

基于语义概念的图像情感分析

杨瀚森¹, 樊养余¹, 吕国云¹, 刘诗雅², 郭哲¹

(1.西北工业大学 电子信息学院, 陕西 西安 710072; 2.虚拟现实内容制作中心, 北京 101318)

摘要:随着越来越多的用户通过社交媒体表达自己的情感,图像情感分析技术受到了研究人员的密切关注。但是由于情感的模糊性和主观性,相比较于其他计算机视觉任务,图像情感分析更具挑战性。该领域既有的工作仅研究了图像到情感之间的直接映射关系。然而,心理学中有关情感感知的理论揭示了人们感知情感的过程是分步式的。因此,提出了一种新的图像情感分析框架,利用情感概念作为中级语义来辅助建立图像和情感之间的关系。将情感和概念的关系用知识图谱来描述并嵌入到语义空间中,再将图像的视觉特征投影至该语义空间与情感进行对齐,从而学习图像和情感之间的关系。另一方面,提出了一种多层次深度度量学习方法,从标记层面以及示例层面同时对模型进行优化。在 2 个情感图像数据集上进行实验,结果表明提出的方法在情感图像检索以及分类任务上,相对于现有方法表现良好。

关键词:图像情感分析;知识图谱;视觉-语义嵌入;深度度量学习

中图分类号:TP391

文献标志码:A

文章编号:1000-2758(2023)04-0784-10

随着互联网技术的快速发展和应用,越来越多的人在社交媒体上表达意见,分享情感和日常生活。大量的文字、视觉内容在不同用户之间传播。相比于文字,视觉内容包含更多的抽象信息,针对图像的分析引起了研究者的密切关注。当下,图像情感分析在舆情监控、意见挖掘^[1]方面发挥了重要作用。此外,该领域开发出的技术存在很多潜在的应用,比如,推荐^[2]、娱乐^[3]、行为评估等。

图像情感分析的目的是预测图像的情感类别,或获得与查询图像具有相同情感的图像。与基于内容的图像分析任务相比,情感的主观性与抽象性使得图像特征和情感之间存在“语义鸿沟”。早期,研究者利用或设计不同的底层特征以学习它们与人类情感之间的映射。Olkiewicz 等^[4]使用边缘、颜色、纹理直方图等人工特征的组合进行情感图像检索。Zhang 等^[5]提取图像的颜色、形状、纹理特征,利用多核学习进行情感图像分类与检索。受到心理学及艺术理论的启发,Zhao 等^[6]基于艺术方面的理论,

提取图像平衡、梯度、运动等特征,并结合底层以及高层特征,进行了图像到情感的多图学习。随着卷积神经网络在物体检测、图像分割、图像识别等领域的成功应用,深度卷积神经网络也逐步被引入图像情感分析领域^[7-8]。近年来,研究人员在卷积神经网络的基础上,提出了各种基于情感的图像检索方法。Yang 等^[9]利用情感之间的层级关系,提出一种具有层级结构的三元组损失函数,实现了情感图像的检索与分类。Yao 等^[10]在 Yang^[9]提出的损失函数的基础上,用自适应间隔代替三元组中的固定间隔,增强了检索与分类性能。Yao 等^[11]利用情感层级关系,设计了一种带有层级结构的 N-Pair Loss 损失函数,对情感图像进行检索和分类,并且在此基础上提出一种自适应的难分负样本生成方法^[12],进一步提升了模型性能。

目前,图像情感分析领域的方法都是建立图像特征和情感之间的直接映射关系。然而,图像情感分析包含了人们感知过程的主观性和抽象性,直接

收稿日期:2022-09-20

基金项目:国家自然科学基金面上项目(62071384)与陕西省重点研发计划(2023-YBGY-239)资助

作者简介:杨瀚森(1986—),西北工业大学博士研究生,主要从事机器学习、图像识别研究。

通信作者:吕国云(1975—),西北工业大学副教授,主要从事音频、视频、图像处理及虚拟现实研究。

e-mail:lvguoyun102@163.com

的映射关系往往难以准确建立。

除计算机视觉领域,图像对情感的影响也得到了心理学家的关注。心理学研究^[13]表明,人类感知情感的过程分为 3 步:刺激(stimuli,S)、机理(organism,O),反应(response,R),即 S-O-R 模型。然而现有的图像情感分析方法可以看作是直接建立 S-O-R 中 S 和 R 的关系,而包含人的感知知识以及经验的机理(O)步骤被忽略了。此外,用独热(one-hot)离散向量来表示情感标记,外部的知识无法融入模型的训练过程。所以,通常的图像情感分析方法会受到模糊性和主观性的影响,而人的情感认知却较少受到影响。

