结合常识知识的无监督情感化图像描述生成

齐雅昀[1),2)] 赵文天[3)] 李 彤[1),2)] 吴心筱[1),2),4)]

[1)](北京理工大学计算机学院 北京 100081)

[2)](智能信息技术北京市重点实验室 (北京理工大学) 北京 100081)

[3)](北京理工大学教育学院 北京 100081)

[4)](广东省智能感知与计算重点实验室 (深圳北理莫斯科大学) 广东 深圳 518172)

摘 要 图像与文本作为跨模态信息交互的两种基本形态,是人类日常生活中传达情感的重要载体。情感化图像描述生成不仅需要准确描述图像的视觉语义内容,还需合理传达出图像所蕴含的情感。现有方法通常依赖大量的图文配对数据进行训练,然而标注这些配对数据需耗费大量人力和物力。为此,本文首次提出无监督情感化图像描述生成任务。针对无监督条件下,数据驱动范式难以引导模型合理利用情感元素表达图像潜在情感的挑战,本文提出一种结合常识知识的无监督情感化图像描述生成方法。该方法从外部语料库中挖掘有关特定情感下实体与描述间关联的情感关系常识,并以视觉信息为线索将与图像高度相关的情感元素引入描述生成过程。为有效利用不成对的图像和情感语料进行训练,本文在强化学习框架中设计了一种新的情感元素奖励机制,鼓励模型采用与常识知识相符的方式生成图像的情感化描述。此外,本文还提出一个新的情感化描述评价指标SentiCLIPScore,用于综合评估描述的事实准确性和情感契合度。在COCO与Flickr30K两个图像数据集上的实验表明,相较于其他无监督基线方法,本文提出的方法在这两个数据集上的SentiCLIPScore分别提升了至少4%和13%。

关键词 无监督情感化图像描述生成;常识知识;情感关系;视觉情感分析

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.02809

Unsupervised Sentimental Image Captioning via Commonsense Knowledge

QI Ya-Yun[1),2)] ZHAO Wen-Tian[3)] LI Tong[1),2)] WU Xin-Xiao[1),2),4)]

[1)](School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081)

[2)](Beijing Key Laboratory of Intelligent Information Technology (Beijing Institute of Technology), Beijing 100081)

[3)](School of Education, Beijing Institute of Technology, Beijing 100081)

[4)](Guangdong Laboratory of Machine Perception and Intelligent Computing (Shenzhen MSU-BIT University), Shenzhen, Guangdong 518172)

Abstract Images and text serve as fundamental carriers for conveying emotions in daily human communication. Sentimental image captioning requires models to not only accurately describe the visual content but also appropriately express underlying visual sentiments. Compared with the conventional image captioning task that focuses purely on factual semantics, sentimental image captioning emphasizes the affective alignment between visual elements and linguistic expressions, making it particularly valuable for applications such as social media recommendation and human-computer interaction. Existing sentimental image captioning methods typically rely on large-scale pairs of images and sentimental captions. However, their annotation process is expensive, labor-intensive, and error-prone. Moreover, existing sentimental-related datasets mainly focus on

收稿日期:2025-07-16;在线发布日期:2025-10-14。本课题得到深圳市科技计划(No. JCYJ20241202130548062)、深圳市自然科学基金(No. JCYJ20230807142703006)、广东省教育厅重点研究平台与项目(No. 2023ZDZX1034)以及国家自然科学基金面上项目(No. 62072041)资助。 **齐雅昀**,博士研究生,主要研究领域为计算机视觉、多模态内容理解。E-mail: qiyayun@bit.edu.cn。**赵文天**,博士,预聘助理教授,主要研究领域为多模态内容理解、人工智能安全、教育技术学。**李 彤**,硕士,主要研究领域为计算机视觉。**吴心筱**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为视觉与语言、机器学习、图像视频理解。E-mail: wuxinxiao@bit.edu.cn。

single modality data with sentiment class labels or paired with texts crawled from social media posts that have large discrepancies with image descriptions, which cannot be used as the training data for sentimental image captioning. To address this limitation, we propose a novel task called unsupervised sentimental image captioning, which aims to generate image descriptions using inherent sentiments without requiring any paired image-sentence data for training. The main challenge lies in how to enable the model to express the underlying sentiment of the image by incorporating appropriate sentimental elements without any supervision. To tackle this challenging task, we propose a method that integrates commonsense knowledge of sentimental relationships into the caption generation process. This is inspired by the fact that human sentimental expressions usually follow certain rules and have specific describing patterns for different entities and emotion combinations. Our method consists of four key components, including a commonsense knowledge base of sentimental relationships, a factual sentence decoder, a sentimental sentence decoder, and a visual information extraction module. Specifically, the commonsense knowledge base of sentimental relationships is constructed from an external corpus, where the sentimental relationship represents the correlation between an entity and a sentimental description in a specific sentiment. Our method adopts a two-phase generation strategy, which first generates a factual sentence with masked sentimental parts, and then fills the masked parts with highly image-relevant sentimental words inferred from the commonsense sentimental relationships. To effectively train the model using unpaired images and sentimental corpus, we design a novel sentimental reward in reinforcement learning that aligns generated sentimental captions with commonsense knowledge. This new reward is calculated by evaluating how reasonable the generated sentimental words are, according to the commonsense knowledge of sentimental relationships, in order to encourage the model to pay more attention to the sentimental part of a sentence. Moreover, to address the problem of existing metrics that independently evaluate the content relevance and sentiment consistency, we propose a new metric called SentiCLIPScore. This novel metric jointly assesses both the factual and sentimental aspects of captions, where the content relevance is measured by the pre-trained multimodal model CLIP, and the sentimental consistency is synthetically measured by the sentence sentiment class and the constructed sentimental relationship knowledge base. Experiments on the COCO and Flickr30k image datasets demonstrate the efficacy of our method. Compared with unsupervised baselines, our method improves SentiCLIPScore by 4% and 13% on COCO and Flickr30K, respectively.

Keywords unsupervised sentimental image captioning; commonsense; sentimental relationship; visual sentimental analysis

1 引 言

图像描述生成是计算机视觉与自然语言处理领域中受到广泛关注的交叉融合课题。目前多数研究聚焦于生成与视觉内容语义一致且语法准确的描述[1-6]，忽视了图像所传达的情感信息。而一系列情感分析的工作[7-14]均指出图像和文本在人类传达情感方面的重要性。与一般的图像描述生成不同，情感化图像描述生成[15]旨在自动生成能体现出图像内在情感的图像内容语言描述，在社交媒体推荐、人机交互等场景中具有广泛的应用前景。

相较于风格化图像描述生成任务[16-21]预先设定描述的语言风格，图1(d)中所示的情感化图像描述生成加深了视觉与文本内容的情感关联，弱化了描述对语言风格的依赖，更注重对图像自身情感信息的探索，并生成更合理、更具情感的图像描述。相较于图1(a)与(b)中仅需预测情感类别的情感分析任务，情感化图像描述生成作为生成式任务，进一步要求对于如何用自然语言正确表达与视觉内容相关情

感的理解。情感化图像描述生成方法的训练过程依赖大量配对的图像与情感化描述数据。标注这些配对数据耗费大量时间精力，且标注结果易出现错误。尽管一些现有的视觉-文本情感分析数据集[22-23]包含从社交媒体帖子中收集的图文数据，其文本通常来自帖子的评论或标题。从图1(b)中的例子可以看出，这些文本与期望的情感化描述存在较大差异，难以直接用作训练数据。

