智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于抽象关系场景图的图像情感识别

康博, 钱艺, 文益民

引用本文:
康博,钱艺,文益民. 基于抽象关系场景图的图像情感识别[J]. 智能系统学报, 2024, 19(2): 335–343.
KANG Bo, QIAN Yi, WEN Yimin. Image sentiment recognition based on the abstract relational scene graph network[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 335–343.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202303009

DOI: 10.11992/tis.202303009
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20231117.1542.004

基于抽象关系场景图的图像情感识别

康博, 钱艺, 文益民
(桂林电子科技大学 广西图像图形与智能处理重点实验室, 广西 桂林 541004)

摘　要: 图像情感识别是通过分析视觉刺激来预测人类情感的抽象过程。现有方法大多缺乏对对象间关系以及对象与场景间相互作用的关注, 并且对象间复杂多样的关系难以得到充分利用, 进而导致难以正确对图像情感进行预测。为解决上述问题, 提出一种基于抽象关系场景图的图像情感识别方法。首先, 构建对象和属性检测器来提取图像中对象及其属性的特征。其次, 使用对象特征推理对象间的亲密度和抽象关系特征, 进而构建抽象关系场景图。再次, 提出抽象关系图卷积网络来推理抽象关系场景图。最后, 设计渐进式注意力机制对多个对象特征进行融合, 以得到图像的整体对象特征。在 FI、EmotionRoI 和 Twitter I 公开数据集上的试验结果表明, 该方法的分类准确率优于现有方法。

关键词: 图像情感识别; 抽象关系; 场景图; 图卷积网络; 注意力机制; 卷积神经网络; 视觉情感分析; 深度学习
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2024)02−0335−09

中文引用格式: 康博, 钱艺, 文益民. 基于抽象关系场景图的图像情感识别 [J]. 智能系统学报, 2024, 19(2): 335–343.
英文引用格式: KANG Bo, QIAN Yi, WEN Yimin. Image sentiment recognition based on the abstract relational scene graph network[J]. CAAI transactions on intelligent systems, 2024, 19(2): 335–343.

Image sentiment recognition based on the abstract relational scene graph network

KANG Bo, QIAN Yi, WEN Yimin
(Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: Image sentiment recognition is an abstract process of forecasting human emotions by analysis of various visual stimuli. Most of the earlier literature does not focus on the relationships among objects and the interactions between objects and scenes, and the complex and diverse relationships among objects are difficult to fully exploit, resulting in difficulty in correctly forecasting image sentiment. To deal with this problem, we develop an abstract relational scene graph network for image sentiment recognition. First, an object and attribute detector is generated to extract object features and their corresponding attribute features from images. Second, the affinities and abstract relationship features among objects are inferred through object features, and then the abstract relational scene graph is generated. Moreover, an abstract relational graph convolutional network is developed for reasoning the abstract relational scene graph. Last, a progressive attention mechanism is designed to fuse multiple object features to acquire the overall object feature of the image. Application on three public datasets, FI, EmotionRoI, and Twitter I, demonstrates that the classification accuracy of the proposed method is better than that of the existing methods.

Keywords: image sentiment recognition; abstract relationship; scene graph; graph convolutional network; attention mechanism; convolutional neural network; visual sentiment analysis; deep learning

收稿日期: 2023−03−04.　**网络出版日期**: 2023−11−20.
基金项目: 国家自然科学基金项目(62366011); 广西重点研发计划项目(桂科 AB21220023); 广西图像图形与智能处理重点实验室项目 (GIIP 2306).
通信作者: 文益民. E-mail: ymwen@guet.edu.cn.

图像情感识别(image sentiment recognition, ISR)是一种通过对图像中的各种视觉对象进行分析, 进而获得图像所传达情感的任务, 该任务是

计算机视觉领域的研究热点和难点。随着社交网络的普及，越来越多的人会选择分享生活图片和旅游照片的方式[1-2]来表达情感。然而，由于大多数图像缺乏对应的情感标签，进行人工标记会造成大量资源的浪费，也为该任务的发展带来巨大挑战。目前该任务面临的最大问题是：图像情感识别是一个主观抽象的过程且高级情感和低级视觉之间存在鸿沟，很难仅使用低级视觉特征对图像做出正确的情感预测[3-6]。图像情感识别性能的提升也有助于图像检索[7]、风格化图像描述[8-10]和观点挖掘[11]等任务的发展。

为解决上述问题，越来越多的研究人员开始关注该领域。早期 ISR[12-16]是通过手工提取特征的方式对图像情感进行识别。Machajdik 等[12]通过手工提取的特征来识别图像情感，包括颜色、纹理、构图和内容 4 种特征。Zhao 等[13]则利用基于艺术原则的中级特征代替基于艺术元素的低级特征来识别图像情感。虽然手工提取特征的方法具有一定的效果，但不具备覆盖重要情感因素的能力。后来，研究人员将卷积神经网络（convolutional neural network, CNN）引入到图像情感识别任务[17-27]中。与手工提取特征的方法相比，尽管这些方法取得了很大的进步，但却忽略了局部区域表达情感的能力。近年来，为了聚焦情感区域，通过结合目标检测与注意力机制提出了一些方法[25-26]，使 ISR 的性能得到进一步提升。Yang 等[25]提出了“情感区域”（affective regions, AR）的概念，并利用 3 种融合策略对情感区域的特征进行融合。Xiong 等[26]利用组稀疏正则化（group sparse regularization, GSR）将低级视觉特征和高级情感特征相结合，对情感区域进行检测。然而，考虑到图像中的不同情感区域可能会传达相反的情感，Yang 等[27]利用对象间的相互作用对图像情感进行预测。

