http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0345

结合上下文关联的图像情感分析

罗改芳[1,2]，张浩[1]，徐丹[1,*]

(1. 云南大学 信息学院，昆明 650500；2. 山西农业大学 软件学院，晋中 030801)

摘 要：图像情感分析旨在分析和理解视觉内容所传达的情感，其挑战在于弥合潜在视觉特征与抽象情感间的情感鸿沟。现有的深度模型试图一次性通过直接在全局范围内学习有辨别力的高级情感表征来弥合鸿沟，但忽略了深度模型各层特征之间的层次关系，导致上下文特征间的关联缺失。为此，提出一种上下文层次交互网络（CHINet）来建立层次结构中的上下文信息和情感之间的相关性模型。该模型包含2个分支：自下而上的主分支直接在高级语义层次上学习全局情感表征，针对该分支的不同层次特征，通过构建浅层风格编码器和情感激活注意力机制来分别提取风格表示并定位潜在情感激活区域；所提取的特征被级联到金字塔结构作为自上而下分支，从而建模上下文层次相关性并为情感表示提供浅层视觉特征。通过全局和局部学习将低级风格属性和高级图像语义整合到一起。实验结果表明：所提模型在FI数据集上较同类方法（包括多层次特征融合方法和结合了局部情感区域的方法）提升了情感识别准确率。

关 键 词：图像情感分析；情感鸿沟；上下文关联；层次结构；情感分类
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2025)07-2304-10

“一图胜千言”，与具体的图像语义感知任务相比，在抽象层面研究视觉情感更具挑战性，这一开创性的课题被称为图像情感分析[1]。根据心理学的描述，图像情感大致可分为2种情感极性，即积极极性和消极极性，其又可以进一步细分为子情感类别[2]。图像情感与复杂的视觉特征密切相关，对于观者而言，不论是图像中的高级语义概念还是低级视觉特征，如颜色、纹理等视觉元素都可以诱发人类情感[3-4]。抽象情感与具体视觉特征紧密相关却又无法被直观地通过数学语言表达，因此，设计一个综合表征来弥合情感鸿沟是情感分析的关键[5-6]。

现有的基于深度学习的方法试图通过在图像全局范围内直接学习得出最终的高级情感表征进行情感表达，然而深度神经网络的特征在信息传递上存在上下文关联，即信息传递越深，特征表达的层次就越高[7-8]。这些方法忽略了视觉特征之间的层次关系，导致深度神经网络层次关系中的上下文信息难以关联，因此，仅使用高层次特征来表示情感是不够的。同时，已有研究也表明颜色、纹理等底层信息也是与视觉情感相关的重要组成部分[9-10]。

为解决上述问题，本文提出一种上下文层次交互网络（context hierarchical interaction network, CHINet）模型用于图像情感推理。本文模型由2个分支组成：自下而上的分支通过主干网络从全局角度对图像语义中的情感进行分类；自上而下的分支通过浅层风格编码器从主干网络中提取细粒度的图像风格。图像风格用于将颜色、纹理和线条等低级视觉特征整合到一起[11-12]，而语义则从高级图像内容形成情感表征。此外，为融合层次上下文特征，设计了一种情感注意力机制（emotional attention mechan-

收稿日期：2023-06-12；**录用日期**：2023-09-08；**网络出版时间**：2023-11-07 14:39
网络出版地址：link.cnki.net/urlid/11.2625.V.20231106.1919.001
基金项目：国家自然科学基金(62162068,62061049)；云南省万人计划“云岭学者”专项(YNWR-YLXZ-2018-022)；云南省教育厅科学研究基金(2025J0008)

*通信作者. E-mail：danxu@ynu.edu.cn

引用格式：罗改芳，张浩，徐丹. 结合上下文关联的图像情感分析[J]. 北京航空航天大学学报，2025，51(7)：2304-2313.
LUO G F，ZHANG H，XU D. Image sentiment analysis by combining contextual correlation[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2304-2313 (in Chinese).

ism, EAM)作用于 2 个分支间的金字塔结构中, 抽取神经网络中的各层级属性, 并构建相邻上下文特征图的相关性。整个模型通过 2 个分支将图像分离为高级语义和低级风格, 并通过浅层风格编码器及情感激活区域提取模块建立网络层级中的上下文关联。引入全局和局部联合学习来优化模型, 用于从全局到局部视角将 CHINet 集成到统一的端到端框架中。本文的主要创新点如下:

1) 提出一种浅层风格编码器将卷积神经网络(convolutional neural networks, CNN)层次结构中不同级别的特征进行分离, 充分考虑浅层风格特征和高级语义的融合来进行情感推理。

2) 设计了一种情感注意力机制用于 CNN 主干与浅层风格编码器之间, 并通过图像金字塔建模层次特征中的上下文关联。

3) 使用全局学习和局部学习的统一框架, 完成上下文特征之间的协调优化, 从而提高情感分类的准确度。

1　相关工作

1.1　情感分析模型

计算机视觉和多媒体领域通常以心理学常用的 2 种方式描述情感: 离散情感状态(categorical emotion states, CES)和维度情绪空间(dimensional emotion space, DES)[13]。本文更关注 CES, 与 DES 相比, CES 可解释性更强。由于情绪的复杂性, 研究者对离散情感状态存在多种界定方式, 分类数量也各不相同。本文选择具有代表性的 2 类情感极性理论和八类情感状态来综合评估本文模型。2 类情感极性将情绪分为积极和消极, 而 8 类情感理论包括愉悦、敬畏、满足、激动、恐惧、悲伤、恶心和愤怒, 不同类型的情感图像示例如图 1 所示。早