图1　情感分析任务、传统图像描述生成任务与情感化图像描述生成任务的区别

为消除对图像-句子成对数据的依赖，本文首次提出无监督情感化图像描述任务，仅利用不存在对应关系的图像数据集和情感语料库，实现反映图像内在情感的描述生成。在没有监督信息支持的情况下，该任务面临如何赋能模型使用恰当情感元素（即常用于表达情感的形容词与副词）表达图像潜在情感的挑战。人类的情感表达通常遵循一定的规律，对不同的实体与情感组合有着公认的特定描述方式。例如，在表达积极情感时，人们倾向于使用“happy”来形容儿童或小动物，而不用其来形容食物。作为合理表达情感的基础，这类体现特定情感下实体与情感元素间关联的情感关系常识知识，可提升模型生成情感化图像描述的能力。

本文提出一种结合常识知识的无监督情感化图像描述生成方法。首先从情感语料库中构建情感关系常识知识库，并设计两阶段的描述生成过程，在生成情感部分被掩码的图像事实描述后，结合知识库中与图像高度相关的情感元素，最终实现情感化图像描述的生成。具体的，利用不成对的图像与事实语句训练一个现有的无监督图像描述模型[24]，实现图像事实描述生成的过程。针对描述中传达情感的部分，以图像的情感类别预测结果和物体检测结果为查询，从知识库中检索一系列视觉相关的情感关系。然后设计一个情感化语句解码器，以情感类别、检索出的情感关系以及上一阶段生成的事实描述为输入，输出最终的情感化描述，并通过重构情感语料库中的句子进行训练。为有效利用有限的不成对图文数据进行训练，引入强化学习并提出一个新的情感元素奖励机制，引导模型生成与常识知识相符的情感化描述。

在方法评估方面，情感化图像描述领域现有的评价指标孤立地评估描述的事实和情感部分。针对该问题，本文设计了一个名为SentiCLIPScore的新指标，借助预训练模型CLIP以及本文构建的知识库，综合评估描述在事实准确性与情感契合度方面的表现。

需要指出的是，情感化图像描述生成作为一个前沿的研究方向，目前已有的研究方法缺乏对无监督场景下情感建模的探索。本文在不依赖成对图文数据的条件下，引入情感关系常识知识来指导描述生成，这不仅在方法上具有独特性，也为缓解情感化图像描述领域的训练数据依赖问题提供了新的解决思路。同时，由于该任务能够帮助机器生成更具情感表达力和人性化的描述，其在社交媒体推荐、人机交互、情感陪伴以及数字内容创作等应用场景中具有潜在价值，拓展了图像描述生成研究在实际应用中的可能性。

本文的主要贡献包括：提出一种结合常识知识的无监督情感化图像描述生成方法，首次尝试在不依赖成对图文数据的情况下，自动生成反映图像内在情感的图像内容语言描述；构建情感关系常识知识库，并设计两阶段的描述生成过程，将情感关系常识知识中与图像高度相关的情感元素，融入情感化描述生成的过程；提出评估情感化图像描述的新指标SentiCLIPScore，在COCO与Flickr30K图像数据

集上的实验结果表明，相较于其他无监督基线方法，本文提出的方法在综合考虑描述事实准确性与情感契合度的新指标上分别提升了4%与13%。

2 相关工作

2.1 结合特定语言风格的图像描述生成

这类任务不仅要求模型对图像中的事实内容进行描述，还需在描述中融入特定的语言风格。主要包含两大研究方向，分别是风格化图像描述[16-21]和情感化图像描述[15]。风格化图像描述假定语言风格是预先设定的，相关方法可大致分为单风格方法与多风格方法。其中，单风格方法[16-17]为每种语言风格训练一个独立的模型，而多风格方法[18-21]则训练一个能够输出多种语言风格的模型。典型的，Guo等人[18]提出一种对抗学习模型MSCap，可同时生成多种风格的描述；Wang等人[20]构建了图像-事实-风格三元语义框架，探索三者的关联关系；Yang等人[21]则利用GAN模型编码风格特征，对具有不同词性的单词进行重组以实现期望情感的表达。对于情感化图像描述，Li等人[15]首次提出通过视觉内容传递潜在情感的图像描述方法，其核心是通过注意力机制融合图像内容信息与情感特征来生成描述，并进一步拓展至视频领域[25]。

值得注意的是，上述方法均依赖于大规模成对的图像-句子数据进行监督训练。为突破这一限制，本文首次提出无监督情感化图像描述生成任务，无需成对图文数据，即可实现利用图像内在感情对其进行描述。

2.2 无监督图像描述生成

目前已有一系列工作[24-28]对无监督图像描述生成展开探索。该任务由Feng等人[24]首次提出，并设计基于GAN的模型对图像和句子进行双向重构，从而学习二者的对应关系。Laina等人[26]将图像映射至学习到的句子流形空间，进而通过预训练的句子解码器从映射后的图像特征中解码出描述。考虑到对抗学习方式的不稳定性，Guo等人[27]提出一种探索图像与句子中共同视觉概念的记忆模型。Qi等人[28]提出一个关系远程监督方法，借助语料库中抽取的物体关系知识连接图像和文本。

与上述方法仅关注视觉语义内容不同，本文方法受启发于人类表达情感的方式，关注如何在图像描述中结合情感元素。为在无可用图像-句子成对数据的情况下实现这一目标，本文从语料库中挖掘情感关系常识知识并构建知识库，从而建立实体与情感化描述之间的关联，指导模型为图像生成合理且通顺的情感化描述。

3 方 法

本节提出结合常识知识的无监督情感化图像描述生成方法。如图2所示，方法整体包含四个关键组件，分别是情感关系常识知识库、事实语句解码器、情感化语句解码器以及视觉信息抽取模块。

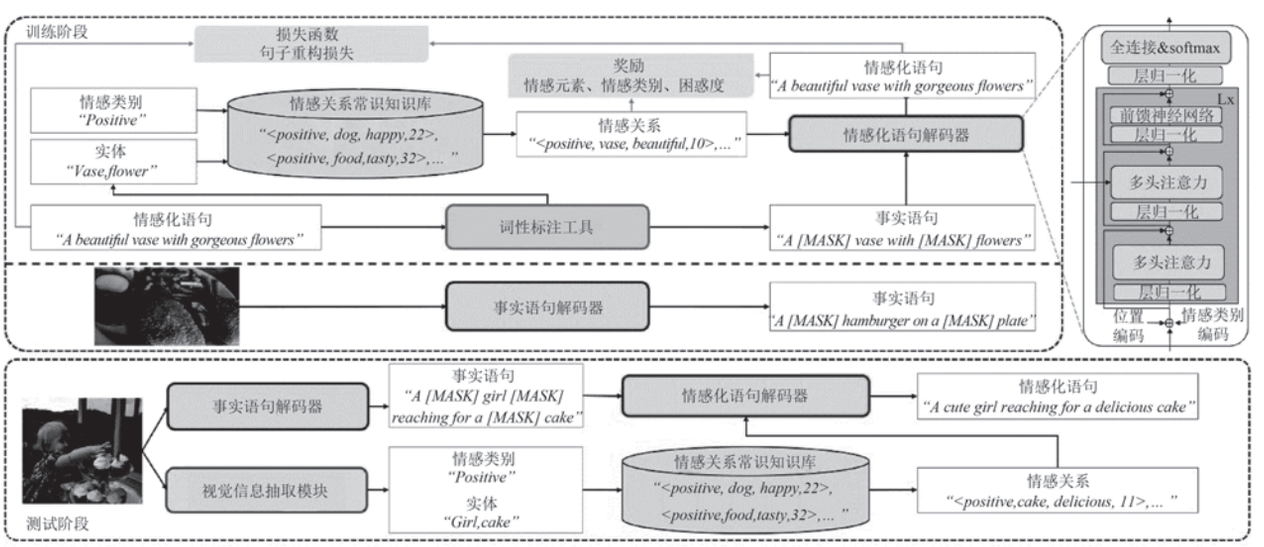

图2 结合常识知识的无监督情感化图像描述生成方法的示意图(图中上方展示的是方法的训练阶段，下方展示的是方法的测试阶段)

