引用格式：郑智龙，张兴茂，王文智，黄清宝，程皓楠. 基于对比语言-图像预训练的情感增强仇恨模因检测[J]. 信息传播研究，2024，31(06)：69-76.
文章编号：2097-4930(2024)06-0069-08

基于对比语言图像预训练的情感增强仇恨模因检测

郑智龙[1]，张兴茂[2]，王文智[1]，黄清宝[1*]，程皓楠[3]
（1.广西大学电气工程学院，南宁 530004；2.广西艺术学院通识教育学院，南宁 530022；3.中国传媒大学媒体融合与传播国家重点实验室，北京市 100024）

摘要：仇恨模因检测是一项具有挑战性的多模态任务，需要模型理解视觉与语言中的隐含语义，并进行跨模态理解交互。针对中文领域的仇恨模因检测任务，本文构建了一个数据集CHmemes，并设计了一个基于CLIP（contrastive language-image pre-training）的情感增强Transformer模型（E2TC，emotion-enhanced Transformer model based on CLIP）作为基线模型。该模型利用图像和文本中的情感信息来增强从CLIP中提取到的特征，然后结合图像中与仇恨相关的人物属性信息以提高模型对于图像中仇恨内容的关注度。最后，采用图像描述作为监督机制以防止模型过拟合。所提出的E2TC模型在CHmemes数据集上以77.67%的AUROC值和72.71%的准确率超越了多个对比模型，验证了情感特征和图像属性信息对于仇恨模因检测的重要性。

关键词：CLIP；中文仇恨模因检测；情感增强
中图分类号：TP391　**文献标识码：**A

Emotion enhanced hateful meme detection based on contrastive language image pre-training

ZHENG Zhilong[1], ZHANG Xingmao[2], WANG Wenzhi[1], HUANG Qingbao[1*], CHENG Haonan[3]
(1. School of Electrical Engineering, Guangxi University, Nanning 530004, China; 2. College of General Education, Guangxi Arts University, Nanning 530022, China; 3. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China)

Abstract: Hateful meme detection is a challenging multimodal task that requires models to comprehend implicit semantics in both visual and textual domains and engage in cross-modal understanding. Addressing hate meme detection in the Chinese context, in this paper a dataset, named CHmemes, was introduced and a CLIP (contrastive language-image pre-training)-based emotion-enhanced Transformer model was designed as a baseline. In this model emotional features from both images and text were leveraged to enhance the features extracted from CLIP, then attribute information related to hate in images was incorporated to boost the model's attention towards hate-related content in images. Finally, image descriptions as a supervision mechanism was employed to prevent overfitting. The proposed model achieves superior performance, surpassing multiple comparison models on the CHmemes dataset with metrics AUROC 77.67% and Acc 72.71%, validating the significance of emotional features and image attribute information for hate meme detection.

Keywords: CLIP; Chinese hateful meme detection; emotion enhancement

基金项目：媒体融合与传播国家重点实验室（中国传媒大学）开放课题（SKLMCC2023KF005）；广西自然科学基金重点项目(2024JJD170001)；国家自然科学基金（62276072）

作者简介(*为通讯作者)：郑智龙（2000-），男，硕士研究生，主要从事多模态情感计算研究。Email:zhilong.zheng@st.gxu.edu.cn；张兴茂（1986-），女，博士，副教授，主要从事语言与文化传播研究。Email:20170024@gxau.edu.cn；王文智(1998-)，男，硕士研究生，主要从事网络语言治理研究。Email:2112391053@st.gxu.edu.cn；黄清宝（1979-），男，博士，教授，主要从事多媒体计算研究。Email:qbhuang@gxu.edu.cn；程皓楠（1994-），女，博士，副研究员，主要从事视听内容计算研究。Email:haonancheng@cuc.edu.cn

1 引言

网络表情包(internet memes)能够生动直观地展现个人的情绪,已经成为交流思想、表达情感和讨论社会问题的主要模式之一。然而由于社交平台过滤机制的不完善,无法有效识别复杂表情包,部分表情包被用来攻击特定个人或群体。因此仇恨表情包检测,即仇恨模因检测已经成为学术社区的研究热点之一。

仇恨模因采取了复杂和微妙的策略来表达一些计算机难以识别的仇恨言论,这些仇恨言论以隐式仇恨居多。该任务的重点和难点在于如何有效结合不同模态的输入,理解多样性和复杂性的表达形式以及识别不同语境下的意图。早期研究[1-2]的一些方法将仇恨模因检测视为多模态任务的一个下游子任务,利用预训练的视觉语言模型在仇恨模因数据上进行微调,并链接外部知识库[3-4]等方法增强模型性能。另外的方法[5-7]是结合预训练模型(如BERT)与特定任务模型进行端到端的调整与改进。随着大模型的兴起,研究工作[8-9]利用视觉语言大模型的零样本能力将图像信息转换为文本信息,与输入文本结合,以便更好利用大语言模型中的上下文背景知识,将多模态任务转化为纯文本任务。但是以上方法依然存在一些问题:1)利用大模型将多模态任务转化为纯文本任务可能会忽略图像中仇恨检测相关的重要信息;2)这些方法忽略了情感信息对于模型效果的影响。