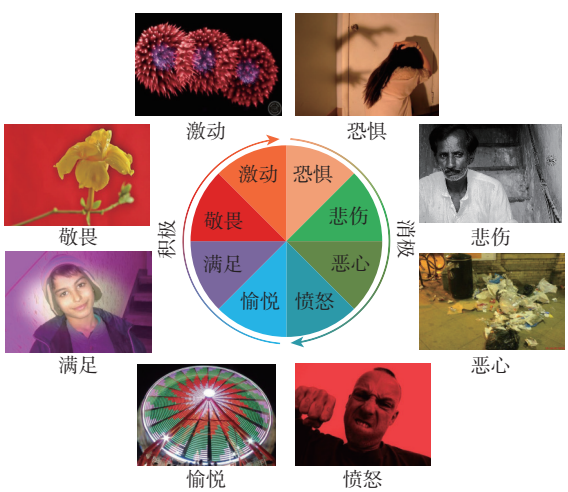


图 1　来自 FI 数据集的情感图像示例

Fig. 1　Affective image examples from FI dataset

期, 研究者从心理学的角度出发, 人为设计并提取视觉特征来表达情感, 与从 CNN 中提取的特征相比, 这些手工制作的特征主要集中在低级视觉线索上, 而高级语义没有得到充分利用[13]。为此, 本文更注重采用深度学习的思路, 构造一种能够提取图像上下文关系的神经网络模型用于情感分析。

1.2　基于深度学习的全局图像情感特征提取

已有大量工作将深度神经网络用于情感分析。Peng 等[14]利用迁移学习将预训练的 CNN 微调, 并在 EmotionROI 数据集上进行情感识别; Campos 等[15]通过消融实验对 AlexNet[16]进行逐层分析。结果均证明 CNN 可以捕获比手动方法更好的情感表征。一些工作将高级语义与低级视觉特征相结合来指导情感分类, 这类方法与本文工作直接相关, 都采用了来自于视觉领域被广泛使用的 CNN, 并根据不同层次设计出针对性的特征提取模块。Zhu 等[17]设计了一个联合 CNN 和循环神经网络(recurrent neural networks, RNN)的模型, 通过卷积层提取视觉和语义特征, 然后, 使用双向循环神经网络进行融合。文献 [10] 和文献 [18] 提出多层次深度网络, 结果表明图像情感可以通过图像语义、美学和低级特征来识别。She 等[19]则从弱监督的角度切入, 提出一种用于视觉情感识别的多核网络。You 等[20]提出一种渐进式 CNN(Progressive CNN, PCNN)架构, 有效利用弱标记数据提高模型泛化能力。DeepSentiBank[21]是一种基于 CNN 的视觉情感概念提取块, 用于发现形容词-名词对, 可提取 2 089 维向量作为中层表示。基于贝叶斯网络, Zhang 等[22]提出一种对象语义情感相关模型来指导图像情感分类。Yamamoto 等[23]提出一种同时使用视觉特征和语义的图像情感识别算法, 通过现有对象检测模型的高级嵌入获得语义, 而 CNN 提取基本视觉特征, 最终实现情感分类。

然而, 目前大多数深度模型均是从输入图像的全局视角中直接学习到网络最后的深层次表征, 并没有仔细考虑情感分析中涉及的情感特征演化过程, 从而忽略了深度模型中的层次关联。基于以上因素, 本文充分考虑网络中不同层次特征的上下文相关性, 针对不同层次的特征设计不同的情感特征提取模块进行最终的图像情感推理。

1.3　融合情感激活区域的局部情感特征提取

结合情感激活区域的方法也与本文模型密切相关, 这类方法结合了图像的局部和全局特征[24-26]。Peng 等[27]提出的监督情感区域检测方法使用情感激活图来表示每个像素对诱发情感的贡献, 该方法

综合考虑了不同区域对情感的影响。

She等[3]提出一种带有侧分支的弱监督耦合网络，通过侧分支检测情感诱发区域并将其反馈到模型主分支用于最终的图像情感分类，取得了良好的效果。Sun等[28-29]提供了一个新的思路，其使用双分支策略结合局部信息和图像整体表征：先检测出若干情感候选区域并对每个区域进行评分；然后，根据评分确定出首选情感激活区域；最后，将图像整体和局部情感激活区域信息进行融合。此外，Zhang等[30]提出一种多层次的情感区域相关性分析（multi-level sentiment region correlation analysis, MSRCA）模型，该模型首先从原始情感图像多个层次中发现与情感潜在相关的区域，然后，使用Transformer编码器从多个角度探索区域之间的相互作用和依赖关系。Xiong等[31]提出的基于群稀疏正则化的区域卷积神经网络（region-based convolutional neural networks using group sparse regularization, R-CNNGSR）也通过自动检测情感区域来提升情感识别准确率。

基于情感区域构建的情感分类器可有效地从情感区域中提取情感表征，增加模型对于局部信息的关注，同时避免接触到非情感区域的冗余噪声。然而，以往的方法大多采用现有的目标检测工具来定位情感激活区域，这类检测方案通常是为目标检测等视觉任务设计的，其丧失了模型自动探索用于情感识别的区域局部特征的可能性。因此，如何定位潜在情感诱发区域也是值得考虑的问题。基于以上挑战，本文提出情感激活注意力机制，探索用于图像局部情感表示的解决方案。

2 情感推理模型

2.1 上下文层次交互网络

相关研究表明，虽然图像语义更能唤起情绪，但纹理、颜色等潜在的底层特征也有利于情感表达[32-33]。现有的用于情感分析的深度模型大多从图像级别学习高级图像语义以弥合情感鸿沟，导致浅层特征的缺失。为此，本文通过分离浅层风格特征和高级语义来进行情感推理。

受图像风格迁移的启发，采用图像风格表示浅层视觉特征，因为其包含了线条、纹理、颜色等。本文提出的上下文层次网络由2个分支组成，模型的整体结构如图2所示，其中，$\boldsymbol{F}_l$为特征图，MLP表示多层感知器。自下而上的分支采用了预训练的ResNet[34]，预训练的模型为提取视觉语义提供了先验知识。ResNet由若干串联的残差块组成，通过极大池化进行下采样。ResNet模型共包含4个阶段，根据模型中特征图的大小是否相同将模型中的层划分为是否在同一阶段。最后的高级语义表征经过一个全局平均池化层所得出。为充分利用中间网络层特征的上下文关系，浅层风格编码器提取每个阶段后的特征进行图像风格表示。不同级别特征的风格表示被合并到自上而下的分支中，并通过层次金字塔结构连接在一起，用于浅层特征表示，以消除情感差距。图3为浅层风格编码器的计算过程，特征具体维度被标识在组件下方以便于查看。

