基于检索增强和模态融合机制的视觉情感分析

程 立[1] 张 虹[1]
CHENG Li ZHANG Hong

摘 要 现有视觉情感分析方法通常直接将视觉特征输入单一视觉模型，使图像与情感之间存在巨大情感鸿沟，为提升视觉情感分析的效果，文章提出了检索增强视觉情感分析模型（multimodal information retrieval-augmented, MIRA），可以有效连接图像与情感。通过图像特征检索相关文本，将检索到的文本特征与图像特征融合表示以进行情感分析。在多个数据集上的实验结果表明，MIRA 模型在情感分类任务中表现优异，尤其在处理模糊或复杂情感图像时显著提高了分类准确性。

关键词 图像情感分析；CLIP；多模态融合；检索增强；交叉注意力

doi：10.3969/j.issn.1672-9528.2025.04.016

0 引言

视觉情感分析（visual aentiment analysis）是计算机视觉领域中的一个重要任务，旨在通过分析图像内容推断其所表达的情感。随着社交媒体和多媒体数据的爆炸式增长，图像情感识别在诸如社交平台用户情感分析、广告推荐和舆情监测等应用场景中具有广泛的应用潜力。然而，情感通常不仅限于图像本身，还依赖于图像内容复杂信息的结合。例如，一张图像所表达的情感可能受其相关的文字描述、背景信息以及其他辅助信息的影响。因此，依靠图像单模态的情感识别方法无法充分捕捉图像背后的复杂情感。

为解决现有视觉情感分析模型在情感复杂或模糊情况下表现不佳的问题，本文提出了一种检索增强的视觉情感分析模 型（multimodal information retrieval-augmented, MIRA）。该模型利用检索机制，从情感文本数据库中获取与图像相关的情感描述，从而在不增加模型训练复杂度的情况下，显著提升了情感分类的准确性。

具体来说，本文首先构建了一个包含 560 000 条文本情感描述的数据集，然后使用 CLIP 模型 [1] 将图像编码为向量表示，并利用 FAISS[2] 建立情感文本向量库索引以进行相似性检索，通过图像特征从构建的数据库中检索出与图像最相似的前 k 条文本描述。随后，图像特征通过预训练的 ResNet50[3] 模型提取，检索到的文本特征通过 BERT 模型 [4] 提取。为了更好地融合图像和文本特征，设计了交叉注意力机制，将图像特征和文本特征进行交互式融合以进行后续分类。

1. 太原师范学院 山西晋中 030012

1 文献综述

1.1 视觉情感分析

图像情感分析（visual sentiment analysis, VSA）最初主要依赖手工设计的低级视觉特征。Machajdik 等人 [5] 提取了颜色、纹理和构图等受心理学启发的特征，用于情感分类。这些方法虽然展示了情感分析的潜力，但由于仅能捕捉图像的浅层特征，难以有效应对复杂的情感表达。为此，Borth 等人 [6] 提出了视觉情感本体（VSO）和 SentiBank，使用形容词－名词对（ANP）连接视觉内容和情感概念，提升了模型的语义理解能力。然而，该方法依赖预定义的情感词汇，灵活性有限，难以适应多样化的情感表达场景。随着深度学习的兴起，卷积神经网络（CNN）成为图像情感分析的核心工具。Chen 等人 [7] 提出的 DeepSentiBank 利用 CNN 自动提取情感特征，显著提高了情感分类性能。但深度模型对大规模标注数据的需求使得情感分类任务面临数据标注成本高和标签不精确的问题。尽管 You 等人 [8] 通过渐进式训练和领域迁移技术改善了模型的泛化能力，但在跨领域迁移时，模型表现仍然存在下降。近年来，研究者引入了注意力机制以更好地识别图像中与情感相关的关键区域。Xu 等人 [9] 开发了多层次依赖注意力网络（MDAN），通过聚焦图像的细粒度区域来提高情感分类的准确性。然而，这类模型虽然提升了情感识别的效果，但在计算资源需求上依然过高，限制了其在大规模场景中的应用。