在训练阶段，事实语句解码器在不成对的图像和情感部分被掩码的句子上训练，学习如何对图像语义部分进行描述。情感化语句解码器则在情感化语料库上通过重构句子进行训练，学习如何以情感关系常识为依据，在事实描述中融入合理的情感元素。此外，针对情感化语句解码器额外设计的情感元素奖励机制，鼓励其生成与知识库中情感关系相符的情感化元素。在测试阶段，视觉信息抽取模块首先识别输入图像中的实体与情感类别。以视觉信息为查询，可在常识知识库中检索视觉相关的情感关系，并综合事实语句解码器生成的事实描述，作为情感化语句解码器生成情感化描述的依据。

3.1　情感关系常识知识库

受限于无监督条件，难以采用数据驱动的方式学习如何利用合理情感元素描述视觉内容。在此情况下，考虑采用结合知识的方式，通过构建关于情感关系的常识知识库，为情感描述提供指导信息。将COCO[29]与SentiCap[16]中的语料库视作知识载体，挖掘不同情感下实体与描述中情感元素间的共现频率，获取体现情感表达规律的情感关系常识知识。

为构建知识库，首先基于NLTK工具包[30]，通过词性标注识别语料库句子中的形容词和副词，将其视作描述中实现情感表达的情感元素。随后统计各情感元素在不同情感类别句子中出现的频率，将情感元素归类至其最频繁出现的情感类别。具体而言，设C_k表示情感类别为k的句子集合，f_j^k表示情感元素j在C_k中出现的频率。则j的情感类别s可在$f_j^k>\eta_f$的前提下，由$s=\arg\max_{k\in S} f_j^k$确定。其中$S$为情感类别集合，$\eta_f$为预定义的阈值。$f_j^k$的计算过程可表示为

$$f_j^k=\frac{N_j^k}{\sum_v N_j^v\times\left\lfloor\frac{N^m}{N^v}\right\rfloor}\times\left\lfloor\frac{N^m}{N^k}\right\rfloor \tag{1}$$

其中，N_j^k表示情感元素j在C_k中出现的次数，N^k为C_k中的句子总数，N^m表示各情感类别句子数量的最大值，即$N^m=\max_{k\in S} N^k$。

得到情感元素所属情感类别后，计算每一对实体与情感元素在对应情感类别句子集合中的共现次数。对于情感类别s下的情感元素j^s，$N_{j^s}^e$表示实体e与j^s在C_s中的共现次数。经过上述步骤，情感关系最终可表示为形如$<s,e,j^s,N_{j^s}^e>$的四元组，例如$<$*positive*, *cat*, *cuddly*, 20$>$。整个情感关系常识知识库的构建过程高度自动化，无需人工标注操作。经统计，常识知识库中最终共收集到105,102个情感关系。在训练和测试过程中，以样例的情感类别和实体为查询，可从知识库中检索出共现频次最高的前N_r个情感关系，指明合理表达当前样例情感的可能方式。NLTK工具包在词性标注方面性能稳定，且情感关系四元组引入了情感元素与实体共现频次作为后续检索依据，上述步骤能够保证在大规模语料上构建可靠的常识库，为后续模型学习情感元素的合理运用奠定坚实基础。

3.2　事实语句解码器

事实语句解码器在不考虑图像内在情感的情况下，对图像的视觉内容进行描述，而描述中的情感部分则借助特定标识符预留下来。本文将现有的无监督图像描述生成模型UIC[24]用作事实语句解码器。该模型采用对抗学习框架构建，由图像编码器、语句生成器与判别器三大核心模块组成。通过视觉-文本内容双向重构的对抗学习机制，该模型能够生成兼具语法准确性与视觉相关性的描述。为使UIC模型适应无监督情感描述任务，将训练语料库中句子里的情感元素替换为特定标识符，并在其他可能的情感元素位置额外插入特定标识符。

3.3　情感化语句解码器

情感化语句解码器负责生成描述中的情感部分，其以事实语句解码器生成的掩码事实语句、检索到的情感关系、情感类别标签以及位置编码作为输入。如图2右上角所示，情感化语句解码器基于标准Transformer解码器[31]构建，由L层堆叠组成，每层包含两个多头注意力模块和一个前馈神经网络，结合残差连接与层归一化设计，实现情感化语句的解码。此处采用多头注意力模块而非掩码多头注意力模块，是因为情感化语句解码器的输入序列在解码开始时已可知。

将情感类别的词嵌入向量与已生成事实语句词嵌入序列的和表示为$\boldsymbol{E}_f\in\mathbb{R}^{T\times D_e}$，其中$T$表示句子长度，$D_e$表示词向量维度。第一个多头注意力模块的计算过程可表示为

$$\boldsymbol{q}_s=\mathrm{MultiHead}\left(\boldsymbol{E}_f,\boldsymbol{E}_f,\boldsymbol{E}_f\right) \tag{2}$$

为进一步将情感关系常识知识融入描述生成过程，第二个多头注意力模块以检索到的四元组中实体和情感元素的词嵌入作为输入数据，计算过程如下：

$$\boldsymbol{v}_s=\mathrm{MultiHead}\left(\boldsymbol{q}_s,\ \mathrm{fc}\left(\left[\boldsymbol{w}_e;\boldsymbol{w}_s\right]\right),\ \mathrm{fc}\left(\left[\boldsymbol{w}_e;\boldsymbol{w}_s\right]\right)\right) \tag{3}$$

其中，$\mathrm{fc}\left(\left[\boldsymbol{w}_e;\boldsymbol{w}_s\right]\right)$用作键向量和值向量，$\boldsymbol{q}_s\in\mathbb{R}^{T\times D_h}$用作查询向量，$D_h$表示注意力层的参数维度。

fc(·)是一个输入维度为$2D_e$、输出维度为D_h的全连接层。$\boldsymbol{w}_e\in\mathbb{R}^{N_r\times D_e}$和$\boldsymbol{w}_s\in\mathbb{R}^{N_r\times D_e}$分别表示实体和情感元素的词嵌入向量。接下来，前馈网络以融合了情感关系常识知识的v_s作为输入，更新后的特征送入后续的全连接层，并经由softmax生成词表上的概率分布。

3.4 视觉信息抽取模块

视觉信息抽取模块由图像情感分类器和目标检测器组成。图像情感分类器用于识别输入图像的情感类别（即描述应具有的情感类别），其结构包含多个带有RELU激活函数的二维卷积层、用于降维的1×1卷积层，以及一个全连接网络和softmax函数。该分类器以ResNet-101[32]最后一个卷积层的图像网格特征作为输入，当最高概率大于阈值η_c（实验中设为0.7）时，将对应类别作为图像的情感类别标签。目标检测器用于检测输入图像中的实体（即物体标签），本文采用基于OpenImages数据集[33]训练的Faster-RCNN模型[34]。