形式化地，记模型第l层的特征图$\boldsymbol{F}_l \in \mathbf{R}^{H\times W\times C}$，其中，$H$和$W$为特征图的高和宽，$C$为通道数，$\boldsymbol{F}_l$中

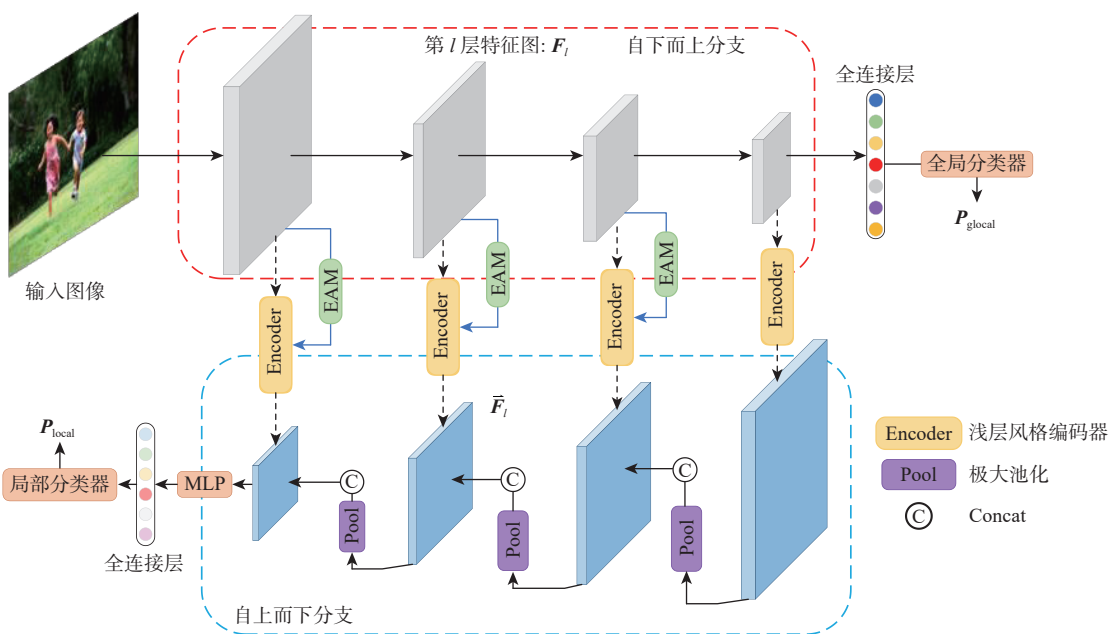


图2 CHINet整体结构

Fig. 2 Overall structure of CHINet

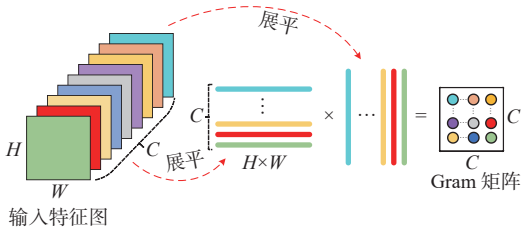


图 3　浅层风格编码器

Fig. 3　Shallow style encoder

的第 i 个通道的特征图为 $\boldsymbol{x}_i \in \mathbf{R}^{H\times W}$。对于第 l 层特征图 $\boldsymbol{F}_l$ 中的每个特征图 $\boldsymbol{x}_i$, 将 $\boldsymbol{x}_i$ 展平为大小为 $H\times W$ 的一维向量。然后, 将变换后的特征图 $\boldsymbol{F}'_l \in \mathbf{R}^{H\times W\times C}$ 乘以其转置形成 Gram 矩阵 $\boldsymbol{G}\in \mathbf{R}^{C\times C}$, 其表达式为

$$\boldsymbol{G} = (\boldsymbol{F}_{l:C} \to \boldsymbol{F}'_l)\otimes(\boldsymbol{F}_{l:C} \to \boldsymbol{F}'_l)^{\mathrm{T}} \tag{1}$$

式中: ⊗表示矩阵乘法。Gram 矩阵中的每个元素 $G_{i,j}$ 是扁平化后的特征 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的内积:

$$G_{i,j} = \frac{1}{C\times H\times W}\sum_{k=1}^{n} x_{ik}x_{jk} \tag{2}$$

其中: $1/(C\times H\times W)$ 为归一化因子, 用于模型优化。

2.2　情感激活区域提取模块

盲目地考虑整幅图像的所有信息可能会削弱特征表达情绪的能力, 从而降低情感分类的准确性[35]。本节提出一种情感注意力机制来获取情感激活图, 模块结构如图 4 所示。与以往的注意力机制不同, 本文 EAM 用于获取情感诱发区域的注意力图, 同时过滤掉大部分噪声, 模块强大的自适应学习能力可以将原始输入特征图映射到若干弱相关的注意力图中。

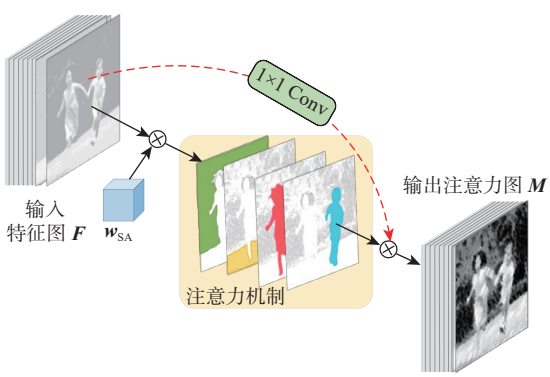


图 4　情感激活注意力机制

Fig. 4　Emotion activation attention mechanism

对于第 l 层特征图 $\boldsymbol{F}_l \in \mathbf{R}^{H\times W\times C}$, 使用逐点卷积将每个位置像素 $\boldsymbol{x}_i \in \mathbf{R}^{H\times W}$ 映射到 n 个情感激活图之一。在每个映射中, 在空间维度上池化像素以获得情感激活图 $\boldsymbol{M} \in \mathbf{R}^{n\times C}(n << HW)$:

$$\boldsymbol{M} = \underset{HW}{\mathrm{softmax}}(\boldsymbol{F}_l\boldsymbol{w}_{\mathrm{SA}})^{\mathrm{T}}\boldsymbol{F}_l \tag{3}$$