基于上述调研与思考,本文设计了一个基于对比语言图像预训练(CLIP, contrastive language image pre-training)[10]的情感增强的多模态模型(E2TC,emotion-enhanced Ttransformer model based on CLIP),将图像和文本情感与对应模态特征显式融合,利用情感信息增强模型对于仇恨相关信息的理解,同时提取图像中的人物属性信息,最后在多模态融合阶段利用自注意力机制联合两种模态特征,建立模态内和模态间关系,进行最终的预测。本文的贡献总结如下:1)构建了第一个中文仇恨模因检测数据集CHmemes,其中包含4254张含有场景文本的图像;2)设计了一个基于CLIP的情感增强多模态模型E2TC作为中文仇恨模因检测的基线模型;3)在CHmemes及另外两个数据集上丰富的实验结果证明了本文所提出模型的有效性。

2 相关工作

前期一些研究将其他多模态任务模型迁移到仇恨模因检测并进行微调。Zhang等人[11]设计了互补视觉和语言网络,在视觉和语言嵌入中添加了上下文和敏感对象信息来增强表示。Deshpande等人[12]通过增强输入特征以提供更丰富的信息进行仇恨模因检测,具体包括文本端的命名实体、情感信息和语义信息,以及图像端的描述、目标检测和网络实体。Pramanick等人[5]构建一个考虑整体和局部特征信息的多模态仇恨模因检测新框架。Kumar等人[6]提出一个端到端模型,利用FIM(feature interaction matrix)对从CLIP编码器获得的图像和文本表示之间的跨模态交互进行显式建模。Cao等人[13]提出基于提示的模型PromptHate,构建了简单的提示并提供一些上下文示例,利用RoBERTa语言模型中的隐式知识进行仇恨模因分类。Burbi等人[7]提出一种多模态仇恨模因分类方法,使用预训练CLIP和文本反转技术来有效捕获模因的多模态语义信息。

仇恨模因检测需要模型拥有丰富的先验知识,因此部分研究将知识引入仇恨模因检测。Kougia等人[3]使用场景图,利用对象及其视觉关系来表达图像,将命名实体识别检测出的实体知识图谱引入仇恨模因分类,通过外部知识库与背景对象的联系增强仇恨模因检测。Liu等[4]人将图卷积神经网络应用于仇恨模因检测,先获取图像中的实体信息,再利用外部知识库资源评估实体与模因文本间的联系,构建跨域图并进行图卷积增强图像和文本特征。

随着大模型的出现,部分研究利用大模型来解决仇恨模因问题。Dai等人[8]对基于预训练的BLIP-2[14]模型的视觉语言指令调优进行了研究,引入一个指令感知的查询转换器,提取不同指令的信息特征,在零样本仇恨模因检测任务上取得较好的性能。Cao等人[9]提出一种基于探测的图像描述方法,通过询问仇恨内容相关问题来提示冻结的预训练视觉语言模型,将答案用作图像标题,使得标题包含对仇恨内容检测的信息。

3 E2TC模型

仇恨模因检测旨在判断一个给定的包含中文文本的图像是否存在仇恨。仇恨模因检测任务要求模型不仅要融合图像和文本特征实现跨模态对齐,还需要有丰富的先验知识。因此本文提出一种基于CLIP的情感增强仇恨模因检测方法,模型框架如图1所示,包括特征提取模块、情感特征提取模块、图像人物属性提取模块和图像描述监督模块。该模型首先利用

冻结的CLIP模型提取图像和文本特征；然后通过一个可训练的投影层，使得CLIP特征能够适应特定的仇恨言论检测任务；接下来将情感特征分别与对应模态的特征融合获得情感增强的图像特征和文本特征；接着模型提取了图像中人物的多种属性信息增强对于图像中仇恨信息的关注；最后为了防止模型过拟合，设计了一个图像描述作为监督信号使得模型能够对全局和局部特征的关注保持良好的平衡。