3.5 训练阶段

事实语句解码器使用不成对的图像和事实语句进行训练，而情感化语句解码器则使用语料库中的情感语句进行训练。视觉信息抽取模块中的图像情感分类器使用图像情感分析数据集进行训练。整个训练过程包含预训练和微调两个阶段。

在预训练阶段，事实语句解码器的训练方式与Feng等人[24]采用的方式相同，唯一的区别在于使用的语料库仅保留句子中的事实部分。情感解码器的训练通过重构情感语句$\bar{y}^s$来实现，其输入包括情感类别标签s、检索到的情感关系R_{sen}以及$\bar{y}^s$情感部分被掩码后的y_m^s。针对情感化解码器输出的句子，采用标准的词级交叉熵损失函数，计算其与掩码前句子$\bar{y}^s$的重构损失：

$$\mathcal{L}_{XE}=-\frac{1}{T}\sum_{t=1}^{T}\log p_\theta\left(\bar{y}_t^s|s,R_{sen},y_m^s\right) \tag{4}$$

与此同时，图像情感分类器基于softmax函数输出的情感类别概率分布，结合图像的真实情感标签计算交叉熵损失，作为优化目标引导模型学习视觉情感分类能力。

在微调阶段，为进一步提升描述质量，引入强化学习并提出一种新的情感元素奖励r_{sen}，鼓励模型使用与常识相符的情感元素来表达情感。r_{sen}通过对比检索出的情感关系和生成的情感化描述二者中的情感元素来计算，其定义如下：

$$r_{sen}=\begin{cases}1, & y^s\cap J_r\neq\varnothing\\ -1, & y^s\cap J_r=\varnothing\end{cases} \tag{5}$$

其中，y^s表示基于概率采样生成的句子，J_r表示从检索出的情感关系中提取的情感元素。当句子包含情感关系提议的情感元素时，该奖励设为1，反之为-1。

除此之外，微调阶段还采用了情感类别奖励r_{cls}和困惑度奖励r_{ppl}来进一步提高句子的准确性和流畅性，其定义分别为

$$r_{cls}=\mathbb{I}_{s_g=s} \tag{6}$$

$$r_{ppl}=\begin{cases}-1, & \mathrm{ppl}(\hat{y})-\mathrm{ppl}(y^s)<0\\ 0, & \mathrm{ppl}(\hat{y})-\mathrm{ppl}(y^s)=0\\ 3, & \mathrm{ppl}(\hat{y})-\mathrm{ppl}(y^s)>0\end{cases} \tag{7}$$

其中，$\mathbb{I}_{s_g=s}$在生成句子的情感类别s_g与原始句子的情感类别s一致时取值为1，否则为-1. $\mathrm{ppl}(y)$表示句子y的困惑度，由SRILM工具包[35]计算得到。y^s和$\hat{y}$分别表示采样生成的句子和贪婪解码生成的句子。总奖励函数定义为

$$r(y^s,\hat{y})=r_{ppl}+\lambda_{r_{cls}}r_{cls}+\lambda_{r_{sen}}r_{sen} \tag{8}$$

其中，$\lambda_{r_{cls}}$和$\lambda_{r_{sen}}$是用于平衡奖励的超参数。微调阶段中强化学习的梯度通过以下方式近似计算：

$$\nabla_\theta\mathcal{J}(\theta)\approx-r(y^s,\hat{y})\nabla_\theta\log p_\theta(y^s) \tag{9}$$

3.6 测试阶段

在测试阶段，给定输入图像后，视觉信息抽取模块首先识别图像的情感类别和实体信息，同时由事实语句解码器生成情感部分被掩码的事实描述。接着，综合情感类别和实体两部分内容作为查询，从常识知识库中检索出二者共现频次最高的前N_r个情感关系。最后，情感化语句解码器以事实语句为输入，结合上述基于视觉信息检索出的情感关系，输出图像的情感化描述。

3.6.1 情感化图像描述评价指标

为衡量情感化图像描述的质量，现有指标大致分为两类，分别从事实准确性和情感契合度两个角度评估。然而，这种分离式的评估方式忽视了内容与情感的协同关系，存在明显缺陷。例如，当模型将一张精彩的棒球比赛图像，描述为语义无关的“An adorable cat is sleeping”时，其情感相关分数（如情感类别准确率和困惑度分数）可能仍处于较高水平，因为该语句通顺且与图像一样都在传达积极情感；另一方面，当模型完全忽略情感信息，生成“A baseball player is

swinging the bat”时，内容相关性指标（如CIDEr）也可能仍然表现良好。这导致现有评估体系难以准确反映情感化图像描述的真实质量，而多个指标在量纲和评分范围上的差异，降低评估结果的可比性。

为解决这一问题，本节提出一种新指标SentiCLIPScore，能够同时反映情感化图像描述的事实准确性和情感契合度。该指标采用无参考评估方式，无需情感化描述真值作为参考，减轻了标注真值的负担。具体而言，事实准确性通过预训练模型CLIP[36]衡量，而情感契合度则综合视觉信息、句子情感内容以及情感关系知识库进行评估。SentiCLIPScore的计算公式如下：

$$\text{Sen-CLIP-S}=\left(\text{CLS}(y,s)\times\text{CLIP-S}(\boldsymbol{c},\boldsymbol{v})\right)^{\frac{1}{\gamma_{rel}}} \tag{10}$$

其中，s表示测试图像的情感类别，$\text{CLS}(y,s)$表示由句子情感分类器给出的生成句子y属于情感类别s的概率，γ_{rel}为情感因子。$\text{CLIP-S}(\boldsymbol{c},\boldsymbol{v})$表示计算出的CLIPScore[37]，$\boldsymbol{c}$和$\boldsymbol{v}$分别表示待评估描述的文本嵌入向量和对应图像的视觉嵌入向量。

具体的，$\text{CLIP-S}(\boldsymbol{c},\boldsymbol{v})$计算过程为

$$\text{CLIP-S}(\boldsymbol{c},\boldsymbol{v})=w\times\max\left(\cos(\boldsymbol{c},\boldsymbol{v}),0\right) \tag{11}$$

其中，$\cos(\cdot)$计算$\boldsymbol{c}$和$\boldsymbol{v}$之间的余弦相似度。w为超参数，设置为2.5。情感因子γ_{rel}综合考虑视觉内容和情感关系常识知识，衡量描述中情感部分的合理性，计算过程为

$$\gamma_{rel}=1+\frac{\hat{N}_{rel}^{y}}{N_{rel}^{y}+\epsilon} \tag{12}$$

其中，N_{rel}^{y}表示由依存句法分析器[38]从句子y中提取的情感元素总数。根据测试图像的视觉内容，可从常识知识库中检索出相关的情感关系，而$\hat{N}_{rel}^{y}$则表示句子y中符合常识知识检索内容的情感元素数量。ϵ是为了避免出现除数为零的常数，实验中设置为1×10^{-4}。