式中: 自适应滤波器 $\boldsymbol{w}_{\mathrm{SA}} \in \mathbf{R}^{C\times n}$ 用于从输入特征 $\boldsymbol{F}_l$ 转换成情感激活图, 经 softmax(·) 激活将其转化为维度为 $H\times W$ 的注意力权重。将情感激活图 $\boldsymbol{M}$ 与图像 Gram 矩阵相乘, 得到输入特征图 $\boldsymbol{F}_l$ 对应的 $\tilde{\boldsymbol{F}}_l$, 最后, 由 1×1 卷积进行不同通道之间的信息交互:

$$\tilde{\boldsymbol{F}}_l = \max(\boldsymbol{0},\ \mathrm{Conv1\times 1}(\boldsymbol{M}\otimes\boldsymbol{G})) \tag{4}$$

2.3　全局和局部联合学习

本文 CHINet 模型包含 2 个分类器, 即全局分类器和局部分类器, 预测概率分别为 $\boldsymbol{P}_{\mathrm{global}}$ 和 $\boldsymbol{P}_{\mathrm{local}}$, 这使得模型不仅能够专注于特定级别的情感表征, 如低级视觉细节或高级语义特征, 同时也能更好地提取层级特征, 这对于不同类型的情感图像至关重要。例如, 社交媒体图像更倾向高级语义, 而抽象绘画更倾向于颜色、线条等低级视觉风格。同时优化 2 个分类器可以帮助模型发挥优势并避免劣势。基于以上因素, 本节采用全局和局部学习来优化模型。

对于局部分类器, 给定训练样本 $\{(x, y)\}$, 其中, x 为图像, y 为对应情感标签, CNN 使用卷积层和全连接层提取输入图像的逐层表示, 在最后的 softmax 层之后, 最后一个全连接层的输出被转换为 c 类图像情感的概率分布 $\boldsymbol{P}\in \mathbf{R}^{1\times c}$。在本文中, c=2 或 8 表示 2 个或 8 个情感类别。情感图像样本属于特定情感的概率被定义如下:

$$p_i = \frac{1}{N}\sum_{j=0}^{N} \mathrm{e}^{h_{i,j}} \Big/ \sum_{i=0}^{c} \mathrm{e}^{h_{i,j}} \qquad i\in[1,c] \tag{5}$$

式中: $h_{i,j}$ 为全连接层的输出; N 为样本数。预测概率的损失值通过交叉熵衡量:

$$L_{\mathrm{cls}} = -\sum_{i=0}^{c} y_i \lg p_i \tag{6}$$

其中: $y_i = \{y_i | y_i \in \{0,1\}, i = 1,2,\cdots,c\}$表示图像的情感标签; $\sum_{i=1}^{c} p_i = 1$。

在本文实验过程中, 首先, 通过全局分类器对主分支(自下而上分支)进行训练, 即固定住自上而下分支的权重, 调整主分支权重; 然后, 通过局部分类器对自上而下的分支进行训练, 即类似地固定住主分支的权重, 调整自上而下分支权重。然而, 单纯的使用以上训练方式将导致模型的预测结果更加偏向最后完成训练的分类器的结果, 即偏向局部分类器的预测概率 $\boldsymbol{P}_{\mathrm{local}}$, 导致情感识别准确率不高(实验结果见 3.2.2 节)。为使得模型作为一个端到端的整体架构, 在优化用于全局学习的分支局部分类器之后, 将 2 个情感表征分类器的预测结果进行加权并同时优化:

$$\boldsymbol{P} = \left[(1-\alpha)\boldsymbol{P}_{\mathrm{global}} + \alpha\boldsymbol{P}_{\mathrm{local}}\right]/N \qquad \alpha\in[0,1] \tag{7}$$

式中：α为超参数，用于平衡全局分类器和局部分类器对于模型最终预测结果的影响。由于 2 个分类器是相互独立事件，类似的，通过最小化交叉熵损失函数来学习最终的情感概率表达：

$$L_{\mathrm{cls}}=-\frac{1}{N}\sum_{j=0}^{N}\sum_{i=0}^{c}y_i\lg p_i \tag{8}$$

其中：c 为情感类别数。

3 实验结果和分析

3.1 实验细节

本文所有实验均在开源深度学习开发平台 MXNet 上实现。模型参数通过小批量梯度下降法进行优化，对于二元情感分类和多元情感识别任务，批量分别设置为 16 和 32。采用动量法加速模型收敛，动量系数设为 0.9。学习率初始化为 0.001，每 30 轮乘 0.1 直至模型收敛。同时，采用图像增强方案用于随机改变训练样本以减少模型对特定属性的依赖，从而提高泛化能力，即将输入图像尺寸调整为 256×256，然后，随机裁剪回 224×224，同时按 0.5 概率镜像翻转。最后，所有输入图像被规范化后输入模型进行训练。所有实验均在 2 块 Nvidia RTX2080ti 显卡上进行。

本文选择 5 个图像情感数据集验证本文模型的情感识别准确率，包括 Abstract[9]、ArtPhoto[9]、Twitter I[20]、EmotionROI[14] 和 FI[1]，表 1 为数据集对应详细信息。为充分验证本文模型的有效性，本文不仅讨论了二元情感分类，还探索了多元情感识别，二元情感分类即将该任务视为积极、消极的情感极性识别问题，多元情感识别则按原始数据集标签进行 8 类情感识别任务。对于二元情感极性分类任务，除 Twitter I 数据集最初用二元情感标记外，其他数据集则根据维度情绪空间中的愉悦度（Valance）将多情感标签转换为积极和消极极性。