现有的图像单模态方法在处理情感分类时，通常无法充分捕捉图像与其他信息（如场景信息）之间的关联，特别是在面对情感复杂或模棱两可的图像时表现不佳。因此，单一依赖视觉特征的情感分析方法仍然存在一定的局限性。为

了解决这些问题，本文提出了检索增强视觉情感分析框架MIRA，通过结合图像和文本的多模态信息，利用预训练模型CLIP进行图像和文本语义空间对齐，并通过检索相关文本描述来增强视觉情感分类的效果。

1.2 检索增强的方法

检索增强生成（retrieval-augmented generation, RAG）模型[10]通过结合外部检索和大语言模型，为知识密集型任务提供了高效方案。RAG先从大规模文本库中检索相关文档，再结合预训练语言模型生成更准确的文本，从而有效减轻生成模型在事实性任务中的“幻觉问题”。此外，通过动态更新检索文档，RAG无需重新训练模型便可更新知识，应对快速变化的知识需求。在大规模语言模型（LLM）[11]领域，RAG的思想广泛应用，解决了传统静态模型难以更新知识的局限。许多LLM在开放领域问答等任务中采用检索增强机制，从外部数据库实时获取知识，提高生成准确性和上下文相关性，并使生成过程更透明、可溯源。本文提出的检索增强视觉情感分析模型受RAG启发，通过从情感文本数据库检索相关文本描述并与图像特征结合，实现多模态特征融合，不仅增强了情感分类的准确性，还能更全面地理解图像情感表达。

2 研究方法

2.1 整体框架

本文提出的检索增强视觉情感分析模型MIRA旨在通过结合视觉和文本的双模态信息提高情感分类的效果，整体框架如图1所示。

由图1中可知，模型骨干部分分为两条路径，第一条路径直接由传统特征提取模型提取图像特征，第二条路径使用CLIP将图像转换为图文向量空间中的特征表示，根据图像特征提取前k条数据（本研究中k取3），随后由Bert提取文本特征，然后将图像特征和文本特征经过交叉注意力机制进行融合。

2.2 文本数据库的构建

为得到富含丰富情感描述的文本，构建了一个情感文本描述数据库，包含560 000余条文本，具体生成过程如图2所示。

该数据库通过COCO caption[12]数据集扩展而来，通过精心构建的英文提示：“Rephrase the following sentence to strongly convey one or more of the following emotions: anger, disgust, fear, happiness, sadness, or surprise. The rephrased