已有方法虽然在图像情感识别任务中取得了不错的效果，但缺乏对对象间关系以及对象与场景间相互作用的考虑。例如一幅图像中有 2 个人，若仅利用对象间的相互作用则无法判定两人的真实状态是处于竞争关系还是争吵关系，需对对象间的关系做进一步探索。当两人存在比赛竞争关系时，会产生兴奋的正向情感；反之，存在争吵关系则会产生负面情感。此外，同类对象在不同场景中会传达出相反的情感。身处花草场景中的小男孩会传达出积极情感；相反，当小男孩在昏暗场景中则会表达出消极情感。基于以上事实，本研究认为对象间关系以及对象与场景间相互作用可以作为情感刺激。

针对上述方法存在的问题，本研究提出了基于抽象关系场景图的图像情感识别方法（abstract relational scene graph network for image sentiment recognition, ARSGN）。首先，通过构建对象和属性检测器从情感图像中提取对象特征及其相应的属性特征。其次，利用对象特征对对象间的亲密度和抽象关系特征进行初步探索，进而构建抽象关系场景图。再次，为了进一步探索对象间的关系，提出抽象关系图卷积网络（abstract relational graph convolutional network, AR-GCN）对抽象关系场景图进行推理，使对象特征具有情感因素。然后，先设计场景特征提取器用于提取图像的场景特征，再提出渐进式注意力机制（progressive attention mechanism, PAM）来融合多个对象特征，以得到图像的整体对象特征。最后，将图像的场景特征和整体对象特征拼接，并送入情感分类器对图像情感进行预测。

1　相关工作

图像情感识别根据不同的心理学模型可分为分类任务和分布式学习任务[28]。由于本研究不涉及分布式学习任务，因此将仅对分类任务进行相关介绍。目前，分类任务根据提取图像特征方式的不同，大致可分为传统方法和深度学习方法两类。

1.1　传统方法

早期的 ISR 方法大多采用手工提取特征的方式来识别图像情感。Machajdik 等[12]通过手工提取颜色、纹理、构图和内容 4 种特征对图像的情感进行识别。Zhao 等[13]根据视觉平衡、和谐和强调等艺术原则来提取中级特征，并用于图像情感的预测。此外，其他研究人员将形容词名词对（adjective noun pairs, ANPs）引入至 ISR 领域。Borth 等[14]筛选出 1 200 个 ANPs 来构成名为 SentiBank 的视觉概念检测器。SentiBank 利用图像对 1 200 个 ANPs 的响应来生成一个 1 200 维的分类向量，进而利用分类向量对图像情感进行识别。Chen 等[15]通过统计图像中频度最高的六类对象，并利用 ANPs 之间的概念相似性建立情感分类模型。Rao 等[16]利用视觉词袋（bags of visual words）对每个图像块（image patch）进行特征提取，得到与情感相关的特征。由于手工提取特征方式的局限性，导致特征包含的噪声较多，并不能很好地弥补低级视觉和高级情感间的鸿沟。

1.2　深度学习方法

随着 CNN 在多个领域取得突破，越来越多的

研究人员倾向于将 CNN 应用到 ISR 领域。基于 Borth 等[14] 工作, Chen 等[17] 利用 CNN 提取图像特征, 并提出一种名为 DeepSentiBank 的视觉情感概念分类器。You 等[18] 利用约 50 万幅带噪声的图像来训练渐进式卷积神经网络(progressive convolutional neural network, PCNN)用于情感图像分类。Rao 等[19] 结合高级图像语义、中级图像美学和低级图像视觉 3 个方面, 提出了多层深度表征网络用于图像情感识别。Zhang 等[20] 提出多层次图像情感识别模型, 该模型包括底层视觉、中层美学和高层语义, 并设计新的损失函数用于解决情感图像数据集中样本不平衡的问题。上述工作虽然取得了一定的效果, 但忽略了局部区域可以表达情感的能力。Yang 等[25]通过计算每一个候选区域的情感分数来选择情感区域, 进而利用情感区域的特征识别情感。Xiong等[26] 利用组稀疏正则化提出一种基于区域的卷积神经网络来自动检测情感区域。

上述方法虽然缩小了低级视觉和高级情感间的鸿沟, 但缺乏对对象间关系以及对象与场景间相互作用的考虑。本研究利用对象间关系以及对象与场景间的相互作用, 提出基于抽象关系场景图的图像情感识别方法。

2　抽象关系场景图网络

本研究提出一种基于抽象关系场景图的图像情感识别方法, 其网络结构如图 1 所示。首先, 构建对象和属性检测器来提取图像中对象及其属性的特征。其次, 利用对象特征对对象间亲密度和抽象关系特征进行推理, 进而构建抽象关系场景图。再次, 提出抽象关系图卷积网络来推理抽象关系场景图, 以使对象特征具有情感因素。然后, 先设计场景特征提取器用于提取图像的场景特征, 再提出渐进式注意力机制, 其利用对象与场景间相互作用来融合多个对象特征, 进而得到图像的整体对象特征。最后, 拼接图像的场景特征和整体对象特征作为图像的情感特征, 并将图像的情感特征送入情感分类器对情感的类别进行预测。