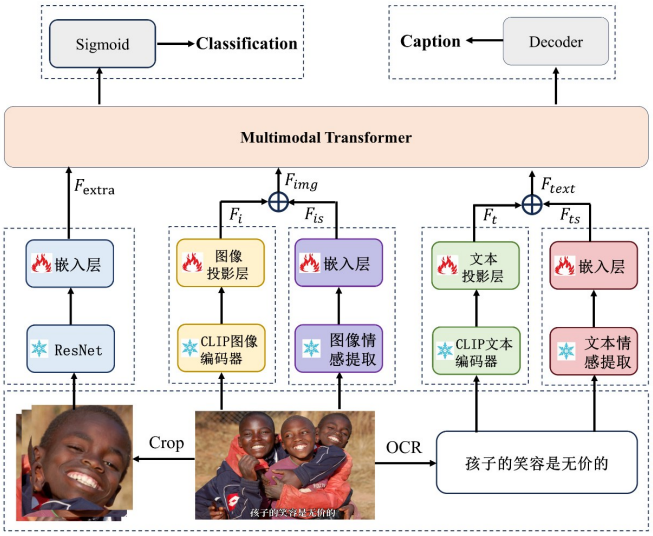

图1 E2TC模型结构图

3.1 特征提取

本文使用Chinese-CLIP提取图像视觉特征。对于视觉特征，首先将图像I缩放到指定的大小，利用冻结的Chinese-CLIP模型的视觉分支的ViT[15]模型来提取模因的视觉特征F_i，如式(1)：

$$F_i = \mathrm{CLIP}(I) \tag{1}$$

对于文本特征，首先利用百度智能云的API提取输入图像中的场景文本，提取出的文本表示为T，然后利用与视觉特征相同的Chinese-CLIP模型的语言分支的BERT[16]模型来提取模因中的文本特征F_t，如式(2)：

$$F_t = \mathrm{CLIP}(T) \tag{2}$$

Chinese-CLIP预训练的图像和文本对通常传达相同的含义，然而在仇恨模因中的部分隐式仇恨是通过讽刺、隐喻等修辞传达仇恨，图像和文本信息可能表达完全相反的信息。因此，为了更好地建模模因的图像和文本特征空间之间的语义关系，本文在CLIP图像和文本编码器的输出处添加可训练的投影层，使得图像和文本之间更准确地对齐。经过投影层的图像特征和文本特征分别表示为F_i'和F_t'，如式(3)、式(4)：

$$F_i' = \mathrm{Project}_i(F_i) \tag{3}$$

$$F_t' = \mathrm{Project}_t(F_t) \tag{4}$$

其中，$\mathrm{Project}_i$和$\mathrm{Project}_t$分别表示图像和文本的可训练投影层。

3.2 图像人物属性提取

本文采用Joo等人[17]提出的人脸表征识别方法，在大规模面部数据集FairFace上训练一个人物属性提取器。将该预训练模型直接应用于本文的仇恨模因图像，假设输入的一张模因图像I中检测到的人脸数量为n，每张图像中的所有任务的三种属性信息f_{race}，f_{gender}和f_{age}分别表示为式(5)、式(6)、式(7)：

$$f_{\mathrm{face}} = \{\ \mathrm{ResNet34}_{\mathrm{face}}(I_{\mathrm{facen}})\ \}_{n=1}^{N} \tag{5}$$

$$f_{\mathrm{gender}} = \{\ \mathrm{ResNet34}_{\mathrm{gender}}(I_{\mathrm{gendern}})\ \}_{n=1}^{N} \tag{6}$$

$$f_{\mathrm{age}} = \{\ \mathrm{ResNet34}_{\mathrm{age}}(I_{\mathrm{agen}})\ \}_{n=1}^{N} \tag{7}$$

将这些信息拼接聚合后通过一个嵌入层，映射到和图像特征相同的维度空间，嵌入层的输出表示为式(8)：

$$f_{\mathrm{extra}} = \mathrm{embed}(cat(f_{\mathrm{race}}, f_{\mathrm{gender}}, f_{\mathrm{age}}) \tag{8}$$

将f_{extra}作为图像的额外知识与其他特征共同输入到编码器中，以辅助模型识别图像中与仇恨相关的实体信息。

3.3 情感特征提取

图像情感特征大致可以分为浅层视觉特征、中层视觉特征和深层视觉特征。本文采用Zhang等人[18]提出的图像情感提取方法，使用在FI[19]上训练的图像情感提取模型获得图像情感特征。图像情感特征提取过程如图2所示：