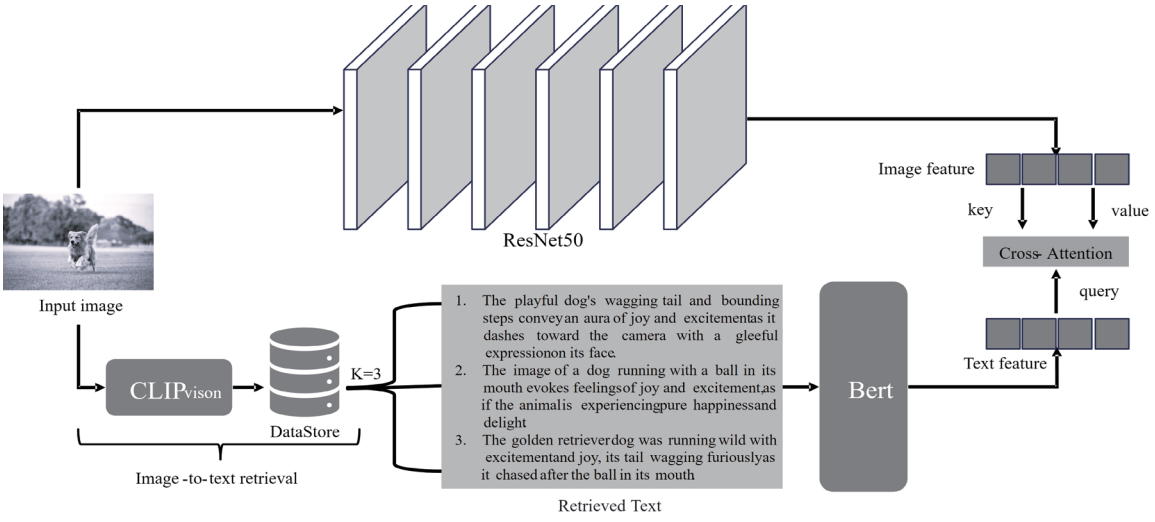

图1 整体框架图

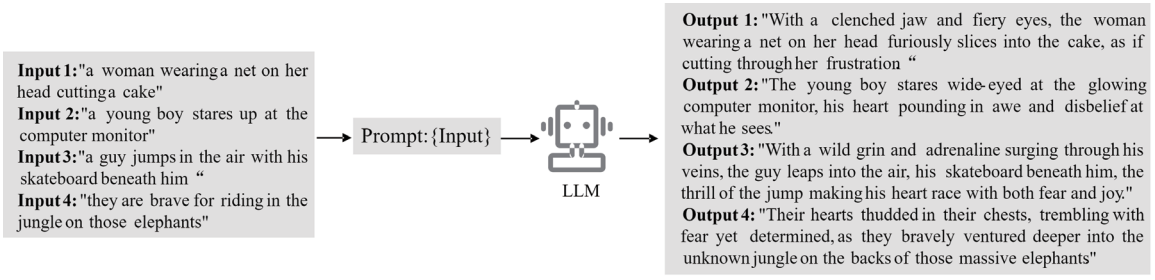

图2 情感文本数据库生成过程

sentence should evoke a clear emotional response, adding vivid emotional depth. Provide only the rephrased sentence: {description}”，利用阿里大模型 Qwen[13] 进行情感文本描述生成。

每条文本都带有基本的情感属性，涵盖了多种情感类别，文本数据库的特征向量表示为：

$$\boldsymbol{D} = \{\boldsymbol{f}_T^i \mid i = 1, 2, \cdots, N\} \tag{1}$$

式中：N 是数据库中情感描述的数量，每个文本描述 T^i 通过 CLIP 的文本编码器进行编码，生成文本特征向量 $\boldsymbol{f}_T^i \in \mathbb{R}^d$，与图像编码的向量共享同一特征空间。

2.3 相似性检索

采用预训练的 CLIP 模型提取图像在图文空间中的特征表示。CLIP 是由 OpenAI 公司提出的用于共同处理图像和文本的大预训练模型，通过对比学习同时训练图像编码器和文本编码器，使得相似的图像和文本在向量空间中更加接近。

给定输入图像，使用预训练的 CLIP 模型对图像进行编码，由此，得到图像在图文空间中的特征表示：

$$\boldsymbol{f}_I = \text{CLIP}(I), \boldsymbol{f}_I \in \mathbb{R}^d \tag{2}$$

式中：I 是输入的图像；$\boldsymbol{f}_I$ 是得到的图像特征向量；d 是图像特征向量的维度。

给定输入图像的特征向量 $\boldsymbol{f}_I$，使用 FAISS 检索系统在文本数据库 $\boldsymbol{D}$ 中建立索引，检索流程如图 3 所示。通过计算图像特征 $\boldsymbol{f}_I$ 与文本特征 $\boldsymbol{f}_T^i$ 之间的相似度：

$$\text{Similarity}(\boldsymbol{f}_I, \boldsymbol{f}_T^i) = \frac{\boldsymbol{f}_I \cdot \boldsymbol{f}_T^i}{\| \boldsymbol{f}_I \| \| \boldsymbol{f}_T^i \|} \tag{3}$$

FAISS 根据相似度从数据库中索引出与输入图像最相似的 k 条文本描述，形成文本集 Γ_k：

$$\Gamma_k = \{T_1, T_2, ..., T_k\},\ T_i = \arg\max_{T_i} \text{sin}(\boldsymbol{f}_I, \boldsymbol{f}_T^i) \tag{4}$$