综上所述，SentiCLIPScore的评分范围由0至1，分数越高表示情感化描述质量越好，分数降低则表明事实准确性或情感契合度存在不足。

3.7　情感关系常识知识库的局限性与扩展性

尽管情感关系常识知识库建立了实体与情感化描述之间的关联，能够为情感化图像描述生成提供指导信息，但其覆盖范围和质量一定程度上取决于构建时所用的语料库（如COCO，SentiCap）。这使得目前版本的知识库对于仅出现在某些特定领域的低频实体或特殊命名实体上可能存在常识缺失或情感偏差的问题。针对上述问题，可以进一步基于上位词（hypernym）关系进行层级化知识检索，当知识库中缺乏某些特定命名实体的情感信息时，模型可以回溯到其上位类别（如“哈士奇”相关知识缺失时可回溯到“狗”），以提升知识推理的覆盖率。此外，还可以结合本文提出的自动化知识库构建过程，引入更大规模和多样化的语料来源（如书籍、社交媒体文本等），以动态扩展知识库并缓解领域偏差，增强方法的鲁棒性。

4　实　验

4.1　数据集

为评估本文方法有效性，在COCO数据集[29]与Flickr30K数据集[39]的图像上进行实验。其中，COCO共包含113287张图像用于训练，5000张图像用于验证，5000张图像用于测试，每张图像都标注了五句描述。COCO训练集中的描述被视作中性语句，与SentiCap[16]中的4892句积极语句和3977句消极语句一起，构成情感语料库。Flickr30K共包含31783张图像，其中29783张用于训练，1000张用于验证，1000张用于测试。图像情感分类器在EmotionROI[40]、ArtPhoto[41]、TwitterI[42]和TwitterII[43]四个图像情感分析数据集上训练，共包含1685张体现积极情感的图像、2308张体现消极情感的图像以及655张无明显情感的中性图像用于训练。

4.2　实验设置

本文方法中的情感化语句解码器由四层相同的多头注意力模块构成，每层具有八个注意力头。词向量维度D_e与注意力层的参数维度D_h均设置为512，前馈网络的隐藏层维度设置为2048。经过在验证集上以SentiCLIPScore为依据、范围为$\{0.5,1.0,2.0,4.0\}$的网格搜索，公式(8)中的超参数$\lambda_{r_{cls}}$和$\lambda_{r_{sen}}$分别被设置为4.0和2.0。用于共现频率过滤的超参数η_f被设定为0.6，而情感关系数量N_r则被固定为30。训练过程采用Adam优化器[44]，预训练阶段将学习率设定为4×10^{-4}，微调阶段将学习率设定为4×10^{-5}。

针对方法生成的情感化图像描述，采用本文提出的SentiCLIPScore指标（缩写为S-CLIP）同时对其与图像的事实准确性和情感契合度两个角度进行

评测，分数越高表明描述质量越高。此外，也利用常用的几种指标进行了评测，包括面向事实准确性的BLEU[45]、METEOR[46]和CIDEr[47]，以及面向情感契合度的情感分类准确率（缩写为cls）和平均困惑度（缩写为ppl）。其中，情感分类准确率表示生成的描述与图像具有相同情感的百分比，由LSTM和一个全连接层构成的句子情感分类器计算，其在情感语料库上训练后具有99%的分类准确率。平均困惑度的计算由SRILM工具包[35]完成，与公式（7）中计算r_{ppl}的方式一致。困惑度越低表明生成的描述越通顺，且越能体现期望的情感。

4.3 方法对比

由于没有其他无监督情感化图像描述方法用于对比，将经典的无监督图像描述方法UIC采用两种不同方式适应至当前任务，形成两个无监督基线方法：（1）UIC*：利用情感语料库和图像数据训练一个可同时应对三种情感的模型。（2）UIC†：为每种情感利用对应的语料训练一个独立的模型。在Flickr30K与COCO的图像上的结果如表1所示。值得注意的是，表中最后的“所有情感”指的是将方法在积极、消极与中性三种情感上的所有输出内容视为一个整体，综合每个样例与其标注真值或视觉内容计算的评价指标而得出的结果，体现方法在三种情感上的综合性能。可以看到，本文方法在两个数据集的所有情感类别上都达到了最高的S-CLIP分数，这表明其无需成对图像-句子数据即可生成高质量情感化描述。与此同时，对照其他两类指标，本文提出的S-CLIP分数也可有效反映出描述在内容和情感方法上的缺陷。此外，实验结果进一步揭示了平均困惑度（ppl）在评估情感化描述性能时的局限性。具体而言，从UIC† 的CIDEr分数可以看出，该基线方法忽略了图像与文本之间的跨模态对应关系，导致生成的描述在事实准确性方面表现较差，而其ppl指标却处于最优水平。类似地，UIC*的情感准确率（cls）表明其在生成反映图像情感的描述方面存在不足，但其ppl分数同样表现优异。上述结果表明，在描述事实性与情感性严重缺失的两种极端情形下，ppl均无法有效区分方法性能以及真实反映情感的描述质量，因而缺乏可比性。这一观察也正是本文提出并采用SentiCLIPScore评价指标的动机所在。

表1 在Flickr30K数据集以及COCO数据集上与无监督基线方法的对比结果

情感	方法	Flickr30K数据集							COCO数据集						
		Bleu-1	Bleu-3	METEOR	CIDEr	ppl(↓)	cls(%)	S-CLIP	Bleu-1	Bleu-3	METEOR	CIDEr	ppl(↓)	cls(%)	S-CLIP
积极	UIC*	41.2	10.0	9.5	10.7	16.2	8.1	0.092	54.1	22.1	15.7	37.8	12.8	1.2	0.020
	UIC†	38.5	7.8	8.7	5.1	8.0	100.0	0.459	35.2	6.6	9.2	5.0	8.6	100.0	0.407
	Ours	41.2	10.6	10.6	13.1	22.0	93.3	**0.545**	54.9	21.5	17.5	37.6	16.5	90.4	**0.598**
消极	UIC*	41.9	11.2	10.2	9.6	15.6	6.2	0.066	55.0	23.1	16.7	45.3	13.5	6.3	0.080
	UIC†	40.7	11.3	10.0	3.0	10.7	100.0	0.547	33.7	2.2	8.4	3.2	6.7	100.0	0.429
	Ours	42.7	11.3	10.8	10.6	19.6	82.2	**0.552**	56.8	23.4	18.4	45.8	16.3	84.1	**0.568**
中性	UIC*	41.7	9.9	10.2	9.7	13.8	84.4	0.367	54.6	21.8	16.4	43.3	9.0	93.6	0.524
	UIC†	46.3	12.6	11.1	13.2	9.3	98.9	**0.507**	60.5	27.6	19.2	60.0	7.8	99.7	0.646
	Ours	44.9	11.5	11.2	12.7	11.7	98.7	0.499	61.0	26.9	18.8	59.4	8.9	99.6	**0.653**
所有情感	UIC*	41.6	10.2	10.0	9.5	10.4	48.6	0.234	54.6	22.1	16.3	43.1	8.0	57.6	0.334
	UIC†	43.4	11.1	10.2	8.7	11.3	99.4	0.503	49.7	18.4	14.8	37.5	11.7	99.8	0.555
	Ours	43.5	11.3	11.0	11.9	16.3	93.9	**0.522**	59.1	25.1	18.5	52.3	11.8	94.7	**0.625**

注：“*”和“†”表示不同训练方式下的基线方法，SentiCLIPScore缩写为S-CLIP，得分最高的加粗标明。

此外，如表2中所示，与全监督方法相比，包括情感化图像描述方法Insenti-Cap[15]和风格化图像描述方法StyleNet[17]、MemCap[19]、MPDCap[48]以及TridentCap[20]，本文方法在没有利用图像-句子监督信息的情况下，依旧在COCO上获得了与全监督方法相当的情感准确率（cls）以及平均困惑度（ppl）。该结果点明了探索情感关系常识知识对于合理表达情感的作用。而监督信息的缺失也的确会导致本文方法在事实准确性指标上落后于全监督方法。