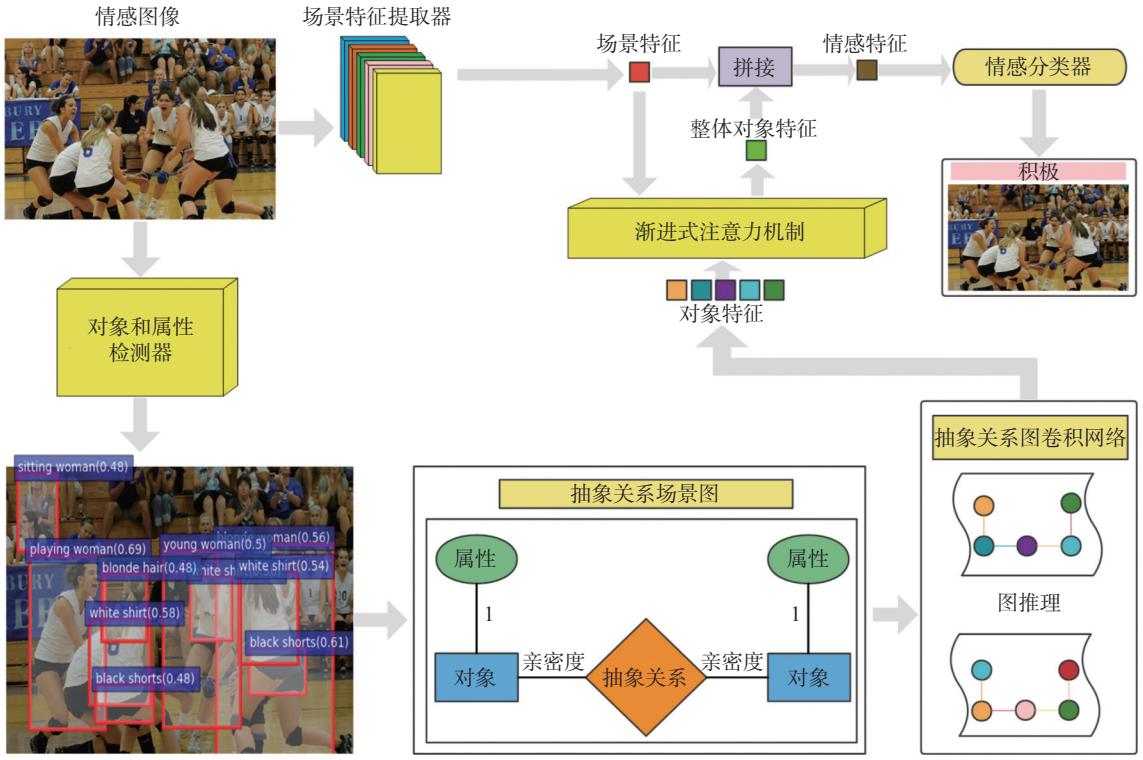

图 1　ARSGN 的网络结构示意

Fig. 1　Architecture of an abstract relational scene graph network for visual sentiment recognition

2.1　对象和属性检测器

为了提取图像中对象及其属性的特征, 本研究构建了一种对象和属性检测器, 如图 2 所示。首先, 使用 Faster R-CNN[29] 提取各对象的特征, 再利用自下而上的注意力机制(bottom-up attention)[30] 对每个对象的类别执行非极大值抑制(non-maximum suppression, NMS)的操作。以此得到图像中各对象特征的集合 $O=\{o_1,o_2,\cdots,o_n\}$, 及其相应的类别标签 $C=\{c_1,c_2,\cdots,c_n\}$ 和置信度 $P=\{p_1,p_2,\cdots,p_n\}$, 其中 $o_i\in\mathbf{R}^{d_1}$, $d_1=2\,048$, $n=10$, 其次, 将各对象的特征按照置信度进行降序重新排列。为了方便后续操作, 仍需将排列后的对象特征集合

与相应的类别标签记为O与C。再次，利用对象类别标签的嵌入和对象特征相加得到属性特征。属性特征的集合记为$A=\{a_1,a_2,\cdots,a_n\}$，其中$a_i\in\mathbf{R}^{d_1}$。然后，把属性特征送入属性分类器中对属性进行分类。最后，额外添加一个多分类交叉熵损失函数完成嵌入层和属性分类器的训练。

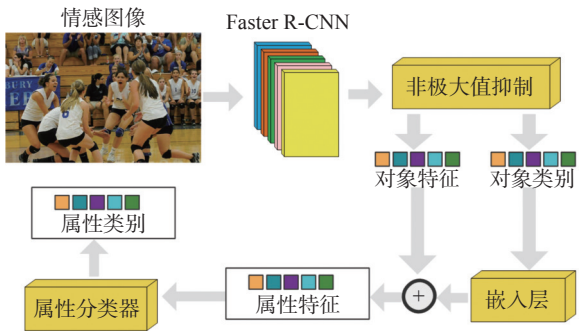

图 2　对象和属性检测器的结构示意

Fig. 2　Architecture of object and attribute detector

对象和属性检测器除了能够检测如狗、人和建筑物等对象外，同时还能对如黄色的、年轻的、高的等属性进行检测。这样做的优势在于：对象特征和属性特征会包含更加丰富的语义信息，并且是后续构建抽象关系场景图的重要前提。此外，利用自下而上的注意力机制能够剔除大量的冗余框，并有利于对包含显著对象的候选框进行选择。

2.2　抽象关系场景图

现有场景图是由对象结点、属性结点和关系结点组成，并需要对对象间关系进行详细标注。由于对象间复杂多样的关系，导致难以训练性能良好的关系分类器。因此，使用对象特征对对象间的亲密度和抽象关系特征进行初步探索，进而构建抽象关系场景图（abstract relational scene graph, ARSG），如图 3 所示。为了进一步探索对象间关系，提出 AR-GCN 对抽象关系场景图进行推理，得到具有情感因素的对象特征。