表 1 图像情感数据集详细信息

Table 1 Details of image emotion datasets

数据集	积极极性数量					消极极性数量					合计
	愉悦	敬畏	满足	激动	小记	愤怒	恶心	恐惧	悲伤	小记	
Abstract[9]	25	15	63	36	139	3	18	36	32	89	228
ArtPhoto[9]	101	102	70	105	378	77	70	115	166	428	806
EmotionROI[14]			330	330	660	330	330	330	330	1 320	1 980
FI[1]	4 942	3 151	5 374	2 963	16 430	1 266	1 658	1 032	2 922	6 878	23 308
Twitter I[20]					769					500	1 269

3.2 消融实验和参数分析

3.2.1 主干选择

为确定主干网络，先分析了几种广泛使用的 CNN 模型，包括 AlexNet[16]、VGG[36]、DenseNet[37] 和 ResNet[34]，因为其在视觉领域稳定且优秀的性能表现和相对统一且一致的框架结构，所有模型参数均来自开源开发平台 MXNet 模型库并在 FI 数据集上进行了微调。表 2 为各模型在 FI 上的情感分类准确率。由于无论是局部学习还是全局学习，CHINet 的模型最后都采用了全连接层 fc，所以表中包含了对最后一个全连接层在情感分析任务中作用的分析，这与本文工作密切相关。由表可知，ResNet 系列取得了最好的效果。特别是 152 层的 ResNet 表现最好，准确率为 85.19%。

相比之下 AlexNet 准确率最差，因为其只有 5 个卷积层，但对于同一系列的网络，具有更多层的模型则效果也更好。为此，本文选择 ResNet-152 作为主干模型。还发现 DenseNet 虽然有 100 多个卷积层，但其效果仍然不够理想，因为每个卷积层中的通道数有限。不仅如此，VGG 优于具有更多卷积层的 ResNet-18，因为最后的全连接层提供了更多可学习的参数，可以将学习到的“分布式表示”映射到特定的情感标签。表 2 中的实验结果表明，网络末端的全连接层有助于情感图像识别，因为其

表 2 FI 上不同层的 CNN 情感分类准确率

Table 2 Emotion classification accuracy of CNN models with different layers on FI

模型	层	分类准确率/%	
		c=2	c=8
AlexNet[16]	5 Conv + 2 fc	65.43	42.78
VGG-16[36]	13 Conv + 2 fc	81.35	63.84
VGG-19[36]	16 Conv + 2 fc	82.62	64.23
DenseNet-100 w/o fc[37]	99 Conv	83.77	64.76
DenseNet-100[37]	99 Conv + 2 fc	84.21	65.08
ResNet-18 w/o fc[34]	17 Conv	80.16	60.76
ResNet-18[34]	17 Conv + 2 fc	81.27	61.55
ResNet-50 w/o fc[34]	49 Conv	83.52	62.81
ResNet-50[34]	49 Conv + 2 fc	84.06	63.74
ResNet-152 w/o fc[34]	151 Conv	84.73	66.27
ResNet-152[34]	151 Conv + 2 fc	85.19	66.44

提供了冗余参数来构建情感推理的高级概念。基于以上原因, 本文同样引入类似的全连接层。

3.2.2　全局和局部学习

在 2.3 节中提到, 应用全局和局部学习来优化本文提出的深度模型。本节讨论优化策略对模型性能的贡献, 结果如表 3 所示, 表中, CHINet w/o G 和 CHINet w/o L 分别表示在没有全局学习和没有局部学习训练情况下模型的情感识别准确率结果, CHINet G&L 则表示同时使用全局和局部学习训练下的模型结果。基线模型是经过微调的 ResNet-152。由表 3 可知: 与局部学习相比, 同时优化局部和全局分类器可提高分类精度 2.38%; 其次, 使用全局学习比局部学习更好, 因为全局学习可以综合 2 个分支不同层次的情感表征。结果证明局部学习和全局学习的优化方法之间能够有效互补。

此外, 分类器权重参数 α 也是关键因素之一, 其决定了模型优化时 2 个分类器对模型的贡献。图 5 为不同 α 下性能的变化趋势, 当 α=0 时, 模型退化为只使用局部学习, 反之则代表只使用了全局学习。可以发现当 α=0.6 时, 模型的情感分类准确率最好, 说明同步优化 2 个分类器有利于模型的情感推理, 在后续实验中遵循了 α=0.6 这一设置。

表 3　基于局部和全局学习的情感分类准确率

Table 3　Emotion classification accuracy based on local and global learning

模型	准确率/%
基线模型	85.19
CHINet w/o G	87.36
CHINet w/o L	88.41
CHINet G&L	89.74

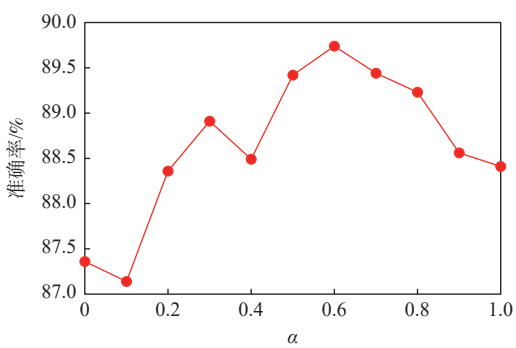


图 5　不同 α 对模型性能的影响

Fig. 5　Effect of different α on model performance

3.3　结果对比

为证明本文模型的有效性, 本节将其与其他方法进行比较, 包括几种经典的手工提取特征算法和最新的深度模型。

首先, 进行二元情感分类。表 4 为本文模型在 FI 上的性能。可以观察到, 本文模型优于为对象分类任务设计的深度 CNN, 如 AlexNet、VGG、DenseNet 和 ResNet。同时, 本文 CHINet 比其他为情感分类设计的模型(包括 PCNN[20] 和 DeepSentiBank[21])拥有更好的性能。与同样构建侧分支, 如情感区域(affective regions, AR)[29]、视觉和情感特征(visual and semantic features, VSF)[23] 相比, CHINet 仍然以 89.74% 的情感识别准确率实现最佳性能。同时, 基于多层次表示的 CHINet 比类似的基于多级特征融合(multi-level representations, MLR)[18] 和 MSRCA[30] 获得更好的性能, 这表明本文模型受益于上下文层次表示和更合适的优化方法。

