,准确率较 LSTM 方法提高了 0.88 百分点。

3.2 多模态特征融合方法消融实验

拼接(concat)和交叉注意力(cross-attention)为多模态特征融合的 2 种方法。简单拼接方法将各模态数据进行简单拼接后使用一个编码器来处理融合后的信息,可以节省计算资源,但会忽略单模态内的交互信息。交叉注意力为每个模态设计一个 Transformer 编码器,提取各模态特征,再交互模态特征,得到综合的多模态表示。可以实现不同模态之间的信息交互,从而获得更丰富的语义信息。其他模块保持不变,更改多模态特征融合方法,实验结果如表 6 所示。

表 6 多模态特征融合方法消融实验结果

Table 6 Multi-mode feature fusion ablation experiment results %

特征融合方法	*Acc*	*F*1	*P*	*R*
concat	76.32	75.19	75.58	74.93
cross-attention	**78.51**	**77.63**	**78.35**	**77.07**

由表 6 可知,交叉注意力方法可以更好地处理各模态的特征,避免了冗余信息问题,因此相对于拼接方法的情感分类效果有了明显的提高。

综上所述,通过对多模态情感分析模型的消融实验进行效果对比,发现交叉注意力机制在 CH-SIMS 数据集上表现较好,验证了该方法的有效性。

3.3 视觉情感分类方法消融实验

本文实验使用 CLIP 模型中的相似度分类方法(similarity-CLS)将视觉特征与类别进行相似度计算,得分最高的类即为分类结果。与常用的线性分类(lineaer-CLS)进行对照实验,使用 CH-SIMS 数据集,结果如表 7 所示。

表 7 图像情感分类方法消融实验结果

Table 7 Ablation experiment of image emotion classification method results %

图像情感分类方法	*Acc*	*F*1	*P*	*R*
linear-CLS	76.75	76.10	76.17	76.04
similarity-CLS	**78.51**	**77.63**	**78.35**	**77.07**

由表 7 可知,Linear-CLS 方法的准确率为 76.75%,比相似度分类方法低 1.76 百分点。线性分类方法需要单独训练线性分类器,在特征空间中寻找一个超平面,将不同类别的数据分离开来,这种方法的表现可能会受到特征空间分布的影响。并且由于 CH-SIMS 的数据量不大、视频中存在噪声干扰,也会导致线性分类方法准确率降低。而相似度计算方法与 CLIP 模型的预训练任务相同,预训练模型所学习到的丰富特征可以直接转移到下游任务,不需要额外的适应过程,减少了模型的训练时间和数据需求。

3.4 损失函数消融实验

为了证明本文使用损失函数的有效性,将其与加权求和损失函数进行对比实验。损失函数消融实验结果如表 8 所示。

表 8 损失函数消融实验结果

Table 8 Loss function ablation experiment results %

损失函数	*Acc*	*F*1	*P*	*R*
加权求和损失	77.19	76.34	76.08	76.63
同方差不确定性损失	**78.51**	**77.63**	**78.35**	**77.07**

由表 8 可知,使用加权求和后的各项指标较单模态而言已经有了较大的提升,这说明将损失加权求和能够在一定程度上平衡不同模态的重要性和难度。但同方差不确定性损失在多模态情感分析中具有更好的效果,其在准确率与 *F*1 值上较加权求和损失提升了 1.32 和 1.29 百分点,说明各模态固有不确定性的重要性以及自动调整各模态权重能够更加准

确地学习到不同模态的信息,提高模型的性能。

4　结论

本文针对多模态情感分析存在的模态融合不充分、信息冗余以及数据量不足等问题,提出一种基于特征融合和不确定性损失的多模态情感分析模型 CLIP-CA-MSA。首先,阐述了 CLIP-CA-MSA 的整体框架。然后介绍了实验所使用的数据集以及参数设置,通过实验验证了该模型的优越性,并探究了不同的视觉模型对该方法的影响,证明了多模态预训练模型对该方法的有效提升。然后,通过消融实验,验证各模块的有效性。但本文只使用了 CH-SIMS 数据集的文本部分和视频的视觉部分。后续研究将加入视频中的音频模态,以确保数据的完整性,进一步提升模型情感分析的准确率和泛化能力。

参考文献:

[1] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? sentiment classification using machine learning techniques [C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). Stroudsburg: ACL, 2002: 79-86.

[2] ZHANG L, LIU B. Sentiment analysis and opinion mining [EB/OL]. (2015-12-31)[2023-04-24]. https://doi.org/10.1007/978-1-4899-7502-7_907-2.

[3] 李勇, 金庆雨, 张青川. 融合位置注意力机制和改进 BLSTM 的食品评论情感分析[J]. 郑州大学学报(工学版), 2020, 41(1):58-62.
LI Y, JIN Q Y, ZHANG Q C. Improved BLSTM food review sentiment analysis with positional attention mechanisms[J]. Journal of Zhengzhou University (Engineering Science), 2020, 41(1):58-62.