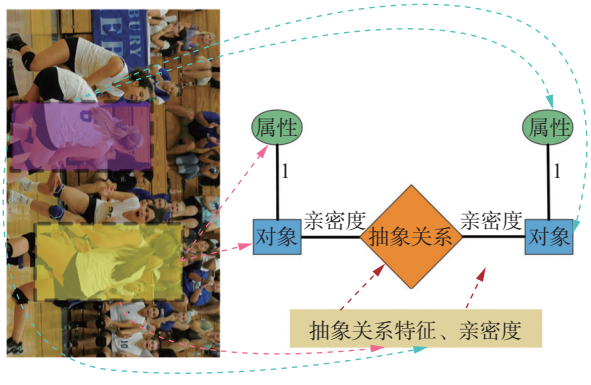

图 3　抽象关系场景图的构建过程

Fig. 3　Construction of abstract relational scene graph

2.2.1　图的构建

本研究将每幅图像的抽象关系场景图定义为$\mathcal{G}=(\mathcal{V},\mathcal{E})$。其中，$\mathcal{V}$和$\mathcal{E}$分别表示点集合与边集合。$\mathcal{V}$包括 3 种结点：对象结点、属性结点和抽象关系结点。首先，将对象特征和属性特征分别作为对象结点和属性结点的值。其次，利用对象特征推理得到对象间的抽象关系特征，并将其作为抽象关系结点的值。具体是将对象特征o_i和o_j相加再进行ℓ_2归一化

$$r_{ij}=\ell_2(o_i+o_j) \tag{1}$$

式中，$r_{ij}\in\mathbf{R}^{d_1}$表示对象$i$和$j$之间的抽象关系特征。最后，由于仅用对象特征无法对对象间关系进行清晰表达，因此需将抽象关系特征r_{ij}从视觉空间投影到情感空间，进一步增强r_{ij}的情感信息

$$F(\cdot)=\ell_2(W_l(\cdot)+b_l) \tag{2}$$

其中，W_l和b_l是可学习的权重和偏置。

$\mathcal{E}$包括两种边，一种存在于对象结点和属性结点之间，另一种存在于对象结点和抽象关系结点之间。本研究先将对象结点和属性结点之间边的权值设为 1。再根据 Li 等[31]提出的方法，计算对象i与j之间的亲密度f_{ij}，并将其作为对象结点与抽象关系结点之间边的权值

$$f_{ij}=\mathrm{sigmoid}(\vartheta(o_i)\times\varphi(o_j)) \tag{3}$$

其中，$\vartheta(o_i)=W_\vartheta o_i$和$\varphi(o_j)=W_\varphi o_j$表示 2 个不同的嵌入函数，且$W_\vartheta$和$W_\varphi$可通过反向传播进行学习。亲密度$f_{ij}$越大，表示对象间关系的强度越强。与现有场景图构建方法不同的是，抽象关系场景图无需依赖对象间关系信息的标注。

2.2.2　图的推理

传统的图卷积网络[32]（graph convolutional network, GCN）定义如下

$$\boldsymbol{X}'=\sigma(\tilde{\boldsymbol{L}}_{\mathrm{sym}}\boldsymbol{X}\boldsymbol{W}_c) \tag{4}$$

$$\tilde{\boldsymbol{L}}_{\mathrm{sym}}=\tilde{\boldsymbol{D}}^{-1/2}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-1/2} \tag{5}$$

$$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I} \tag{6}$$

$$\tilde{\boldsymbol{D}}_{ii}=\sum\nolimits_j\left(\tilde{\boldsymbol{A}}_{ij}\right) \tag{7}$$

式中：$\tilde{\boldsymbol{L}}_{\mathrm{sym}}$为重归一化拉普拉斯矩阵；$\boldsymbol{W}_c$为 GCN 的可学习权重；$\boldsymbol{A}$和$\boldsymbol{D}$分别表示邻接矩阵（adjacency matrix）和度矩阵（degree matrix）。

与传统 GCN 不同的是，本研究在构建抽象关系场景图时缺乏对象间关系信息的标注，因此需要设计新的规则对抽象关系场景图进行推理。由于对象的属性仅被用作描述对象的精确信息且不会随着对象间的相互作用而改变，以及抽象关系特征是由对象特征推理得到，因此每层 AR-GCN 推理时仅对对象结点的值进行更新。AR-GCN 被定义为

$$o_i^{(l+1)} = \sigma(W_g^{(l)} o_i^{(l)} + W_r^{(l)}(a_i + \sum_{j=1}^{n} f_{ij}^{(l)} r_{ij}^{(l)})) \tag{8}$$

式中: σ表示非线性激活函数 ReLU; $W_g^{(l)}$和$W_r^{(l)}$属于 AR-GCN 中的可学习权重。对于第l层, $o_i^{(l)}$表示对象结点的值, $f_{ij}^{(l)}$和$r_{ij}^{(l)}$分别表示对象间的亲密度和抽象关系结点的值。鉴于对象间的相互作用会改变对象间的亲密度和关系, 因此在每层 AR-GCN 推理后, 都会根据式 (1)~ (3) 更新对象间的亲密度$f_{ij}^{(l)}$和抽象关系特征$r_{ij}^{(l)}$。