表 4　FI 数据集上二元情感极性分类准确率比较

Table 4　Classification accuracy comparison of binary emotion polarity on FI dataset

模型	网络	准确率/%
基准模型	AlexNet[16]	60.54
	VGGNet[36]	70.64
	ResNet[34]	72.22
	Fine-tuned AlexNet[6]	72.43
	Fine-tuned VGGNet[36]	83.05
	Fine-tuned ResNet-50[34]	85.19
对比方法	PCNN (VGGNet)[20]	75.34
	DeepSentiBank[21]	61.54
	AR[29]	86.35
	VSF[23]	88.11
	MSRCA[30]	87.40
	MLR[18]	87.87
本文模型	CHINet w/o G	87.36
	CHINet w/o L	88.41
	CHINet G&L	89.74

此外, 表 5 为当前最先进的方法在几个小规模数据集上的二元情感表现, 表中, 无数据表示本文未能复现出的结果。由表可知, CHINet 在所有数据集上始终优于其他模型。对于 ArtPhoto 和 Abstract 数据集, 本文模型有相对明显的性能提升, 分类准确率分别为 83.27% 和 82.33%。值得一提的是, ArtPhoto 和 Abstract 数据集由缺乏对象或语义但充满低级特征(如纹理、颜色等)的抽象绘画组成。这证明了浅层视觉细节的重要性。除此之外, 全局和局部学习也是必不可少的组成部分, 可以使模型的性能全面超越其他方法。

为进一步评估本文模型的有效性, 本文还在 8 类别情感识别任务上进行了实验, 表 6 为具体实验结果。表中手工特征算法被删除对比, 因为其效果较弱不具备直接对比的价值。实验中采用预训

表5 小规模数据集上的二元情感极性分类准确率比较

Table 5 Classification accuracy comparison of binary emotion polarity on small-scale datasets %

模型	网络	ArtPhoto[9]	Abstract[9]	Twitter I[20]			EmotionROI[14]
				Twitter I 5	Twitter I 4	Twitter I 3	
手工特征	PAEF[13]	67.85	70.05	72.90	69.61	67.92	75.24
	SentiBank	67.74	64.95	71.32	68.28	66.63	66.18
深度学习	DeepSentiBank[21]	68.73	71.19	76.35	70.15	71.25	70.11
	PCNN (VGGNet)[20]	70.96	70.84	82.54	76.52	76.36	73.58
	Fine-tuned VGG[36]	70.09	72.48	84.35	82.26	76.75	77.02
	AR[29]	74.80	76.03	88.65	85.10	81.06	81.26
	R-CNNGSR[31]	75.02	75.89				81.36
	VSF[23]	81.62	81.82			83.10	83.17
	MLR[18]	75.63	77.85	89.77	85.72	81.49	83.08
本文模型	CHINet w/o G	78.61	79.07	86.54	83.04	81.36	82.26
	CHINet w/o L	81.12	81.15	89.51	82.95	82.17	82.47
	CHINet G&L	83.27	82.33	90.52	86.43	84.37	83.53

练的CNN，然后，在FI上进行微调。与结合了情感激活区域的模型，包括对象语义情感关联模型（object semantics sentiment correlation model, OSSCM）[22]、弱监督耦合网络（weakly supervised coupled networks, WSCNet）[3]和VSF[23]相比，CHINet取得了最高的识别精度，因为本文模型可以自适应地选择浅层视觉特征或深层语义，以更好的建模层次结构的上下文关联。此外，与基于多级特征融合的MSRCA[30]和MLR[18]相比，CHINet达到超越2.51%的性能提升，同时，即使在不考虑全局与局部学习的统一学习框架下，其表现也能取得近0.9%的准确率提升，证明了本文模型的有效性。总的来说，本文模型可以有效地学习更具辨别力的情感表示，提高分类性能。

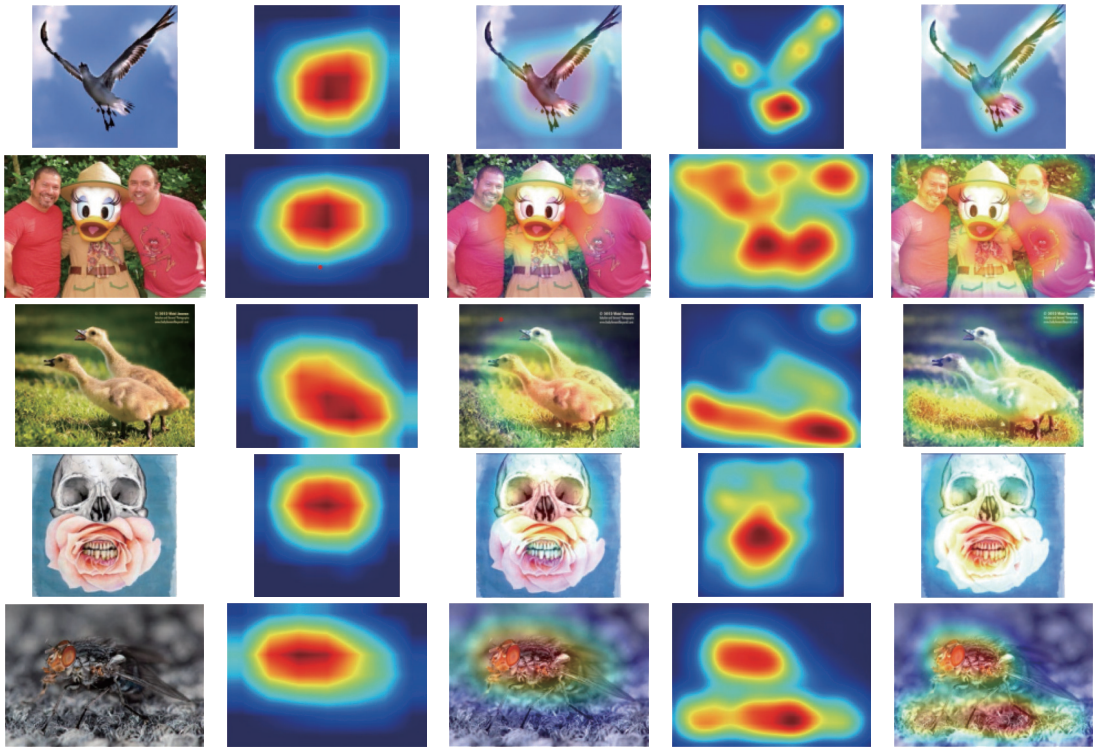
表6 FI数据集上多类别情感分类性能比较

Table 6 Comparison of multi-category emotion classification performance on FI dataset