考虑到预训练视觉语言模型的飞速发展，针对经典的BLIP2[49]、LLaVA[50]以及MiniGPT4[51]，以“Describe this image with the inherent sentiment reflected by the image”为提示词，实现零样本的情感化图像描述生成，并将结果与本文方法进行对比。COCO数据集上的结果如表3所示。从结果可以看

表2　在COCO数据集上与情感化描述方法(Insenti-Cap)和风格化图像描述方法(StyleNet, MemCap, MPDCap, TridentCap)的对比结果

情感	方法	监督训练	方法类别	Bleu-1	Bleu-3	METEOR	CIDEr	ppl(↓)	cls(%)	S-CLIP
积极	StyleNet[17]	是	风格化	45.3	12.1	12.1	36.3	24.8	45.2	-
	MemCap[19]	是	风格化	51.1	17.0	16.6	52.8	18.1	96.1	-
	MPDCap[48]	是	风格化	52.3	18.2	17.0	54.8	13.2	99.3	-
	TridentCap[20]	是	风格化	55.1	23.5	18.7	69.0	14.9	100.0	-
	Insenti-Cap[15]	是	情感化	59.7	25.3	20.9	61.3	13.0	98.5	0.725
	本文	否	情感化	54.9	21.5	17.5	37.6	16.5	90.4	0.598
消极	StyleNet[17]	是	风格化	43.7	10.6	10.9	36.6	25.0	56.6	-
	MemCap[19]	是	风格化	49.2	18.1	15.7	59.4	18.9	98.9	-
	MPDCap[48]	是	风格化	49.3	18.4	16.3	55.0	13.0	96.5	-
	TridentCap[20]	是	风格化	55.7	24.5	18.9	71.3	13.6	100.0	-
	Insenti-Cap[15]	是	情感化	59.1	24.3	19.4	53.3	12.3	95.5	0.632
	本文	否	情感化	56.8	23.4	18.4	45.8	16.3	84.1	0.568
中性	Insenti-Cap[15]	是	情感化	73.5	41.2	24.7	97.5	8.4	98.9	0.690
	本文	否	情感化	61.0	26.9	18.8	59.4	8.9	99.6	0.653
所有情感	Insenti-Cap[15]	是	情感化	69.0	34.7	22.9	82.5	9.2	97.5	0.686
	本文	否	情感化	59.1	25.1	18.5	52.3	11.8	94.7	0.625

出，本文方法在S-CLIP分数上明显高于预训练视觉语言模型。而所有的预训练视觉语言模型都在传达了积极或消极情感的图像上，获得了极低的情感准确率。上述对比结果表明，对于预训练视觉语言模型而言，生成能准确反映图像情感的描述依然极具挑战，这也进一步体现出本文方法的价值。

表3　在COCO数据集上与预训练视觉语言模型的对比结果

情感	方法	Bleu-1	Bleu-3	METEOR	CIDEr	ppl(↓)	cls(%)	S-CLIP
积极	BLIP2[49]	65.8	40.0	25.7	101.6	34.6	0.5	0.077
	LLaVA[50]	17.0	7.4	17.8	0.0	155.9	8.4	0.083
	MiniGPT4[51]	17.7	6.9	14.0	0.4	134.5	2.8	0.069
	本文	54.9	21.5	17.5	37.6	**16.5**	**90.4**	**0.598**
消极	BLIP2[49]	60.0	36.1	23.6	98.5	32.5	1.3	0.079
	LLaVA[50]	16.7	6.9	17.0	0.0	167.7	2.2	0.083
	MiniGPT4[51]	18.3	6.9	14.3	0.2	145.6	2.4	0.070
	本文	56.8	23.4	18.4	45.8	**16.3**	**84.1**	**0.568**
中性	BLIP2[49]	62.5	37.3	24.3	99.6	14.0	99.1	0.078
	LLaVA[50]	16.7	7.1	17.2	0.0	246.2	91.0	0.083
	MiniGPT4[51]	18.3	7.1	14.3	0.3	126.1	94.1	0.070
	本文	61.0	26.9	18.8	59.4	**8.9**	**99.6**	**0.653**
所有情感	BLIP2[49]	62.8	37.6	24.4	99.7	14.3	59.8	0.078
	LLaVA[50]	16.7	7.1	17.3	0.0	240.5	56.7	0.083
	MiniGPT4[51]	18.2	7.0	14.2	0.3	130.4	57.5	0.070
	本文	59.1	25.1	18.5	52.3	**11.8**	**94.7**	**0.625**

4.4　消融实验

针对本文方法中的情感关系常识知识以及情感元素奖励，在COCO数据集上进行消融实验来验证其有效性，结果如表4所示，其中“K”代表情感关系常识，“R”代表情感元素奖励。值得注意的是，由于情感元素奖励以情感关系常识为基础，情感关系常识的移除也会导致该奖励无法实现。对比表4中各情感的第一行与第二行结果可以看出，在移除情感关系常识的情况下，本文方法在各项指标上的结果均有所下降，体现出引入知识对生成高质量情感化描述的重要性。对比表4中各情感的第二行与第三行结果可以看出，

移除情感元素奖励会导致情感相关的指标性能大幅下降，例如cls、ppl以及S-CLIP，验证了该奖励在确保描述情感契合且表述通顺上的重要作用。除自动化评价指标外，还招募十位志愿者人工评价这些方法生成的描述质量如何。人工评价对描述与图像的事实准确性和情感契合度进行打分。情感契合度分数范围为1分至3分，分别代表描述所传达的情感与图像的完全不符以及高度一致。相似的，事实准确性分数范围为1分至3分，分别代表描述的内容与图像完全无关以及完全一致。人工测评结果见图3，纵轴表示方法得到不同分数的占比，横轴括号内是方法的平均分数。本文方法同时在内容和情感两个方法都收获了最高分，而移除情感关系常识知识或情感元素奖励中的任何一项，都会导致分数下降，体现出情感关系常识知识以及情感元素奖励的有效性。

表4 COCO数据集上情感关系常识知识(K)与情感元素奖励(R)的消融实验结果

情感	K	R	Bleu-1	Bleu-3	METEOR	CIDEr	ppl(↓)	cls(%)	S-CLIP
积极	×	×	52.6	20.5	17.3	35.1	20.7	88.0	0.574
	√	×	53.9	20.9	17.4	36.9	19.1	88.5	0.588
	√	√	**54.9**	**21.5**	**17.5**	**37.6**	**16.5**	**90.4**	**0.598**
消极	×	×	56.1	23.0	18.1	44.6	17.0	79.6	0.533
	√	×	56.1	23.1	18.3	44.8	17.5	78.9	0.533
	√	√	**56.8**	**23.4**	**18.4**	**45.8**	**16.3**	**84.1**	**0.568**
中性	×	×	59.9	26.2	**18.8**	56.9	11.6	99.2	0.650
	√	×	60.3	26.5	**18.8**	58.4	10.0	**99.6**	0.653
	√	√	**61.0**	**26.9**	**18.8**	**59.4**	**8.9**	**99.6**	**0.653**
所有情感	×	×	57.8	24.4	18.3	49.9	15.0	93.1	0.611
	√	×	58.4	24.7	18.4	51.4	13.4	93.3	0.616
	√	√	**59.1**	**25.1**	**18.5**	**52.3**	**11.8**	**94.7**	**0.625**