2.3　渐进式注意力机制

为了融合多个对象特征, 本研究通过对象与场景间的相互作用来设计渐进式注意力机制, 如图 4 所示。

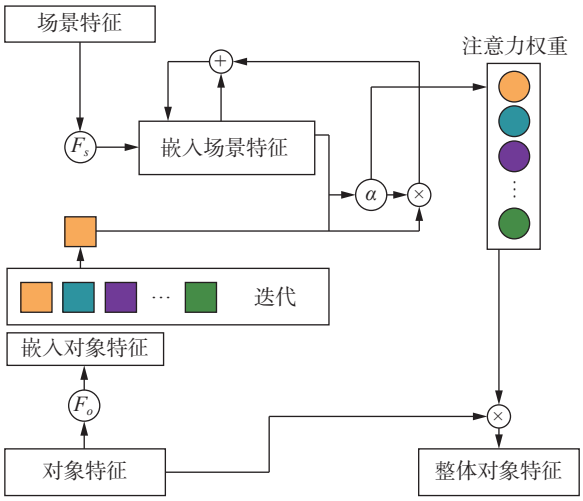

图 4　渐进式注意力机制的结构示意

Fig. 4　Architecture of progressive attention mechanism

首先, 将 ResNet-101 作为场景特征提取器, 提取图像的场景特征f_{sce}, 其中$f_{\text{sce}} \in \mathbf{R}^{d1}$。其次, 为了防止过拟合, 对 AR-GCN 推理后的对象特征做ℓ_2归一化处理, 并利用线性函数F_o和F_s将对象特征o_i和场景特征f_{sce}映射到相同的特征空间, 以缩小对象特征与场景特征间的鸿沟

$$f'_{\text{sce}} = F_s(f_{\text{sce}}) \tag{9}$$

$$O' = F_o(\ell_2(O)) \tag{10}$$

$$F_s(\cdot) = \ell_2(W_s(\cdot)) \tag{11}$$

$$F_o(\cdot) = \ell_2(W_o(\cdot)) \tag{12}$$

式中: W_s和W_o是可学习的权重; f'_{sce}是嵌入场景特征; $O' = \{o'_1, o'_2, \cdots, o'_N\}$是嵌入对象特征的集合。再次, 由于场景是由多个对象依次加入而逐渐形成, 因此需要逐个计算各对象的注意力权重

$$\alpha_i = \text{sigmoid}(f'_{\text{sce}} \times o'_i) \tag{13}$$

在每得到一个对象的注意力权重后, 根据下式对f'_{sce}进行更新, 直至所有对象计算完毕。

$$f'_{\text{sce}} = f'_{\text{sce}} + \alpha_i \times o'_i \tag{14}$$

然后, 利用注意力权重α_i与对象特征o_i按照下式对整体对象特征f_{obj}进行计算

$$f_{\text{obj}} = \sum_{i=1}^{n} \alpha_i \times o_i \tag{15}$$

因为场景和对象都可以被独立作为情感刺激, 所以拼接f_{sce}和f_{obj}作为图像的情感特征f_{emo}

$$f_{\text{emo}} = \text{concate}[f_{\text{sce}}, f_{\text{obj}}] \tag{16}$$

紧接着, 将f_{emo}送入情感分类器

$$p(c_l | f_{\text{emo}}, \boldsymbol{W}) = \frac{\exp(f_{\text{emo}} w_{c_l})}{\sum_{j=1}^{C_L} \exp(f_{\text{emo}} w_j)} \tag{17}$$

式中: c_l表示 ARSGN 预测的情感类别; C_L表示情感类别的个数。$\boldsymbol{W} \in \mathbf{R}^{(d_1+d_1)\times C_L}$表示情感分类器中可学习的权重。最后, 添加一个交叉熵损失函数对整个网络进行训练。

3　试验及结果分析

本节首先对试验所用数据集以及网络实施细节进行介绍, 其次与现有其他图像情感识别方法进行对比, 最后对网络结构和超参数进行试验分析。

3.1　数据集

本研究使用 3 个公开情感图像数据集 FI[33]、Twitter I[18] 和 EmotionRoI[34] 对所提网络进行评估。各个数据集的信息详见表 1, 并在图 5 中给出对应数据集的情感图像示例。

表 1　情感图像数据集的统计数据

Table 1　Statistics of affective images datasets

数据集	积极	消极	总数	类别数
FI	16430	6878	23308	8
Twitter Ⅰ	769	500	1269	2
EmotionRoI	660	1320	1980	6

图 5　来自 FI、Twitter I 和 EmotionRoI 的情感图像示例

Fig. 5　Affective images from FI, Twitter I, and EmotionRoI

FI 是一个规模庞大且标注良好的数据集，共有 23 308 幅图像。该数据集中每张图片都由 5 个 AMT（amazon mechanical turk）对其进行情感标注，且每张图片至少拥有 3 个相同的标注。FI 中的情感被划分为 8 个不同的类别，包括娱乐、敬畏、满足、兴奋、愤怒、厌恶、恐惧和悲伤。Twitter I 中的图片同样由 5 个 AMT 给出情感分类，该数据集划分为积极与消极两类，且仅包含 1269 张情感图像。其根据 AMT 对同一张图片给出相同标注意见的个数，被划分为 3 个子集："Five Agrees"、"At Least Four Agrees"和"At Least Three Agrees"。EmotionRoI 经常被用作图像情感识别的基准，共有 1980 张图片。其被分为 6 个情感类别（喜悦、惊讶、愤怒、厌恶、恐惧和悲伤）。

本研究聚焦图像情感二分类问题，参考 Yang 等[25] 的工作，将多情感类别分为积极和消极两类。其中，FI 中存在 8 种情感类别，将娱乐、敬畏、满足和兴奋作为积极情感，愤怒、厌恶、恐惧和悲伤作为消极情感；EmotionRoI 有 6 种情感类别，喜悦和惊讶被划分为积极情感，愤怒、厌恶、恐惧和悲伤被划分为消极情感；Twitter I 仅含有积极和消极两类，无需对其进行处理。训练集和测试集的划分也参考 Yang 等[25] 的做法，除 EmotionRoI 存在固定的训练集和测试集之外，将 FI 随机分为 80% 的训练集、5% 的验证集和 15% 的测试集，并将 Twitter I 随机分为 80% 的训练集和 20% 的测试集。