模型	网络	准确率/%
基线模型	AlexNet[16]	49.54
	VGGNet-19[36]	61.74
	ResNet[34]	53.08
	Inception[38]	60.12
	Inception-Resnet[39]	62.77
对比模型	DeepSentiBank[21]	51.29
	PDANet[40]	69.42
	MKN[19]	63.92
	CycleEmotionGAN[41]	66.79
	Deep metric learning[42]	68.37
	MldrNet[16]	67.24
	OSSCM[22]	69.32
	WSCNet[3]	70.07
	VSF[23]	70.46
	MSRCA[30]	69.05
	MLR[18]	67.49
本文模型	CHINet w/o G	69.94
	CHINet w/o L	70.32
	CHINet G&L	71.56

3.4 可视化分析和讨论

为观察模型的情感推理过程并通过可视化分析模型的潜在缺陷，采用Grad-类激活图（class activation map, CAM）[43]展示微调的ResNet和CHINet对情感图像样本预测的CAM可视化结果。预测结果如图6所示，图6(a)为情感图像样本，图6(b)和图6(d)是ResNet和CHINet在Grad-CAM方法下的预测结果，图6(c)和图6(e)是预测结果与原图的叠加。可以观察到ResNet的激活图更为聚焦，表明其更注重高度集成的语义信息和显著的对象或目标，这类对象具有足够的语义代表性，可以进行初步的情绪预测，然而其无法获得进一步的背景信息。相比之下，本文模型不仅可以捕获具有强烈语义意义的对象，还可以捕获对预测情绪很重要的空间细节，如背景、颜色、纹理细节等。如第2行的结果，ResNet只聚焦于黛西，而CHINet还注意到黛西旁的2个人物的笑脸及人物衣服上的人物动作；对于第4行，本文模型不仅注意到图片上半部分的骷髅，还考虑了下半部分莲花中龇牙咧嘴的怪相。以上可视化的结果表明，CHINet的情感激活图不仅关注特定的对象，而且充分考虑到所有可能潜在的对象，以上可视化分析结果证实了关注层次结构中的上下文关联是有效的。



(a) 情感图像　(b) ResNet CAM　(c) ResNet CAM+情感图像　(d) CHINet CAM　(e) CHINet CAM+情感图像

图 6　用于情感推理的类激活图可视化

Fig. 6　Visualization of class activation map for emotion reasoning

4　结　论

1) 本文提出一种上下文层次交互网络来构建网络中间特征中的上下文关系, 并结合不同的视觉属性进行视觉情感分析。自下而上的分支学习全局的高级语义表示, 而自上而下的分支分离出图像风格以获取浅层情感特征, 最后, 通过全局和局部学习同时优化 2 个局部分类器。

2) 实验细节分析和消融实验结果验证本文模型的有效性。可视化的相关结果进一步表明, 本文模型可以同时捕获显著性对象和潜在情感诱发区域。

未来, 为使本文模型能够适应各种类型的情感图像, 仍需考虑如何构造更加有效且扩展性更好的模块对于不同层次特征的挖掘。同时, 如何更加有效地建模上下文特征关联也是值得考虑的方向。

参考文献（References）

[1] 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]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016, 308-314.

[2] XU L W, WANG Z T, WU B, et al. MDAN: multi-level dependent attention network for visual emotion analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 9469-9478.

[3] 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.

[4] YANG H S, FAN Y Y, LV G Y, et al. Exploiting emotional concepts for image emotion recognition[J]. The Visual Computer, 2023, 39(5): 2177-2190.

[5] YANG J Y, LI J, WANG X M, et al. Stimuli-aware visual emotion analysis[J]. IEEE Transactions on Image Processing, 2021, 30: 7432-7445.

[6] LIANG Y, MAEDA K, OGAWA T, et al. Chain centre loss: a psychology inspired loss function for image sentiment analysis[J]. Neurocomputing, 2022, 495: 118-128.

[7] 张浩, 李海鹏, 彭国琴, 等. 多层次特征融合表征的图像情感识别[J]. 计算机辅助设计与图形学学报, 2023, 35(10): 1566-1576.
ZHANG H, LI H P, PENG G Q, et al. Image emotion recognition via fusion multi-level representations[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(10): 1566-1576(in Chinese).

[8] ZHANG H M, XU M. Multiscale emotion representation learning for affective image recognition[J]. IEEE Transactions on Multimedia, 2022, 25: 2203-2212.

[9] 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. New York: ACM, 2010.

[10] RAO T R, LI X X, XU M. Learning multi-level deep representations for image emotion classification[J]. Neural Processing Letters, 2020, 51(3): 2043-2061.

[11] 詹明. 融合风格特征的抽象图像情感识别[D]. 吉林: 吉林大学, 2023.

ZHAN M. Affective analysis of abstract images using style representation[D]. Jilin : Jilin University, 2023(in Chinese).

[12] 尹朝. 基于内容生成与特征提取的图像情感识别模型研究[J]. 系统仿真技术, 2023, 19(2): 141-147.
YIN C. Research on image emotion recognition model based on feature extraction and content generation[J]. System Simulation Technology, 2023, 19(2): 141-147(in Chinese).

[13] ZHAO S C, YAO X X, YANG J F, et al. Affective image content analysis: two decades review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6729-6751.

[14] PENG K C, CHEN T, SADOVNIK A, et al. A mixed bag of emotions: model, predict, and transfer emotion distributions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 860-868.

[15] CAMPOS V, SALVADOR A, GIRO-I-NIETO X, et al. Diving deep into sentiment: understanding fine-tuned CNNs for visual sentiment prediction[C]//Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia. New York: ACM, 2015.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[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. New York: ACM, 2017: 3595-3601.