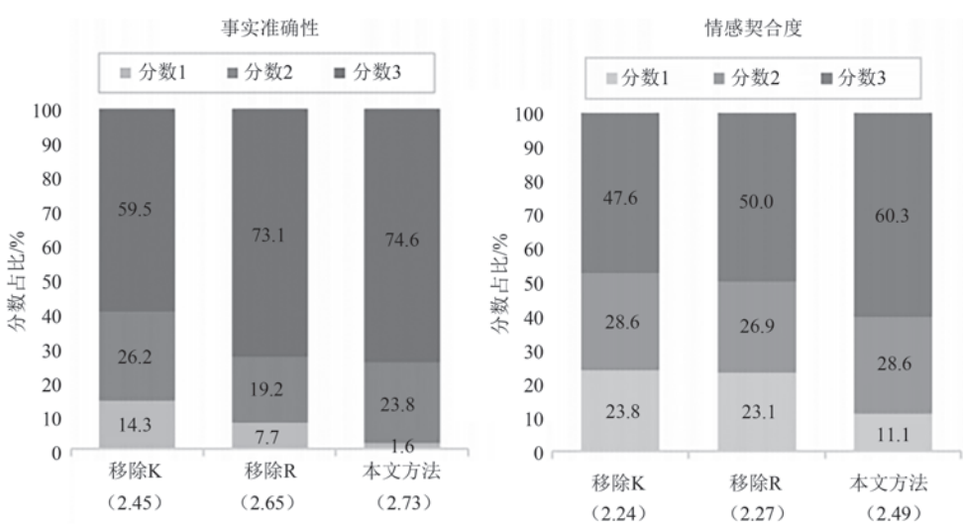

图3 消融实验的人工评测结果。K代表情感关系常识，R代表情感元素奖励

4.5 可视化分析

图4展示了无监督基线方法与本文方法生成的情感化图像描述。相比之下，本文方法生成的描述质量更高，能够利用“gorgeous flower”、“lonely street”准确反映图像内容。此外，“UIC†”只是一味地生成与图像情感相符的情感词“nice”、“dead”，而整体句子内容是与图像内容完全无关的。图5展示了移除情感关系常识方法以及情感元素奖励方法和本文方法的生成结果。相较于不引入情感关系常识，本文方法能够利用与图像相关且合理的情感元素，例如“pretty woman”和“rotten banana”，来传达图像情感。相较于不使用情感元素奖励，本文方法能够生成更贴近视觉实体且更生动的描述，例如生成了“cuddly cat”而不是宽泛的“nice cat”。

积极情感图像	事实/情感化描述	消极情感图像	事实/情感化描述
	COCO： A close up of a vase with many flowers. UIC*： A vase filled with flower on a table. 本文方法：Gorgeous flower in a vase on a table.		COCO： A woman walking down the road with a pink umbrella. UIC*： A woman holding an umbrella in the rain. 本文方法：A person with an umbrella walking down the lonely street.
	COCO：Many different dishes of food on a table. UICt†： A group of people are traveling down a nice street. 本文方法：A table topped with plate of awesome food and a drink.		COCO： A large jetliner flying over a small farm near a forest. UICt†： A dead woman standing in a field with her horse. 本文方法： An airplane flying in the gloomy sky above a tree.

图4　本文方法与无监督基线方法“UIC*”、“UIC †”在COCO数据集上的情感化描述结果可视化（其中“COCO”代表COCO数据集中图像对应的事实描述）

积极情感图像	事实/情感化描述	消极情感图像	事实/情感化描述
	COCO：A woman hitting a ball with a tennis racquet. 移除情感关系常识：A woman is swinging a tennis racket at a ball. 本文方法：A pretty woman is swinging a tennis racket at a ball.		COCO： A basket of bananas and some apples on a table. 移除情感关系常识：A bunch of banana sitting on a table. 本文方法： A bunch of rotten banana sitting on a table.
	COCO：A black and white cat laying on a desk by a laptop. 移除情感元素奖励：A nice cat laying on a desk next to a laptop. 本文方法：A cuddly cat laying on a desk next to a laptop.		COCO： A blue boat docked next to a table full of people. 移除情感元素奖励：A boat in a body of shallow water next to the building. 本文方法：A boat in a body of dirty water next to the building.

图5　本文方法与移除情感化常识、情感元素奖励的方法在COCO数据集上的情感化描述结果可视化（其中“COCO”代表COCO数据集中图像对应的事实描述）

5　结论与展望

本文提出了结合常识知识的无监督情感化图像描述生成方法。通过挖掘外部语料中的情感关系常识构成知识库，并检索出与图像视觉相关的情感元素用于描述的参考和提供强化学习奖励，该方法能够在不依赖图像-句子成对数据的情况下，利用恰当的情感元素表达图像潜在情感，进而生成与图像内容相关且情感契合的描述。与此同时，本文提出的SentiCLIPScore指标，能够同时反映出情感化图像描述与图像的事实准确性与情感契合度，降低该任务原先需在多个指标上综合评估的难度。在COCO和Flickr30K数据集上的大量对比实验与消融实验验证了本文方法的有效性。

本文的研究集中于积极、消极和中性三类宏观情感。然而，人类情感表达往往更加细腻，例如喜悦、悲伤、愤怒与惊奇等。未来工作可对现有方法进行扩展，尝试通过更精细地划分语料中的情感类别，构建细粒度情感关系常识知识库，以提升生成描述的情感多样性与表现力。然而，这一研究也面临着诸多挑战，包括细粒度情感语料的获取难度、不同情感类别之间的语义模糊性，以及更复杂的知识库表示与推理开销。此外，探索场景级的情感关系常识收集，以及利用更丰富的视觉信息检索常识中的情感元素，同样是未来重要的研究方向。

参　考　文　献

[1] Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory

transformer for image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10578-10587

[2] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6077-6086

[3] Yang X, Liu Y, Wang X. Reformer: The relational transformer for image captioning//Proceedings of the ACM International Conference on Multimedia, Lisboa, Portugal, 2022:5398-5406

[4] Wang Y, Xu J, Sun Y. End-to-end transformer based model for image captioning//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022: 2585-2594

[5] Fu Z, Song K, Zhou L, et al. Noise-aware image captioning with progressively exploring mismatched words//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 12091-12099

[6] Li J, Mao Z, Li H, et al. Exploring visual relationships via transformer-based graphs for enhanced image captioning. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(5): 1-23

[7] Park S, Kim Y. Building thesaurus lexicon using dictionary-based approach for sentiment classification//Proceedings of the IEEE International Conference on Software Engineering Research, Management and Applications. Towson, USA, 2016: 39-44

[8] Al Amrani Y, Lazaar M, El Kadiri K E. Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Computer Science, 2018, 127: 511-520

[9] Rao G, Huang W, Feng Z, et al. Lstm with sentence representations for document-level sentiment classification. Neurocomputing, 2018, 308: 49-57

[10] Basiri M E, Nemati S, Abdar M, et al. Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. Future Generation Computer Systems, 2021, 115: 279-294

[11] Siersdorfer S, Minack E, Deng F, et al. Analyzing and predicting sentiment of images on the social web//Proceedings of the ACM International Conference on Multimedia. Firenze, Italy, 2010: 715-718