3.2　试验设置

本研究以 ResNet-101 作为骨干网络实现 Faster R-CNN[29]，进而构建对象和属性检测器。对象和属性检测器在 Visual Genome[35] 数据集上进行预训练。场景特征提取器在 ImageNet[36] 数据集上预训练。本研究使用深度学习框架 PyTorch 实现 ARSGN，并使用 Adam 优化器对网络参数进行优化。在训练时，设置 batch_size 大小为 16，总迭代次数为 30，权重衰减（weight decay）为 0.001，学习率初始化为 0.000 1，且每 7 次迭代衰减为当前学习率的 0.1 倍。对于 Twitter I 和 EmotionRoI 这 2 个小规模数据集，本研究先用 FI 训练的参数初始化网络，再用 Twitter I 和 EmotionRoI 对网络参数进行微调。

3.3　与其他方法比较

为了验证 ARSGN 的性能，本研究将与其他现有方法进行对比，包括传统方法和深度学习方法，试验结果如表 2 所示。试验结果以图像情感的分类准确率进行评估，分类准确率被定义为情感图像正确分类的数量占情感图像总数量的比例。本研究将与以下传统方法进行对比。Borth 等[14] 和 Zhao 等[15] 利用手工提取得到情感相关特征，这是图像情感识别领域的初步探索。此外，本研究采用 Yang 等[25] 提取的几种底层视觉特征进行试验，包括 GIST、SIFT 和 HOG 等底层视觉特征。本研究还使用情感图像数据集微调参数，对 VGG-16[35] 进行了试验。对于深度学习方法，本研究首先与 Chen 等[17] 提出的 DeepSentiBank 以及 You 等[18]提出的 PCNN 进行比较；其次对比了 2 种聚焦于情感区域的方法，分别是 Yang 等[25] 提出的 AR 和 Xiong 等[26] 提出的 R-CNNGSR；最后与 Zhang 等[20] 提出的多层次情感识别模型进行比较。可以看出表 2 中存在一些缺失数据，原因是 ARSGN 与对比方法缺乏相同的试验结果或者对比方法的源代码未公开。由表 2 可知，ARSGN 在 EmotionRoI 和 FI 数据集的分类准确率分别达到了 83.47% 和 88.21%，并且在 Twitter I 3 个子集的分类准确率分别达到了 89.91%、86.20% 和 82.36%。通过与上述方法进行对比，ARSGN 的分类效果均优于现有方法。通过对其进行分析，这得益于 ARSGN 考虑到对象间关系以及对象与场景间相互作用对情感的影响，而非将对象看作独立个体。

表 2　ARSGN 与已有方法的分类准确率进行比较

Table 2　Classification accuracy of ARSGN compare with other methods　%

方法	Twitter I			Emotion-RoI	FI
	Twitter I 5	Twitter I 4	Twitter I 3		
Gist[25]	65.87	61.47	60.68	60.38	—
SIFT+BoW[25]	63.15	63.71	60.36	65.30	—
SIFT+VLAD[25]	70.29	68.91	67.14	72.15	—
SIFT+FisherVector[25]	71.09	67.29	65.56	70.92	—
HOG+BoW[25]	68.48	61.92	60.99	61.05	—
HOG+VLAD[25]	71.99	67.74	66.43	63.38	—
HOG+FisherVector[25]	76.07	70.34	68.32	65.33	—
SentiBank[14]	71.32	68.28	66.63	66.18	—
PAEF[13]	72.90	69.61	67.92	75.24	—
DeepSentiBank[17]	76.35	70.15	71.25	70.11	61.54
PCNN(VGGNet)[18]	82.54	76.52	76.36	73.58	75.34
VGG-16[37]	83.44	78.67	75.49	72.25	70.64
Fine-tuned VGG-16[37]	84.35	82.26	76.75	77.02	83.05
AR[25]	88.65	85.10	81.06	81.26	86.35
R-CNNGSR[26]	—	—	—	81.36	—
Zhang[20]	89.77	85.72	81.49	83.08	87.87
ARSGN(本研究)	**89.91**	**86.20**	**82.36**	**83.47**	**88.21**

3.4　消融试验

3.4.1　网络结构分析

由表 3 可知, 本研究在 FI、Twitter I 和 EmotionRoI 3 个数据集上进行消融试验, 共包括 5 组。前 2 组试验“Multi-objects”和“Multi-objects+Scene”是本领域深度学习方法的一种基础做法, 其中“Multi-objects”表示多个对象直接累加, “Multi-objects+Scene”表示将对象和场景进行结合。为了验证所提模块的有效性, 本研究设计了第 3、4、5 组试验。ARSGN 主要由抽象关系场景图(ARSG)和渐进式注意力机制(PAM)2 个模块组成, 其中 ARSG 模块包括抽象关系场景图的构建以及推理过程。基于第 2 组试验, 添加 ARSG 模块进行第 3 组试验。从试验结果得出, 通过探索对象间关系能够有效提升图像情感的分类效果。基于第 2 组试验, 添加 PAM 模块进行第 4 组试验。由试验结果可知, 对象与场景间相互作用对图像情感识别性能的提升具有一定贡献。然后, 在第 2 组试验的基础上, 通过引入 ARSG 模块和 PAM 模块得到第 5 组试验。试验结果证明, 通过考虑对象间关系以及对象与场景间相互作用可以提升图像情感识别的分类准确率。综上所述, ARSG 和 PAM2 个模块是相辅相成、不可或缺的, 其结合起来造就了 ARSGN 的有效性。