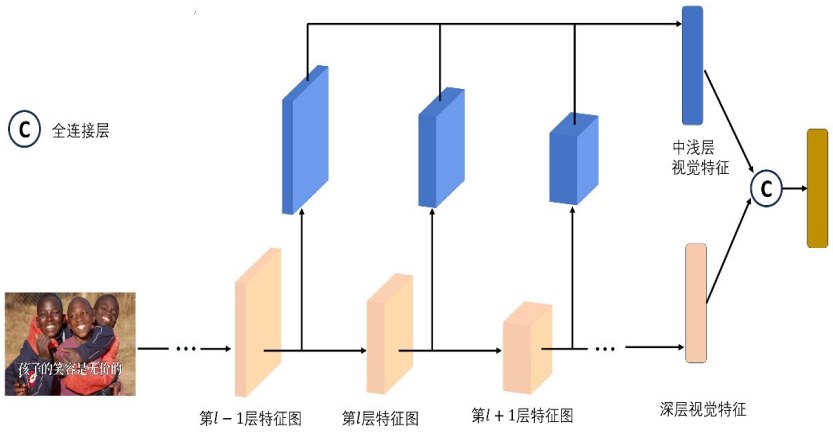

图2 图像情感特征提取模型结构图

本文利用ResNet50作为主干网络提取高级语义特征，在ResNet浅层网络延伸分支提取低级特征和中级特征，并将浅层网络的输出特征与高级语义特征融合联合判断图像的情感。具体来说，分别从ResNet的5个不同网络深度的卷积层提取出不同尺度的图像特征来捕获浅层情感特征f_{ci}，其中i = 1,2,3,4,5，表示网络深度的不同程度。将分支网络得到的5层情感特征与主干网络输出的情感特征相连接聚合，得到一组情感特征表示，将联合情感特征送入全连接层得到最终的融合多层视觉特征的图像情感特征。

对于文本情感，本文直接使用预训练的RoBERTa模型来提取文本的句子级情感特征。

3.4 多模态融合

多模态融合模块首先将输入信息预处理中获取的图像特征和文本特征分别与情感提取模块中的图像情感特征和文本情感特征映射到相同的维度聚合连接，得到图像特征与文本特征的最终表示F_{img}和F_{text}。然后将图像特征、文本特征和额外的图像知识特征映射到相同维度的特征空间拼接融合，多模态编码器的输入F_{multi}为三种特征的拼接，如式(9)：

$$F_{\text{multi}} = cat\left(LN\left(W_1 F_{\text{img}}\right), LN\left(W_2 F_{\text{text}}\right), LN\left(W_3 F_{\text{extra}}\right)\right) \tag{9}$$

其中W_1，W_2和W_3是可学习参数，LN是归一化。最后利用Transformer的自注意力机制融合输入特征，输出融合后的图像特征。

3.5 图像描述监督

视觉编码器可能只学习对仇恨模因检测有利的简化图像特征，忽略部分细节或者背景语义，因此可能会导致模型过拟合，进而影响模型在其他数据集上的泛化性。为了减少这种潜在偏差，本文利用模因图像的图像描述作为监督信号，让模型能够同时关注图像的全局特征而不是仅仅关注图像中对仇恨模因检测有利的局部特征。监督模块的具体实现如图3所示，首先利用Koutlis等人[20]提出的视觉部分利用率(VPU, visual part utilization)模型对模因图像进行裁剪，获取去除模因文本信息的纯视觉图像;然后将其输入冻结的预训练视觉语言大模型，设计prompt利用大模型的零样本能力生成裁剪后图像的图像描述，将其作为图像描述监督模块的参考答案。最后，在分类分支之外额外添加一个图像描述生成分支，以编码器输出的融合图像特征作为输入，经过一个Transformer解码器生成图像描述。

3.6 损失函数

E2TC模型包括两个训练目标，一个是分类性能的准确性，一个是分支图像描述生成的准确性。对于第一部分，使用分类任务中常用的二元交叉熵损失函数，计算公式如式(10)所示：

$$L_{\text{class}} = y\log\left(\hat{y}\right) + (1 - y)\log\left(1 - \hat{y}\right) \tag{10}$$

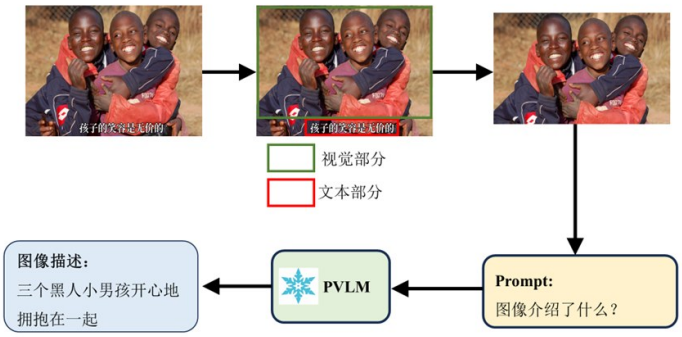

图3 图像描述监督模块

其中，y表示样本的真实标签，$\hat{y}$表示模型的预测概率。

第二部分是生成模型的损失函数，图像描述是以自回归的方式生成，假设t时刻输出结果的条件概率分布为$P(y_t|y_{<t},I,T)$，如式(11)所示：

$$P(y_t|y_{<t},I,T)=Softmax([D_{\text{out}}]_t) \tag{11}$$

其中，$[D_{\text{out}}]_t$表示解码器输出的矩阵特征的第t个特征向量。根据该条件概率分布可以获得t时刻生成单词y_t；然后将t时刻生成的特征向量与解码器的输入特征拼接作为下一时刻的输入特征；重复执行该过程直到生成内容满足长度要求或生成结束符号为止。最终生成结果的概率分布可以表示为式(12)：