式中：k=3 表示本文选择与图像最相似的前 3 条文本描述进行后续的特征提取和融合。

2.4 图文特征提取

使用预训练的 ResNet 模型对图像进行特征提取，生成的图像特征 $\boldsymbol{F}_I \in \mathbb{R}^{h \times w \times d}$ 是一个高维度的特征图，其中，h 和 w 分别表示特征图的高度和宽度；d 表示通道数：

$$F_I = \text{ResNet}(I) \tag{5}$$

对于检索到的 k 条文本描述 Γ_k，本文使用预训练的 BERT 模型提取其文本特征。假设每条文本的特征表示为 $F_{T_i} \in \mathbb{R}^{d_T}$，则提取出的文本特征表示为：

$$F_T = \{F_{T_1}, F_{T_2}, F_{T_3}\}, F_T \in \mathbb{R}^{(k \times d_T)} \tag{6}$$

式中：d_r 是文本特征向量的维度。

2.5 交叉注意力融合

为融合图像和文本的多模态特征，采用了交叉注意力机制。交叉注意力最早被使用在 Transformer[14] 中，其核心思想是通过查询（Quary）、键（Key）和值（Value）的交互，捕捉图像和文本特征之间的依赖关系。

在交叉注意力机制中，图像特征 $\boldsymbol{F}_I$ 被用作查询向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$，文本特征 $\boldsymbol{F}_T$ 被用作查询向量 $\boldsymbol{Q}$。首先计算查询 $\boldsymbol{K}$ 与 $\boldsymbol{Q}$ 的点积相似度，并对结果进行缩放和 softmax 函数处理：

$$A = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}) \tag{7}$$

式中：A 表示注意力权重；d_k 表示键向量的维度，用于防止点积值过大而引起梯度不稳定。

将注意力权重 A 与值 $\boldsymbol{V}$ 相乘，得到融合后的特征 $\boldsymbol{F}_{\text{fusion}}$：

$$\boldsymbol{F}_{\text{fusion}} = A \cdot \boldsymbol{V} \tag{8}$$

融合后的特征 $\boldsymbol{F}_{\text{fusion}} \in \mathbb{R}^{h \times w \times d}$ 包含图像特征和文本，通过