表 3　ARSGN 网络结构的消融试验

Table 3　Ablation experiment of ARSGN network structure　%

方法	Twitter I			Emotion RoI	FI
	Twitter I 5	Twitter I 4	Twitter I 3		
Multi-objects	87.19	82.79	78.27	78.65	85.64
Multi-objects + Scene	88.10	84.68	80.71	80.57	86.62
Multi-objects + Scene + ARSG	88.89	85.22	80.94	81.68	87.90
Multi-objects + Scene + PAM	89.12	85.75	81.50	81.52	87.39
Multi-objects + Scene + ARSG + PAM	**89.91**	**86.20**	**82.36**	**83.47**	**88.21**

3.4.2　超参数分析

为了准确挖掘对象间的关系, 本研究通过在 FI 数据集上进行试验来确定 AR-GCN 层数的取值, 结果如图 6 所示。试验过程中, 将 AR-GCN 层数的初始值设为 1, 最大值为 5, 共进行 5 组试验。试验结果表明, 当 AR-GCN 的层数取值为 1 时, 网络性能达到最优。传统 GCN 层数的取值在 2~4, 与 AR-GCN 层数的设置存在差异。本研究认为造成这种现象的原因主要可以归为: 传统 GCN 建立多层的目的在于完成各结点之间的信息交互, 而抽象关系结点本身就包含对象间的信息传播和交互, 因此单层 AR-GCN 便可以完成对象间的信息交互。

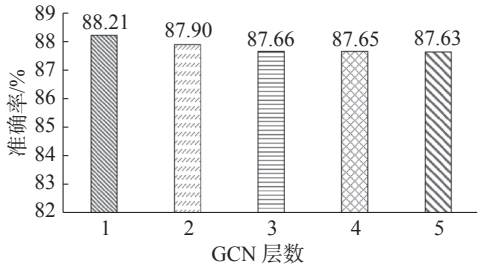

图 6　在 FI 数据集上关于 AR-GCN 层数的消融试验

Fig. 6　Ablation experiment of AR-GCN layers on FI dataset

4　结束语

本研究提出基于抽象关系场景图的图像情感识别方法。该方法在识别图像情感过程中考虑了对象间的关系以及对象与场景间的相互作用, 显著提升了图像情感识别的分类效果。首先, 为了初步探索对象间关系, 本研究利用对象特征对对象间的亲密度和抽象关系特征进行推理, 并构建抽象关系场景图。其次, 提出 AR-GCN 来推理抽象关系场景图, 对对象间关系做进一步探索。最后, 通过对象与场景间的相互作用设计渐进式注意力机制, 将多个对象特征融合进而形成图像的整体对象特征。试验结果表明, 本研究方法能够有效缩小低级视觉和高级情感间的鸿沟, 且在 3 个公开数据集上的分类准确率均优于多个现有算法。未来, 将对具体视觉和抽象关系相结合的方法做进一步探索, 以提高图像情感识别的分类效果。

参考文献:

[1]　ZHAO Sicheng, GAO Yue, DING Guiguang, et al. Real-time multimedia social event detection in microblog[J]. IEEE transactions on cybernetics, 2018, 48(11): 3218–3231.

[2]　吴佩谕, 黄远水. 旅游照片的符号属性对旅游意向的影响研究: 以微信朋友圈旅游照片为例 [J]. 资源开发与市场, 2019, 35(7): 993–1000.
WU Peiyu, HUANG Yuanshui. Study on influence of symbolic attributes of travel photos on travel intention—taking travel photos of WeChat friends circle as an example[J]. Resource development & market, 2019, 35(7):

993–1000.

[3] ZHAO Sicheng, YAO Xingxu, YANG Jufeng, 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.

[4] 赵思成, 姚鸿勋. 图像情感计算综述 [J]. 智能计算机与应用, 2017, 7(1): 1–5.
ZHAO Sicheng, YAO Hongxun. A survey of image emotion computing[J]. Intelligent computer and applications, 2017, 7(1): 1–5.

[5] 王仁武, 孟现茹. 图片情感分析研究综述 [J]. 图书情报知识, 2020(3): 119–127.
WANG Renwu, MENG Xianru. Review of image sentiment analysis[J]. Documentation, information & knowledge, 2020(3): 119–127.

[6] 姚鸿勋, 邓伟洪, 刘洪海, 等. 情感计算与理解研究发展概述 [J]. 中国图象图形学报, 2022, 27(6): 2008–2035.
YAO Hongxun, DENG Weihong, LIU Honghai, et al. An overview of research development of affective computing and understanding[J]. Journal of image and graphics, 2022, 27(6): 2008–2035.

[7] PANG Lei, ZHU Shiai, NGO C W. Deep multimodal learning for affective analysis and retrieval[J]. IEEE transactions on multimedia, 2015, 17(11): 2008–2020.

[8] GUO Longteng, LIU Jing, YAO Peng, et al. MSCap: multi-style image captioning with unpaired stylized text[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 4199−4208.

[9] ZHAO Wentian, WU Xinxiao, ZHANG Xiaoxun. MemCap: memorizing style knowledge for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12984−12992.

[10] FENG Junlong, ZHAO Jianping. Improving stylized caption compatibility with image content by integrating region context[J]. Neural computing and applications, 2022, 34(6): 4151–4163.

[11] LI Zuhe, FAN Yangyu, JIANG Bin, et al. A survey on sentiment analysis and opinion mining for social multimedia[J]. Multimedia tools and applications, 2019, 78(6): 6939–6967.

[12] 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: 83−92.