$$P(Y)=\sum_{t=1}^{M}p(y_t|y_{<t},I,T) \tag{12}$$

图像描述模块最终的损失函数为生成的描述与大模型生成的描述之间的交叉熵损失，如式(13)：

$$L_{\text{cap}}=-\sum_{t=1}^{M}\hat{p}(y_t)\log P(y_t|y_{<t},I,T) \tag{13}$$

其中，$\hat{p}(y_t)$为大模型生成的图像描述中第t个中文字符的特征编码。

模型最终的损失函数是上述两部分函数的结合，计算公式如式(14)所示：

$$L=L_{\text{class}}+\mu L_{\text{cap}} \tag{14}$$

4 实验

4.1 数据集及评估指标

CHmemes是本文为中文仇恨模因检测任务构建的数据集，数据包括微博中爬取的图像和英文数据集的翻译图像两部分，共4254张图像，每张图像都有嵌入图像的文本描述。

微博图像爬取过程：利用种族、性别、地域、LGBTQ和其他五个主题下的关键词爬取仇恨相关数据，关键词搜索只关注与特定词语相关的微博数据，能够从海量的噪声数据中快速识别并定位到感兴趣的相关内容。对于图像中包含场景文本的数据直接作为数据集中数据；对于图像中不存在文本数据的图像，将微博文本嵌入图像作为数据集数据，最终得到2754张图像。

英文翻译数据：微博爬取图像中数据分布不平衡，仇恨图像远多于非仇恨数据。为了平衡数据集分布，我们从英语公共数据集Hateful Memes中筛选出1500张非仇恨图像，将图像中英文翻译为中文作为数据集的一部分，最终的数据集包含4254张图像，其中包含仇恨图像2352张，非仇恨数据1902张。

图4展示了数据集的一部分样本，图中上半部分的两张图像来自于微博的数据，下半部分的两张图像来自于Hateful Memes。

图4 CHmemes数据集样本

为了验证所提出的E2TC模型的泛化性和可迁移性，本文还在另外两个公共英文数据集上进行了实验。Hateful Memes[1]数据集(FHM)是Facebook专门为研究仇恨模因检测任务设计的，包括12000张带有场景文本的图像，每张图像都进行了仇恨和非仇恨的二元标注，且对数据集中的仇恨和非仇恨样本进行了较好的平衡。HarMeme[2]数据集是一个针对仇恨模因检测构建的基准数据集，包含3544个从现实来源中收集的COVID-19相关的模因。

延续先前仇恨模因检测任务，本文使用准确率Acc和AUROC值两个指标来评估模型性能。

4.2 参数设置

对于图像，训练集图像被重新调整大小并统一填充到16*16，然后使用随机裁剪、缩放、翻转等多种手段对图像进行数据增强，增强模型的泛化性。对于文本，所有文本中的空格和标点符号都被删除。词汇库的大小由大语言模型生成的图像描述中至少出现3次的单词数量决定，通过统计由大语言模型生成的图像描述的句子长度，设置实验中使用的最大分词长度为30。对于输入编码器的所有特征向量，设置其投影特征空间的维度。为了生成裁剪后的图像描述，本文使用BLIP2作为图像描述生成模型。在训练中，本文利用小批量优化策略，使用AdamW优化器，模型的初始学习率设置为1×10^{-5}，使用学习率衰减策略，随着训练时间和结果自动调整学习率。模型中提取视觉和语言特征的Chinese-CLIP和生成图像描述的预训练模型BLIP2的权重是完全冻结的。对于不同的数据集采用不同的迭代轮次：在本文提出的数据集上总共迭代30个epoch，在FHM和另一个数据集上总共迭代50个epoch。

4.3 实验结果

在CHmemes数据集上的对比实验结果如表1所示（黑体表示性能最优）。本文所提出的E2TC模型在验证集和测试集上的准确率Acc和AUROC值指标都超越了所有的对比模型，说明对于中文仇恨模型检测，E2TC模型的性能优于其他对比模型。

其中，在测试集上，E2TC与单文本模态的BERT[16]模型相比，准确率和AUROC指标分别提升了7.51%和9.68%；与单图像模态的Image-Region模型相比，准确率和AUROC指标分别提升了19.66%和19.96%；与多模态的三个模型VisualBERT[21]、Hate-CLIPper[6]和Pro-Cap[9]相比，准确率分别提升了4.81%、2.43%和3.87%，AUROC值分别提升了5.80%、3.75%和2.05%。

表1 Chmemes数据集上E2TC与其他方法的性能比较

	模型	验证集		测试集	
		AUROC（%）	Acc（%）	AUROC（%）	Acc（%）
单模态	BERT	64.77	62.62	62.43	61.73
	Image-Region	54.19	50.48	52.15	49.58
多模态	VisualBERT	68.15	66.50	66.31	64.43
	Hate-CLIPper	73.11	69.47	68.36	66.81
	Pro-Cap	74.55	70.21	70.06	65.37
	E2TC	**77.67**	**72.71**	**72.11**	**69.24**