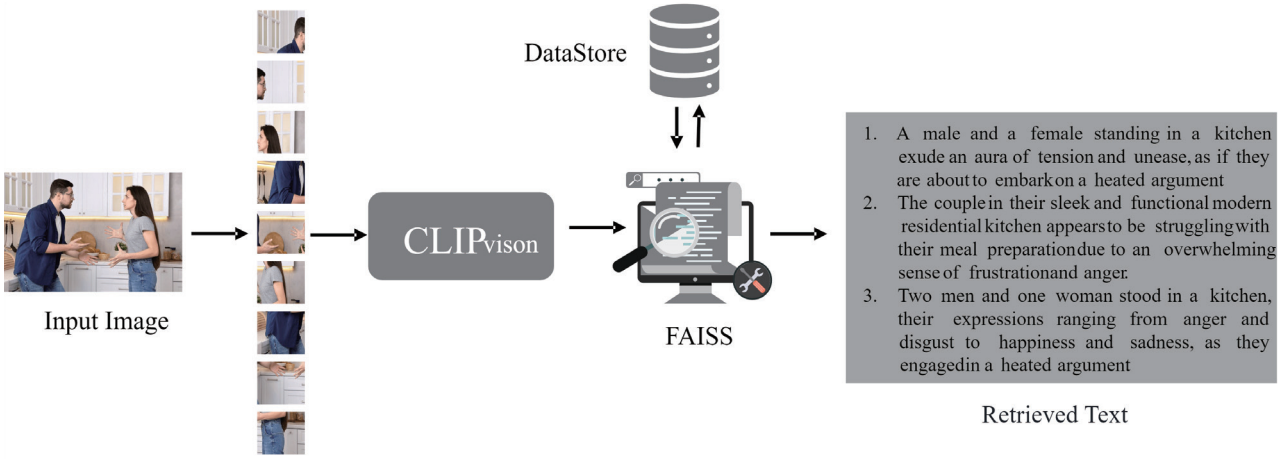

图 3 图像到文本检索过程

注意力权重调整图像的多模态交互信息。

2.6 结果预测

融合后的多模态特征 $\boldsymbol{F}_{\text{fusion}}$ 被输入到全连接层（fully connected, FC），并通过 softmax 函数进行情感分类。全连接层用于将多模态特征映射到情感类别空间，softmax 函数输出每个情感类别的概率分布：

$$\hat{y} = \text{softmax}(\text{FC}(F_{\text{fusion}})) \tag{9}$$

式中：$\hat{y}$是预测的情感类别概率分布。

3 实验分析

3.1 数据集

在本实验中，本文选用了 6 个公开的图像情感分类数据集，涵盖了抽象艺术、艺术摄影以及社交媒体等多种场景类型，如表 1 所示，包含 Abstract、ArtPhoto、Twitter I、Twitter II[15]、EmotionROI[16] 和 FI 数据集。这些数据集类型多样，规模不一，为本文所提出的情感分类模型的评估提供了坚实的基础。

表 1 数据集数量分布

数据集	类型	规模
Abstract	abstract	280
ArtPhoto	artistic	806
Twitter I	social	1 269
Twitter II	social	603
EmotionROI	social	1 980
FI	social	23 308

3.2 对比试验

为了全面验证本文提出的 MIRA 模型的有效性，本文在多个常用的图像情感分类数据集上与各种基准模型进行了对比实验。这些模型代表了不同的技术路线和创新点，能够全面评估 MIRA 在图像情感分类中的性能优势。

如表 2 所示，选用经典的视觉情感模型如 SentiBank 和 DeepSentiBank，以及结合深度学习的高级方法如 PCNN、CNN-RNN[17]、Affective region[18]、WSCNet[19]、MCPNet[20]、MDAN、PT-DPC[21] 和最新的多模态情感分析方法 SimEmotion[22] 作为对比模型。采用准确率（Accuracy）作为主要评估指标，并通过与其他方法的性能对比，全面评估各方法在不同数据集上的表现。

由表 2 可知，MIRA 模型在各个数据集上的表现均优于现有方法，尤其是在 FI 和 TwitterII 等复杂数据集上取得了显著的准确率提升（分别为 95.05% 和 85.10%）。

这说明，通过检索增强的文本信息可以为图像提供丰富的情感支持，显著提升了模型在复杂场景下的判断能力。

3.3 消融实验

为了验证本文提出的 MIRA 模型中各个模块的有效性，本文设计了多项消融实验，分别去除或修改某些模块，以此评估它们对最终模型性能的影响。主要针对以下 4 个设置进行对比实验：