[13] ZHAO Sicheng, GAO Yue, JIANG Xiaolei, et al. Exploring principles-of-art features for image emotion recognition[C]//Proceedings of the 22nd ACM iNternational Conference on Multimedia. New York: ACM, 2014: 47−56.

[14] BORTH D, JI Rongrong, CHEN Tao, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C]//Proceedings of the 21st ACM International Conference on Multimedia. New York: ACM, 2013: 223−232.

[15] CHEN Tao, YU F X, CHEN Jiawei, et al. Object-based visual sentiment concept analysis and application[C]//Proceedings of the 22 nd ACM International Conference on Multimedia. New York: ACM, 2014: 367−376.

[16] RAO Tianrong, XU Min, LIU Huiying, et al. Multi-scale blocks based image emotion classification using multiple instance learning[C]//2016 IEEE International Conference on Image Processing. Phoenix: IEEE, 2016: 634−638.

[17] CHEN Tao, BORTH D, DARRELL T, et al. DeepSentiBank: visual sentiment concept classification with deep convolutional neural networks[EB/OL]. (2014−10−30)[2022−01−01]. https://arxiv.org/abs/1410.8586.pdf.

[18] YOU Quanzeng, LUO Jiebo, JIN Hailin, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Austin: AAAI, 2015: 381−388.

[19] RAO Tianrong, LI Xiaoxu, XU Min. Learning multi-level deep representations for image emotion classification[J]. Neural processing letters, 2020, 51(3): 2043–2061.

[20] ZHANG Hao, XU Dan, LUO Gaifang, et al. Learning multi-level representations for affective image recognition[J]. Neural computing and applications, 2022, 34(16): 14107–14120.

[21] 李志义, 许洪凯, 段斌. 基于深度学习 CNN 模型的图像情感特征抽取研究 [J]. 图书情报工作, 2019, 63(11): 96–107.
LI Zhiyi, XU Hongkai, DUAN Bin. Research on image emotion feature extraction based on deep learning CNN model[J]. Library and information service, 2019, 63(11): 96–107.

[22] 蔡国永, 贺歆灏, 储阳阳. 基于空间注意力和卷积神经网络的视觉情感分析 [J]. 山东大学学报(工学版), 2020, 50(4): 8–13.
CAI Guoyong, HE Xinhao, CHU Yangyang. Visual sentiment analysis based on spatial attention mechanism and convolutional neural network[J]. Journal of Shandong University (engineering science edition), 2020, 50(4): 8–13.

[23] 白茹意, 郭小英, 贾春花. 基于特征融合的小样本抽象画图像情感预测 [J]. 计算机应用, 2020, 40(8): 2207–2213.
BAI Ruyi, GUO Xiaoying, JIA Chunhua. Sentiment prediction of small sample abstract painting image based on feature fusion[J]. Journal of computer applications, 2020,

40(8): 2207–2213.

[24] 蔡国永, 储阳阳. 基于双注意力多层特征融合的视觉情感分析 [J]. 计算机工程, 2021, 47(9): 227–234.
CAI Guoyong, CHU Yangyang. Visual sentiment analysis based on multi-level features fusion of dual attention[J]. Computer engineering, 2021, 47(9): 227–234.

[25] YANG Jufeng, SHE Dongyu, SUN Ming, et al. Visual sentiment prediction based on automatic discovery of affective regions[J]. IEEE transactions on multimedia, 2018, 20(9): 2513–2525.

[26] XIONG Haitao, LIU Qing, SONG Shaoyi, 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): 1–9.

[27] YANG Jingyuan, GAO Xinbo, LI Leida, et al. SOLVER: scene-object interrelated visual emotion reasoning network[J]. IEEE transactions on image processing, 2021, 30: 8686–8701.

[28] ZHAO Sicheng, DING Guiguang, GAO Yue, et al. Discrete probability distribution prediction of image emotions with shared sparse learning[J]. IEEE transactions on affective computing, 2020, 11(4): 574–587.

[29] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137–1149.

[30] ANDERSON P, HE Xiaodong, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6077−6086.

[31] LI Kunpeng, ZHANG Yulun, LI Kai, et al. Visual semantic reasoning for image-text matching[C]//IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2020: 4653−4661.

[32] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2016−11−03)[2022−01−01]. https://arxiv.org/abs/1609.02907.pdf.

[33] YOU Quanzeng, LUO Jiebo, JIN Hailin, et al. Building a large scale dataset for image emotion recognition: the fine print and the benchmark[J]. Proceedings of the AAAI conference on artificial intelligence, 2016, 30(1): 308–314.

[34] PENG Kuanchuan, SADOVNIK A, GALLAGHER A, et al. Where do emotions come from? Predicting the emotion stimuli map[C]//2016 IEEE International Conference on Image Processing. Phoenix: IEEE, 2016: 614−618.

[35] KRISHNA R, ZHU Yuke, GROTH O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations[J]. International journal of computer vision, 2017, 123(1): 32–73.

[36] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248−255.

[37] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014−12−23)[2022−01−01]. https://arxiv.org/abs/1409.1556.pdf.

作者简介:

康博, 硕士研究生, 主要研究方向为计算机视觉和视觉情感分析。E-mail: 1981480003@qq.com。

钱艺, 博士研究生, 主要研究方向为计算机视觉与零样本学习。E-mail: qyizos@163.com。

文益民, 教授, 博士生导师, 博士, 中国计算机学会 杰出会员, 主要研究方向为机器学习、推荐系统和大数据分析。主持国家自然科学基金项目 3 项, 发表学术论文 50 余篇。E-mail: ymwen@guet.edu.cn。