为了验证提出的E2TC模型的泛化性和可迁移性，本文还在常用的英文数据集FHM[1]和HarMeme[2]上进行了对比实验，表2展示了测试集上的结果。

虽然E2TC模型并没有在这两个数据集上取得最优结果，但是在FHM数据集上，E2TC的AUROC性能仅次于最优的Hate-CLIPper模型，结果相差2.11%；在HarMemes数据集上，E2TC的AUROC值和准确率与最优的Pro-Cap模型相当，分别相差0.23%和0.17%。

综合分析两个英文数据集（表2）和CHmemes数据集（表1）上的测试集实验结果，E2TC模型在FHM和CHmemes上的AUROC值和准确率Acc差距分别为9.07%（81.18% vs.72.11%）和4.05%（73.29% vs.69.24%），Pro-Cap模型在两个指标上的差距分别为10.81%和6.91%。在HarMemes数据集和CHmemes数据集上，E2TC模型AUROC值和准确率差距分别为17.91%和13.84%；Pro-Cap模型在两个指标上的差距分别为20.19%和17.88%。相对于其他模型，E2TC模型在不同数据集上的性能差距更小，表明针对不同语言和不同数据分布的数据集，E2TC有更高的鲁棒性和更强的泛化性。

表2 FHM和HarMemes数据集上的对比实验结果

	模型	FHM数据集		HarMeme数据集	
		AUROC（%）	Acc（%）	AUROC（%）	Acc（%）
单模态	BERT	66.10	57.12	81.39	75.68
	Image-Region	56.69	52.34	76.46	73.05
多模态	VisualBERT	68.71	61.48	80.46	75.31
	Hate-CLIPper	**83.29**	–	–	–
	Pro-Cap	80.87	72.28	**90.25**	**83.25**
	E2TC	81.18	**73.29**	90.02	83.08

为了验证可训练投影层、情感特征提取模块、图像人物属性提取和图像监督模块的影响，本节对E2TC模型进行了消融实验，实验结果如表3所示，所有消融实验都在验证集上进行且参数设置与实验设置中的参数一致，以此来证明设计模块的有效性。

表3 四个主要成分的消融实验结果

可训练投影层	情感特征	人物属性	图像描述监督	AUROC(%)
	√	√	√	73.41
√	√		√	74.91
√		√	√	75.33
√	√	√		76.22
√	√	√	√	**77.67**

表3的第一行显示了移除可投影训练层的实验结果，不使用投影层对从CLIP中提取的特征进行训练，准确率下降了4.26%，说明原始的CLIP特征并不完全适用于仇恨模因检测，可训练的投影层有利于文本和图像在特征空间实现更好的对齐，增强模型对于输入信息的理解。第二行是移除图像人物属性提取模块的实验结果，移除该模块后模型的输入将不包含图像中人物的种族、年龄、性别等属性信息，在准确率上下降了2.76%。仇恨模因大多是针对特定群体或者个人的，图像中人物的属性信息对于仇恨检测至关重要，因此移除该模块后模型整体性能出现明显的下降。第三行是移除情感特征提取模块的实验结果，移除该模块后模型将不会把图像和文本的情感特征显示融合到CLIP特征中，在准确率上下降了2.34%。仇恨可以看做是一种消极的情感，虽然CLIP提取的特征能够以隐式的方式获取情感信息，但是将情感特征显式融合能够增强模型对于情感特征的关注，提高模型的检测性能。第四行是移除图像描述监督模块的实验结果，移除该模块后模型的训练目标只有一个分类损失函数，模型在正确率和AUROC上下降了1.45%。图像监督模块能够使模型关注图像的全局信息，而不是只关注于仇恨相关的局部特征，能够防止模型过拟合，同时提高模型在其他数据集上的泛化能力。

4.4 定性分析

表4展示了E2TC模型和Pro-Cap模型在CHmemes数据集上的预测结果对比，直观地介绍了该任务的形式和本文所提出的E2TC模型相比于Pro-Cap模型的优势。

表4 E2TC与Pro-Cap案例分析比较

图像		
文本	就喜欢看这样的情侣	伊斯兰教是和平的宗教， 不要批判我的宗教
Pro-Cap	非仇恨	仇恨
E2TC	仇恨	非仇恨
真实标签	仇恨	非仇恨

表4左边的图像描述的是在闹矛盾的情侣，其中的女生漫不经心，文本内容是“就喜欢看这样的情侣”，整体表达的是对别人闹矛盾的幸灾乐祸，属于比较中性的仇恨，Pro-Cap错误地预测为“非仇恨”，主要原因是Pro-Cap利用的大模型生成了错误的图像描述“一对男女坐在长椅上”，由于对女生负面面部表情描述的缺失，导致了错误预测。而本文的E2TC没有忽视视觉特征，准确判断了图像中人物表情传达的情感，正确预测为“仇恨”。对于表4右边的图像，Pro-Cap错误地预测为“仇恨”，可能是因为在训练中出现穆斯林相关的模因仇恨的居多，因此Pro-Cap会倾向于预测这个模因是仇恨的，这也是目前大多数仇恨言论检测模型的局限性，在检测特定内容时容易出现偏见，导致检测结果错误。而本文的E2TC模型将情感特征显式地融入图像和文本特征，以情感特征作为辅助信息来避免这种偏见。在这个例子中，图像描述的是“一个面带笑容的穆斯林男人”，图像情感是高兴，属于积极情感；文本内容是“伊斯兰教是和平的宗教，不要批判我的宗教”，文本是呼吁别人理解自己的宗教，总体是正面情感，E2TC能够结合情感信息正确判断模因是属于“非仇恨”的。