[12] Song K, Yao T, Ling Q, et al. Boosting image sentiment analysis with visual attention. Neurocomputing, 2018, 312: 218-228

[13] Yang J, She D, Sun M, et al. Visual sentiment prediction based on automatic discovery of affective regions. IEEE Transactions on Multimedia, 2018, 20(9): 2513-2525

[14] Lee S, Ryu C, Park E. Osanet: Object semantic attention network for visual sentiment analysis. IEEE Transactions on Multimedia, 2022, 25: 7139-7148

[15] Li T, Hu Y, Wu X. Image captioning with inherent sentiment//Proceedings of the IEEE International Conference on Multimedia and Expo. Shenzhen, China, 2021: 1-6

[16] Athews A, Xie L, He X. Senticap: Generating image descriptions with sentiments//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 3574-3580

[17] Gan C, Gan Z, He X, et al. Stylenet: Generating attractive visual captions with styles//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA. 2017: 3137-3146

[18] Guo L, Liu J, Yao P, et al. Mscap: Multi-style image captioning with unpaired stylized text//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4204-4213

[19] Zhao W, Wu X, Zhang X. Memcap: Memorizing style knowledge for image captioning//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 12984-12992

[20] Wang L, Qiu H, Qiu B, et al. Tridentcap: Image-fact-style trident semantic framework for stylized image captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(5): 3563-3575

[21] Yang C, Wang Y, Han L, et al. Fine-grained image emotion captioning based on generative adversarial networks. Multimedia Tools and Applications, 2024, 83(34): 81857-81875

[22] Niu T, Zhu S, Pang L, et al. Sentiment analysis on multi-view social data//Proceedings of the International Conference on Multimedia Modeling. Miami, USA, 2016: 15-27

[23] Yang X, Feng S, Wang D, et al. Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia, 2020, 23: 4014-4026

[24] Feng Y, Ma L, Liu W, et al. Unsupervised image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4125-4134

[25] Wu X, Li T. Sentimental visual captioning using multimodal transformer. International Journal of Computer Vision, 2023, 131(4): 1073-1090

[26] Laina I, Rupprecht C, Navab N. Towards unsupervised image captioning with shared multimodal embeddings//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 7414-7424

[27] Guo D, Wang Y, Song P, et al. Recurrent relational memory network for unsupervised image captioning//Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence. Yokohama, Japan, 2020: 920-926

[28] Qi Y, Zhao W, Wu X. Relational distant supervision for image captioning without image-text pairs//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 4524-4532

[29] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[30] Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with The Natural Language Toolkit. O'Reilly Media, Inc., 2009

[31] Aswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[32] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[33] Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision, 2020, 128(7): 1956-1981

[34] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[35] Stolcke A. Srilm-an extensible language modeling toolkit// Proceedings of the International Conference on Spoken Language Processing. Denver, USA, 2002: 901-904

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

[37] Hessel J, Holtzman A, Forbes M, et al. Clipscore: A reference-free evaluation metric for image captioning// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 7514-7528

[38] Manning C D, Surdeanu M, Bauer J, et al. The stanford corenlp natural language processing toolkit//Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, USA, 2014: 55-60

[39] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2641-2649

[40] Peng K C, Sadovnik A, Gallagher A, et al. Where do emotions come from? Predicting the emotion stimuli map//Proceedings of the IEEE International Conference on Image Processing. Phoenix, USA, 2016: 614-618

[41] Machajdik J, Hanbury A. Affective image classification using features inspired by psychology and art theory//Proceedings of the ACM International Conference on Multimedia. Firenze, Italy, 2010: 83-92

[42] You Q, Luo J, Jin H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 381-388

[43] Borth D, Ji R, Chen T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs//Proceedings of the ACM International Conference on Multimedia. Barcelona, Spain, 2013: 223-232

[44] Kingma D P. Adam: A method for stochastic optimization// Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015:1-15

[45] Papineni K, Roukos S, Ward T, et al. Bleu: A method for automatic evaluation of machine translation//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[46] Banerjee S, Lavie A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments// Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, USA, 2005: 65-72

[47] Edantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4566-4575

[48] Wu X, Zhao W, Luo J. Learning cooperative neural modules for stylized image captioning. International Journal of Computer Vision, 2022, 130(9): 2305-2320

[49] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models// Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 19730-19742

[50] Liu H, Li C, Wu Q, et al. Visual instruction tuning// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023: 34892-34916

[51] Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models// Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024: 1-17

QI Ya-Yun, Ph. D. candidate. Her research interests include computer vision and multimodal content analysis.

ZHAO Wen-Tian, Ph. D., assistant professor. His research interests include multimodal content analysis, artificial intelligence security, and educational technology.

LI Tong, M. S. His research interests include computer vision and video content analysis.

WU Xin-Xiao, Ph. D., professor, Ph. D. supervisor. Her research interests include vision and language, machine learning, and video understanding.

Background

The research presented in this paper belongs to the field of vision-language understanding and generation, which lies at the intersection of computer vision and natural language processing. Within this field, image captioning has been a long-standing core problem, aiming to generate natural language descriptions that accurately reflect visual content. Over the past decade, this task has evolved from describing factual semantics to capturing higher-level attributes such as style and sentiment.

In particular, the task of sentimental image captioning represents an important step toward sentiment-aware artificial intelligence. It requires models to generate textual descriptions of an image with the underlying emotion expressed by the image itself, enabling more human-like communication and interaction. However, the reliance on large-scale annotated image-sentence pairs limits its scalability and adaptability to diverse real-world scenarios.

In this paper, we make the first attempt on unsupervised sentimental image captioning, which aims to generate image descriptions using inherent sentiments without any image-sentence pairs for training. The available data for this task is an image set and an external sentimental corpus, where the image set and the sentimental corpus have no correlation. This paradigm not only reduces dependence on costly human labeling but also provides a pathway toward more autonomous and scalable sentiment-aware systems.

Our method integrates external commonsense knowledge of sentimental relationships into the caption generation process, enabling the inference of appropriate emotional expressions aligned with visual content. To acquire essential knowledge, we investigate on constructing a commonsense knowledge base of sentimental relationships, through mining the co-occurrence frequencies between entities and sentimental words with different sentiments from the corpus. In total, we collect 105, 102 sentimental relationships to build the knowledge base, where each sentimental relationship is represented by a quadruple. With the support of the acquired knowledge, our unsupervised sentimental image captioning method adopts a two-phase generation strategy. It first produces factual captions with masked sentimental components and then fills these masked parts with sentimentally relevant words derived from the constructed knowledge base that links entities to emotional expressions. Furthermore, we design a sentimental reward within a reinforcement learning framework to guide the model toward generating captions sentimentally aligned with commonsense knowledge. To more accurately evaluate this task, we also propose a novel metric, SentiCLIPScore, which jointly measures the factual accuracy and sentimental expressiveness of captions. Extensive experimental results on COCO and Flickr30k demonstrate that our approach significantly improves sentimental relevance and descriptive quality over unsupervised baselines. Notably, it also surpasses strong VLMs pre-trained on extensive paired image-caption data, even without utilizing any paired image-caption data. These results confirm the effectiveness of our method in bridging factual and emotional understanding in visual captioning, offering a new step toward sentiment-aware visual-language intelligence.

This work was supported in part by the grants from the Shenzhen Science and Technology Program under Grant No. JCYJ20241202130548062, the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, the Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No. 2023ZDZX1034, and the National Natural Science Foundation of China under Grant No. 62072041.