（1）无文本辅助：仅使用 ResNet 提取图像特征，进行单模态图像分类，不包含文本信息。

（2）简单拼接图文特征：将图像特征与检索到的文本特征进行简单拼接（Concatenation），而不进行交叉注意力融合。

（3）仅使用一条检索到的文本描述：在检索模块中，只保留一条检索到的文本描述，而不是多条文本融合。

（4）使用完整模型：即包含检索增强、交叉注意力机制，并使用多条检索到的文本描述的完整 MIRA 模型。

表 2 对比实验图

方法	Abstract	ArtPhoto	Twitter I	Twitter II	EmotionROI	$\boldsymbol{F}_1$
SentiBank	0.643 0	0.673 3	0.666 3	0.659 3	0.352 4	0.564 7
DeepSentiBank	0.690 7	0.702 6	0.712 5	0.702 3	0.425 3	0.643 9
PCNN	0.702 6	0.714 7	0.825 4	0.776 8	—	0.735 9
CNN-RNN	0.738 8	0.755 0	—	—	—	0.842 6
Affective region	0.760 3	0.748 0	0.810 6	0.804 8	—	0.863 5
WSCNet	—	—	0.842 5	0.813 5	0.582 5	—
MCPNet	—	0.792 4	0.897 7	0.811 9	—	0.897 7
MDAN	0.833 0	0.781 2	—	—	0.616 6	0.833 0
PT-DPC	—	—	0.896 5	0.834 6	0.697 0	0.938 9
SimEmotion	—	—	0.897 6	0.842 1	0.704 0	0.897 6
MIRA (Ours)	**0.865 4**	**0.821 5**	**0.910 3**	**0.851 0**	**0.735 0**	**0.950 5**

还是在 6 个基准数据集上进行了消融实验，实验设置保持一致，其他参数保持不变，以确保不同模块对模型性能的独立影响。

在消融实验中，本文针对不同的模块进行了性能评估。如表 3 所示，整体模型在任何一个数据集上都优于消除某个模块的情况。移除检索生成模块并仅使用图像特征的情况下，模型在所有数据集上的表现显著下降。例如，在 EmotionROI 数据集中，准确率从 73.50% 下降至 65.40%，表明情感相关的文本描述为图像情感分类提供了关键的情感信息支持。通过这一模块的增强，模型能够更有效地捕捉图像中的复杂情感表达。

当本文移除融合机制时，仅通过简单拼接图像和文本特征时，模型的表现也有所下降。在 Twitter I 和 EmotionROI 数据集中，准确率分别下降了 4.68% 和 4.63%。交叉注意力机制能够更好地融合图像和文本的情感信息，使图文模态情感的交互更加紧密。简单拼接特征未能充分捕捉到图像与文本之间的复杂情感关系，反而导致分类性能的下降。

只使用一条检索到的文本描述时，模型表现略微下降。在 $\boldsymbol{F}_1$ 和 EmotionROI 数据集中，准确率分别从 82.50% 降至 80.01% 和从 73.50% 降至 71.25%。这表明多条文本描述能够提供更加多样化的情感信息，有助于增强模型在处理复杂情感时的鲁棒性。

完整的 MIRA 模型在所有数据集上的表现最为优越。例如，在 $\boldsymbol{F}_1$ 和 Twitter II 数据集上，模型的准确率分别达到了 95.05% 和 85.10%。这表明检索生成、交叉注意力融合机制以及多条文本描述的协同作用，使得 MIRA 在多模态情感分类任务中具有强大的适应性和表现力。综上所述，完整的 MIRA 模型在情感分类任务中具有明显优势，尤其是在处理复杂情感信息时，表现出了强大的鲁棒性和适应性。

4 结论

本文提出的检索增强视觉情感分析模型 MIRA 通过结合图像与文本的模态信息，利用检索技术从大规模情感文本数据库中提取与图像情感相关的描述，显著提升了情感分类的准确性。实验结果表明，该模型在多个数据集上超越了现有方法，尤其在处理情感复杂或模糊的图像时表现优越。交叉注意力机制有效地融合了图像和文本特征，使用文本特征增强了图像的情感表达，使模型能够更深刻地理解图像情感内容。未来工作将致力于提高文本数据库质量，优化模型结构以及提出更优的检索策略，增强视觉情感分析的效率和适用性。

参考文献：

[1] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[DB/OL].(2021-02-26)[2024-03-19].https://doi.org/10.48550/arXiv.2103.00020.

[2] DOUZE M, GUZHVA A, DENG C Q, et al. The faiss library[DB/OL].(2024-01-16)[2024-06-22].https://doi.org/10.48550/arXiv.2401.08281.

[3] 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, 1: 770-778.

[4] DEVLIN J. CHANG M W, LEE K,et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Brussels: ACL, 2019: 4171-4186.

[5] MACHAJDIK J, HANBURY A. Affective image classification using features inspired by psychology and art theory[C]//Proceedings of the 18th ACM international conference on Multimedia. NewYork: ACM,2010: 83-92.