[4] MUNIKAR M, SHAKYA S, SHRESTHA A. Fine-grained sentiment classification using BERT [EB/OL]. (2019-10-04) [2023-04-24]. https://arxiv.org/abs/1910.03474.

[5] ZHU X G, LI L, ZHANG W, et al. Dependency exploitation: a unified CNN-RNN approach for visual emotion recognition [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. New York: ACM, 2017:3595-3601.

[6] YOU Q Z, JIN H L, LUO J B. Visual sentiment analysis by attending on local image regions [C]//Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence. New York:ACM, 2017: 231-237.

[7] WANG H H, MEGHAWAT A, MORENCY L P, et al. Select-additive learning: improving generalization in multimodal sentiment analysis [C]//2017 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2017: 949-954.

[8] 吴思思, 马静. 基于感知融合的多任务多模态情感分析模型[J]. 数据分析与知识发现,2023(10):74-84.
WU S S, MA J. Multi-task & multi-modal sentiment analysis model based on aware fusion[J]. Data Analysis and Knowledge Discovery, 2023(10):74-84.

[9] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26)[2023-04-24]. https://arxiv.org/abs/2103.00020.

[10] 赖宇斌, 陈燕, 胡小春,等. 基于提示嵌入的突发公共卫生事件微博文本情感分析[J]. 数据分析与知识发现,2023,7(11):46-55.
LAI Y B, CHEN Y, HU X C. et al. Emotional analysis of public health emergency micro-blog based on prompt embedding[J]. Data Analysis and Knowledge Discovery, 2023,7(11):46-55.

[11] YU W M, XU H, MENG F Y, et al. CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 3718-3727.

[12] ZADEH A, CHEN M H, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis [EB/OL]. (2017-07-23) [2023-04-24]. https://doi.org/10.48550/arXiv.1707.07250.

[13] LIU Z, SHEN Y, 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. Stroudsburg: ACL, 2018: 2247-2256.

[14] TSAI Y H H, BAI S J, 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. Stroudsburg: ACL, 2019: 6558-6569.

[15] YU W M, XU H, YUAN Z Q, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto :AAAI, 2021: 10790-10797.

[16] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04) [2023-04-24]. https://arxiv.org/abs/1409.1556.

[17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition

(CVPR). Piscataway: IEEE, 2016: 770-778.
[18] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11966-11976.
[19] BALTRUSAITIS T, ZADEH A, LIM Y C, et al. OpenFace 2.0: facial behavior analysis toolkit[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Piscataway: IEEE, 2018: 59-66.
[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2023-04-24]. https://arxiv.org/abs/2010.11929.
[21] DESAI S, RAMASWAMY H G. Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization[C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2020: 972-980.
[22] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. (2019-09-26)[2023-04-24]. https://arxiv.org/abs/1909.11942.
[23] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-11-11)[2023-04-24]. https://doi.org/10.48550/arXiv.1810.04805.
[24] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration[EB/OL]. (2019-04-19)[2023-04-24]. https://doi.org/10.48550/arXiv.1904.09223.
[25] CUI Y M, CHE W X, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[EB/OL]. (2020-04-29)[2023-04-24]. https://doi.org/10.48550/arXiv.2004.13922.
[26] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2023-04-24]. https://doi.org/10.48550/arXiv.1907.11692.
[27] LUO H S, JI L, ZHONG M, et al. CLIP4Clip: an empirical study of CLIP for end to end video clip retrieval and captioning[J]. Neurocomputing, 2022, 508: 293-304.

Multimodal Sentiment Analysis Model Based on CLIP and Cross-attention

CHEN Yan[1,2], LAI Yubin[1], XIAO Ao[1], LIAO Yuxiang[1], CHEN Ningjiang[1]

(1. School of Computer and Electronic Information Science, Guangxi University, Nanning 530000, China; 2. Guangxi Key Laboratory of Multimedia Communication and Network Technology, Guangxi University, Nanning 530000, China)

Abstract: In response to the issues of limited annotated data, insufficient fusion between modalities, and information redundancy in multimodal sentiment analysis, a multimodal sentiment analysis model called CLIP-CA-MSA based on contrastive language-image pretraining(CLIP) and cross-attention mechanism was proposed in this study. This model employed models such as BERT which was pre-trained by CLIP, and PIFT to extract feature vectors from videos and textual content. Subsequently, a cross-attention mechanism was applied to facilitate interaction between image feature vectors and text feature vectors, enhancing information exchange across different modalities. Finally, the uncertainty loss was utilized to compute the fused features, and the ultimate sentiment classification results were generated from the outputs. The experimental results showed that the model could increase accuracyrate by 5 percentage points to 14 percentage points and the $F1$ value by 3 percentage point to 12 percentage point over other multimodal models, which verifieed the superiority of the model in this study. And uses of ablation experiments to verified the validity of each module of the model. This model could effectively utilize the complementarity and correlation of multimodal data, and utilize uncertainty loss to improve the robustness and generalization ability of the model.

Keywords: sentiment analysis; multimodal learning; cross-attention; CLIP model; Transformer; feature fusion