受到情感认知过程的启发,本研究尝试在图像到情感的学习过程中加入类似于人们认知的中间特征。

一类图像情感概念为本研究提供了思路。这些概念通过名为 DeepSentibank^[14] 的图像概念检测器得到。给定一张图片,DeepSentibank 能够输出多个跟情感有关的语义概念,比如“beautiful flower”或“smiling girl”等。这些概念能够很好地描述图像内容,同时具有明显的情感倾向性,因此可以很好地作为图像和情感之间的中间语义。

综合以上事实,本研究提出了一种新的图像情感分析方法,利用情感概念作为中间语义来连接图

像与情感之间的关系,旨在克服图像情感分析中的模糊性和主观性问题。首先,本研究建立了一个情感专属知识图谱来描述概念和情感之间的关系。然后将知识图谱中的节点嵌入到一个共同的语义空间中。模型的训练采用端到端的视觉-语义嵌入方式,将图像特征投影至语义空间与情感进行匹配。同时,情感图像在语义上具有较大的差异性,本研究还提出了一个多级可变间隔损失函数,从而有效地学习图像特征和情感之间的关系。在该损失函数中,间隔大小可以根据相应的概念相似度进行自动调节。实验表明,本文提出的方法具有较好的分类及检索性能,模型输出特征具有较高的可区分性。

1 方 法

图 1 展示了本研究所提出方法的框架:①训练图像经过 DeepSentibank 检测器得到一系列情感概念;②利用情感概念以及情感构建知识图谱;③用图嵌入方法将知识图谱嵌入语义空间;④利用卷积神经网络提取图像特征,将图像特征投影至语义空间,通过多级相似度损失函数优化图像特征和情感特征之间的相似度。测试时,计算图像投影特征与不同情感向量之间的相似度,将相似度最大的情感作为预测结果。

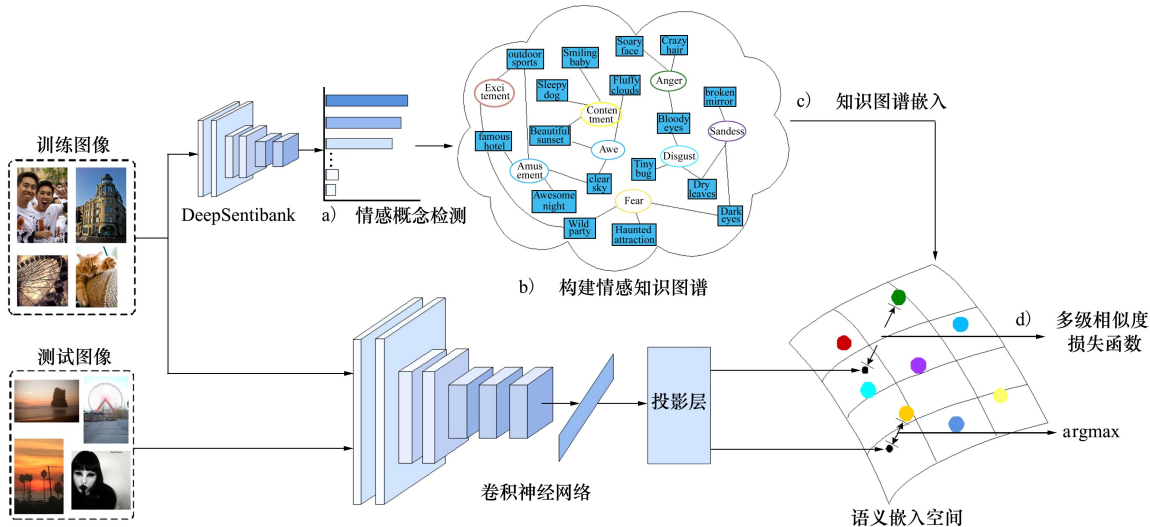


图 1 系统框架图

1.1 DeepSentibank 简介

本研究中的情感语义概念通过视觉概念检测器 DeepSentibank 得到。基于在 ImageNet 数据集上预

训练的深度模型,Chen 等^[14]在 VSO^[15]数据集上对深度模型进行微调。VSO 是一个通过基于情感词语检索和标记分析建立的情感本体数据集,该数据

集包含 3 000 个情感概念标记和 100 多万张图像。DeepSentibank 的最后一个全连接层有 2 089 个神经元,每个神经元对应一个情感概念。输入一张图片,DeepSentibank 能够输出 2 089 个概念的概率分布。图 2 给出了情感概念的示例。当人们看到图 2a) 所示的图片时,不禁对历史遗迹发出赞叹。所以,惊叹是与这张图片最匹配的情感响应。当人们看到图 2b) 这样的场景时,相比较于其他情感,恐惧可能是最为接近人们反应的情感。很明显,除了图像的视觉信息,人们经验当中的先验知识会在很大程度上影响情感认知。