[6] BORTH D, JI R R, CHEN T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C]//Proceedings of the 21st ACM international conference on Multimedia. NewYork:ACM,2013: 223-232.

表 3 消融实验结果对比

Ablation Setting	Abstract	ArtPhoto	Twitter I	Twitter II	EmotionROI	$\boldsymbol{F}_1$
完整模型 (MIRA)	0.865 4	0.821 5	0.910 3	0.851 0	0.735 0	0.950 5
无文本辅助	0.666 4	0.710 8	0.781 3	0.782 3	0.654 0	0.890 4
简单拼接图文特征	0.789 4	0.744 3	0.863 5	0.815 0	0.688 7	0.912 0
仅使用一条检索文本	0.811 0	0.768 4	0.880 1	0.832 1	0.712 5	0.930 2

[7] CHEN T, BORTH D, DARRELL T, et al. DeepSentiBank: visual sentiment concept classification with deep convolutional neural networks[DB/OL]. (2014-10-30)[2024-06-23]. https://doi.org/10.48550/arXiv.1410.8586.

[8] YOU Q Z, LUO J B, JIN H L, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]//Proceedings of the AAAI conference on Artificial Intelligence. NewYork: ACM, 2015: 381-388.

[9] XU L W, WANG Z T, WU B, et al. MDAN: multi-level dependent attention network for visual emotion analysis[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 9479-9488.

[10]LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. NewYork: ACM, 2020: 9459-9474.

[11]GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: a survey[DB/OL]. (2024-03-27)[2024-06-13].https://doi.org/10.48550/arXiv.2312.10997.

[12]CHEN X L, FANG H, LIN T Y, et al. Microsoft COCO captions: data collection and evaluation server[DB/OL].(2015-04-03)[2024-09-11].https://doi.org/10.48550/arXiv.1504.00325.

[13]YANG A, YANG B S, HUI B Y, et al. Qwen2 technical report[DB/OL].(2024-09-10)[2024-10-13].https://doi.org/10.48550/arXiv.2407.10671.

[14]VASWANI A, SHAZEER N, PARMAR N,et al.Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.NewYork: ACM, 2017: 6000-6010.

[15]PENG K C, SADOVNIK A, GALLAGHER A, et al.Where do emotions come from? predicting the emotion stimuli map[C/OL]//2016 IEEE International Conference on Image Processing (ICIP). Piscataway:IEEE,2016[2024-01-13]. https://ieeexplore.ieee.org/document/7532430.DOI:10.1109/ICIP.2016.7532430.

[16]YOU Q Z, LUO J B, JIN H L,et al.Building a large scale dataset for image emotion recognition: the fine print and the benchmark[C]//AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.NewYork: ACM, 2016: 308-314.

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

[18]YANG J F, SUN M,CHENG M M,et al. Visual sentiment prediction based on automatic discovery of affective regions[J].IEEE transactions on multimedia, 2018, 20(9): 2513-2525.

[19]SHE D Y, YANG J F, CHENG M M, et al. WSCNet: weakly supervised coupled networks for visual sentiment classification and detection[J].IEEE transactions on multimedia, 2020,22(5):1358-1371.

[20]OU H C, QING C M, XU X M, et al. Multi-level context pyramid network for visual sentiment analysis[J]. Sensors, 2021, 21(6): 2136.

[21]DENG S N, WU L F, SHI G, et al. Learning to compose diversified prompts for image emotion classification[J]. Computational visual media, 2024,10:1169-1183.

[22]DENG S N, WU L F. SHI G,et al.Simple but powerful, a language-supervised method for image emotion classification[J]. IEEE transactions on affective computing, 2022(14): 3317-3331.

【作者简介】

程立（1998—），通信作者（email:charnlee@163.com），男，安徽亳州人，硕士研究生，研究方向：计算机视觉、大语言模型、多模态大语言模型。

张虹（1977—），女，山西太原人，博士，副教授，研究方向：人工智能、区块链与智能数据。

（收稿日期：2024-12-05）