表5展示了一些本文模型预测错误的例子。表5左边图像的真实标签是“仇恨”，E2TC预测错误的原因可能是模型推理能力的不足，该例子中的图像是“两个女性穿着婚纱在一起亲吻”，图像情感是快乐幸福，属于积极情绪；文本内容是“科学家正在努力治愈她们”，传达的也是正面的情感信息，因此E2TC错误地预测为“非仇恨”。模型需要推理出图像认为同性恋是一种病，科学家治愈的是“同性恋”这种病，才能理解该模因是对“同性恋”群体的不尊重和仇恨，并做出正确的答案预测。表5右边的真实标签也是“仇恨”，图像中的人是一个变性人，其原始的生理性别是男，但由于外观特征发生变化导致E2TC模型识别图像中的人为女性，E2TC的图像人物属性模块错误地识别其性别信息，导致预测为“非仇恨”。这表明人脸属性提取模块对于一些类别数量较少的人物识别成功率较低，需要提升识别性能，才能解决长尾分布的问题。

表5 典型类别的错误示例

图像		
文本	科学家正在努力治愈她们	世界很精彩 不需要性别多样性了
E2TC	非仇恨	非仇恨
真实标签	仇恨	仇恨

5 总结与展望

本文针对仇恨模因检测中隐式仇恨难以检测的问题，构建了一个中文仇恨模因数据集并提出一种基于CLIP的情感增强模型，将情感特征与CLIP特征显示融合，使得模型能够高效融合图像与文本语义特征。此外，还设计了一个图像人物属性提取器，使得模型能够关注与仇恨相关的图像人物属性，并利用图像描述作为监督模块，保证模型能够同时关注图像整体和局部特征。在多个数据集上的对比实验和消融实验证明了提出模型的有效性。

参考文献（References）：

[1] Kiela D，Firooz H，Mohan A，et al. The hateful memes challenge：detecting hate speech in multimodal memes[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems，2020，33：2611-2624.

[2] Pramanick S，Dimitrov D，Mukherjee R，et al. Detecting harmful memes and their targets[C]// Findings of the Association for Computational Linguistics：ACL-IJCNLP 2021，2021：2783-2796.

[3] Kougia V，Fetzel S，Kirchmair T，et al. Memegraphs：linking memes to knowledge graphs[C]// International Conference on Document Analysis and Recognition，2023：534-551.

[4] 刘旭东，杨亮，张冬瑜，等. 结合图卷积网络的多模态仇恨迷因识别研究[J]. 重庆理工大学学报（自然科学），2024，38(1)：169-179.

[5] Pramanick S，Sharma S，Dimitrov D，et al. MOMENTA：a multimodal framework for detecting harmful memes and their targets[C]// Findings of the Association for Computational Linguistics：EMNLP 2021，2021：4439-4455.

[6] Kumar G K，Nandakumar K. Hate-CLIPper：multimodal hateful meme classification based on cross-modal interaction of CLIP features[C]// Proceedings of the Second Workshop on NLP for Positive Impact（NLP4PI），2022：171-183.

[7] Burbi G，Baldrati A，Agnolucci L，et al. Mapping memes to words for multimodal hateful meme classification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision，2023：2832-2836.

[8] Dai W，Li J，Li D，et al. InstructBLIP：towards general-purpose vision-language models with instruction tuning[C]// Advances in Neural Information Processing Systems，2024，36.

[9] Cao R，Hee M S，Kuek A，et al. Pro-cap：leveraging a frozen vision-language model for hateful meme detection[C]// Proceedings of the 31st ACM International Conference on Multimedia，2023：5244-5252.

[10] Radford A，Kim J W，Hallacy C，et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning，2021：8748-8763.

[11] Zhang W，Liu G，Li Z，et al. Hateful memes detection via complementary visual and linguistic networks[DB/OL]. arXiv：2012.04977，2020.

[12] Deshpande T，Mani N. An interpretable approach to hateful meme detection[C]// Proceedings of the 2021 International Conference on Multimodal Interaction，2021：723-727.

[13] Cao R，Lee R K W，Chong W H，et al. Prompting for multimodal hateful meme classification[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing，2022：321-332.

[14] Li J，Li D，Savarese S，et al. BLIP-2：bootstrapping language-image pre-training with frozen image encoders and large language models[C]// Proceedings of International Conference on Machine Learning，2023：19730-19742.