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

[19] SHE D Y, SUN M, YANG J F. Learning discriminative sentiment representation from strongly- and weakly supervised CNNs[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15(3s): 1-19.

[20] 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. Palo Alto: AAAI Press, 2015, 29(1): 1-8.

[21] CHEN T, BORTH D, DARRELL T, et al. DeepSentiBank: visual sentiment concept classification with deep convolutional neural networks[EB/OL]. (2014-10-30)[2023-06-01]. https://arxiv.org/abs/1410.8586.

[22] ZHANG J, CHEN M, SUN H, et al. Object semantics sentiment correlation analysis enhanced image sentiment classification[J]. Knowledge-Based Systems, 2020, 191: 105245.

[23] YAMAMOTO T, TAKEUCHI S, NAKAZAWA A. Image emotion recognition using visual and semantic features reflecting emotional and similar objects[J]. IEICE Transactions on Information and Systems, 2021, 104(10): 1691-1701.

[24] YANG J Y, GAO X B, LI L D, et al. SOLVER: scene-object interrelated visual emotion reasoning network[J]. IEEE Transactions on Image Processing, 2021, 30: 8686-8701.

[25] DENG Z L, ZHU Q R, HE P, et al. A saliency detection and gram matrix transform-based convolutional neural network for image emotion classification[J]. Security and Communication Networks, 2021, 2021: 6854586.

[26] 彭国琴. 基于深度学习的图像情感语义分析关键问题研究[D]. 昆明: 云南大学, 2021.
PENG G Q. Research on key issues of image emotional semantic analysis based on deep learning[D]. Kunming: Yunnan University, 2021(in Chinese).

[27] PENG K C, SADOVNIK A, GALLAGHER A, et al. Where do emotions come from? predicting the emotion stimuli map[C]//Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2016: 614-618.

[28] SUN M, YANG J F, WANG K, et al. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2016: 1-6.

[29] YANG J F, SHE D Y, SUN M, et al. Visual sentiment prediction based on automatic discovery of affective regions[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2513-2525.

[30] ZHANG J, LIU X Y, CHEN M, et al. Image sentiment classification via multi-level sentiment region correlation analysis[J]. Neurocomputing, 2022, 469: 221-233.

[31] XIONG H T, LIU Q, SONG S Y, et al. Region-based convolutional neural network using group sparse regularization for image sentiment classification[J]. EURASIP Journal on Image and Video Processing, 2019, 2019(1): 30.

[32] 申朕, 崔超然, 董桂鑫, 等. 基于深度多任务学习的图像美感与情感联合预测研究[J]. 软件学报, 2023, 34(5): 2494-2506.
SHEN Z, CUI C R, DONG G X, et al. Unified image aesthetic and emotional prediction based on deep multi-task learning[J]. Journal of Software, 2023, 34(5): 2494-2506(in Chinese).

[33] ZHANG H, LUO G F, YUE Y Y, et al. Affective image recognition with multi-attribute knowledge in deep neural networks[J]. Multimedia Tools and Applications, 2024, 83(6): 18353-18379.

[34] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE Press, 2016: 770-778.

[35] 张浩, 徐丹. 基于深度学习的少数民族绘画情感分析方法[J]. 中国科学: 信息科学, 2019, 49(2): 204-215.
ZHANG H, XU D. Ethnic painting analysis based on deep learning[J]. Scientia Sinica (Informationis), 2019, 49(2): 204-215 (in Chinese).

[36] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-06-01]. https://arxiv.org/abs/1409.1556.

[37] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2261-2269.

[38] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2818-2826.

[39] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on

learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017, 31(1): 1-7.

[40] ZHAO S C, JIA Z Z, CHEN H, et al. PDANet: polarity-consistent deep attention network for fine-grained visual emotion regression[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019.

[41] SUN Z K, SARMA P, SETHARES W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34(5): 8992-8999.

[42] YAO X X, SHE D Y, ZHANG H W, et al. Adaptive deep metric learning for affective image retrieval and classification[J]. IEEE Transactions on Multimedia, 2020, 23: 1640-1653.

[43] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks *via* gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

Image sentiment analysis by combining contextual correlation

LUO Gaifang[1, 2], ZHANG Hao[1], XU Dan[1, *]

(1. School of Information Science & Engineering, Yunnan University, Kunming 650500, China;
2. School of Software, Shanxi Agricultural University, Jinzhong 030801, China)

Abstract: Image sentiment analysis aims to analyze the emotions conveyed by visual content. A key challenge in this field is to bridge the affective gap between latent visual features and abstract emotions. Existing deep learning models attempt to address this issue by directly learning discriminative high-level emotional representations globally at once but overlook the hierarchical relationship between features at each layer of the deep model, resulting in a lack of correlation between contextual features. Therefore, this paper proposed a context-hierarchical interaction network (CHINet) to model the correlation between contextual information and sentiment within the hierarchy. The model consists of two branches: a bottom-up branch, which first directly learns the global emotional representation at the high-level semantic level; then, for different feature level within the branch, it extracts the style representation and localizes potential emotion activation regions by shallow style encoder and emotion activation attention mechanism. The extracted features are then cascaded into a pyramid structure as top-down branches, modeling contextual hierarchical dependencies and providing shallow visual features for emotion representation. Finally, global and local learning integrate shallow image styles with high-level semantics. Experiments show that the proposed model improves emotion recognition accuracy on the FI dataset compared with related methods, including multi-level feature fusion methods and approaches incorporating local emotional regions.

Keywords: image sentiment analysis; affective gap; contextual correlation; hierarchy structure; emotion classification

Received: 2023-06-12; **Accepted**: 2023-09-08; **Published Online**: 2023-11-07 14:39

URL: link.cnki.net/urlid/11.2625.V.20231106.1919.001

Foundation items: National Natural Science Foundation of China (62162068,62061049); Yunnan Province Ten Thousand Talents Program and Yunling Scholars Special Project (YNWR-YLXZ-2018-022); Yunnan Provincial Department of Education Scientific Research Fund (2025J0008)

*** Corresponding author**. E-mail: danxu@ynu.edu.cn