[15] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16x16 words：Transformers for image recognition at scale[C]// Proceedings of International Conference on Learning Representations，2020.

[16] Devlin J，Chang M W，Lee K，et al. BERT：pre-training of deep bidirectional Transformers for language understanding[C]// Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2019：4171-4186.

[17] Kärkkäinen K，Joo J. FairFace：face attribute dataset for balanced race，gender，and age for bias measurement and mitigation[C]// IEEE Winter Conference on Applications of Computer Vision，2021：1547-1557.

[18] Zhang H，Xu D，Luo G，et al. Learning multi-level representations for affective image recognition[J]. Neural Computing and Applications，2022，34(16)：14107-14120.

[19] You Q，Luo J，Jin H，et al. Building a large scale dataset for image emotion recognition：the fine print and the benchmark[C]// Proceedings of the AAAI Conference on Artificial Intelligence，2016，30(1).

[20] Koutlis C，Schinas M，Papadopoulos S. MemeTector：enforcing deep focus for meme detection[J]. International Journal of Multimedia Information Retrieval，2023，12(1)：11.

[21] Li L H，Yatskar M，Yin D，et al. Visualbert：a simple and performant baseline for vision and language[DB/OL]. arXiv：1908.03557，2019.

编辑：赵志军

引用格式：李枝勇，管悦，魏馨，王栋晗，颜迎晨. AIGC驱动下视听内容定制化：前沿探索与趋势分析[J]. 信息传播研究，2024，31(06)：77-88.

文章编号：2097-4930(2024)06-0077-12

AIGC驱动下视听内容定制化：前沿探索与趋势分析

李枝勇[1,2]，管悦[2]，魏馨[2*]，王栋晗[1,2]，颜迎晨[3,4]

（1. 中国传媒大学媒体融合与传播国家重点实验室，北京 100024；
2. 中国传媒大学经济与管理学院，北京 100024；
3. 北京航空航天大学经济管理学院，北京 100191；
4. 复杂系统分析与管理决策教育部重点实验室，北京 100191）

摘要：随着数字技术和人工智能的迅猛发展，视听内容定制化已成为现代传媒和娱乐领域的重要趋势。在需求侧，定制化的视听内容能够满足用户的个性化需求。在供给侧，定制化的视听内容一方面为视听内容创作者提供更加灵活和高效的创作工具，另一方面为视听内容平台吸引和留住更多的视听内容用户和视听内容创作者。本文旨在系统回顾和总结人工智能生成内容技术如何推动视听内容定制化的研究进展，重点讨论AIGC技术与数字艺术，及AIGC在视听内容定制化中的应用。文章从AIGC技术发展与数字艺术出发，进一步梳理视听内容与视听平台、数字产品定制化及AIGC治理角度下数字版权管理的相关研究，并讨论了面临的挑战及未来的研究方向。

关键词：视听内容；定制化；AIGC；数字艺术

中图分类号：G220.7;TN948　**文献标识码**：A

Audiovisual content customization driven by AIGC：frontier exploration and trend analysis

LI Zhiyong[1,2], GUAN Yue[2], WEI Xin[2*], WANG Donghan[1,2], YAN Yingchen[3,4]

(1.State Key Laboratory of Integration and Communication, Communication University of China, Beijing 100024, China; 2. School of Economics and Management, Communication University of China, Beijing 100024, China; 3. School of Economics and Management, Beihang University, Beijing 100191, China; 4. Key Laboratory of Complex System Analysis, Management and Decision (Beihang University), Ministry of Education, Beijing 100191, China)

Abstract: With the rapid advancement of digital technology and artificial intelligence, audio-visual content customization (ACC) has emerged as a significant trend in modern media and entertainment. From the demand side, customized audio-visual content is able to satisfy users' personalized needs. From the supply side, ACC not only provides more flexible and efficient creative tools for audio-visual content creators but also attracts and retains a larger user base and content creators for audio-visual content platforms. This review aimed to systematically summarize and synthesized the research progress on how

基金项目：国家自然科学基金面上项目(72472143)；国家自然科学基金重点项目(U21B20102)；国家自然科学基金青年项目(72202220)；国家自然科学基金青年项目(72001011)

作者简介(*为通讯作者)：李枝勇(1986-)，男，博士，副教授，主要从事平台商业模式与数字内容定价相关研究。Email：zyli@cuc.edu.cn；魏 馨(1996-)，女，博士，讲师，主要从事数字化管理决策与知识管理研究。Email：xinwei124@yeah.net；管悦(1993-)，女，博士，副教授，主要从事多模态数据分析与商务智能研究。Email：yueguan@cuc.edu.cn；王栋晗(1973-)，男，博士，教授，主要创新管理与战略管理相关研究。Email：dhwang@cuc.edu.cn；颜迎晨(1993-)，女，博士，副教授，主要从事线上平台运营模式与产品创新管理研究。Email：ychyan@buaa.edu.cn