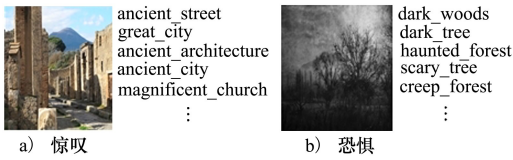


图2 情感概念示例

1.2 问题描述

假设有 M 张训练图片 $\mathbf{X} = \{(x_i, y_i^j)\} (i = 1, 2, \dots, M, j = 1, 2, \dots, C)$, 这里 x_i 表示第 i 张图片, 并且每张图片带有一个标记 y_i^j , $\mathbf{Y} = \{y^1, y^2, \dots, y^C\}$, 数据集共有 C 类标记。视觉语义嵌入模型通过最大化视觉向量和标记向量之间的相似度, 建立从输入到输出的映射 $F: \mathbf{X} \rightarrow \mathbf{Y}$ 。给定一张图片 x 和相应的类别标记 y 在语义空间中的表示 $\delta(y)$, 模型的学习过程使如下先验概率最大化:

$$F(x, y) = \underset{y \in \mathbf{Y}}{\operatorname{argmax}} P(\delta(y) | x, \theta) \quad (1)$$

θ 是视觉语义嵌入模型中需要学习的参数。

1.3 图像表示

预训练的深度学习神经网络模型可以作为良好的图像表示。在 ResNet50^[16] 的基础上建立视觉语义嵌入模型, 并在训练集上对其进行微调。为了消除 ResNet50 最后一个全连接层与图嵌入空间中节点维数的差异, 用另一个全连接层替换 ResNet50 本身的最后一个全连接层。基于其他深度模型的性能将会在之后进行讨论。

1.4 知识图谱构建与表示

为了将外部知识融入到模型的训练过程中, 首先建立一个情感专属的知识图谱。在该知识图谱中, 情感和概念之间的相互关系得以描述。知识图谱作为一种实体的组织方式, 能够将数据以图的形式保存和表示, 并且保留各实体之间的结构化信息。

一个图可以表示为 $G = (V, E)$, V 表示节点, E 表示边。节点之间的关系通过边来表示。在本研究构建的情感专属知识图谱中, 有 C 个情感类别和 S 个概念, 所以一共有 $C + S$ 个节点。边 $e = (V_M, V_P) \in E$ 表示情感节点 (V_M) 和概念节点 (V_P) 之间的连接权重, 边 $e = (V_P, V_P) \in E$ 表示 2 个概念节点之间的连接权重。首先用检测器对训练集中的每个图片进行检测。因为检测概率较低的概念不能够正确反映图像中的内容, 所以对每张图像保留前 k 个检测概念。于是, 可以通过统计概念和情感之间的共现次数, 以及概念和概念之间的共现次数, 并且进行归一化, 得到两类边的权值。如图 2a) 所示, ancient street 和惊叹以及 ancient city 分别共现一次。任意 2 个情感节点之间没有关联。

在构建知识图谱后, 利用 Node2vec^[17] 将图谱中的所有实体以及图关系嵌入到一个低维的语义空间中, Node2vec 是一种基于 Skip-gram 的图嵌入方法, 采用一种有偏的随机游走采样方式, 通过参数设置来控制搜索策略, 从而有效地平衡了学习过程中的同质性和同构性。并且这种游走方式可以挖掘出结构相似性等序列本身没有的信息。同质性和同构性较强的节点在该语义空间中距离较近, 而相似度较低的语义节点距离则较远。

给定当前节点 v , 访问下一个节点 u 的概率为

$$P(c_i = u | c_{i-1} = v) = \begin{cases} \frac{\pi_{vu}}{Z}, & (v, u) \in E \\ 0, & \text{其他} \end{cases} \quad (2)$$

式中: c_i 表示游走中的第 i 个节点; Z 是归一化常数; π_{vu} 是节点 v 和节点 u 之间的未归一化转移概率。为平衡同质性和同构性, Node2vec 引入 2 个超参数 p 和 q 来控制随机游走产生的方式, 并且定义未归一化转移概率为: $\pi_{vu} = \alpha_{pq}(t, x) w_{vu}, w_{uv}$ 为节点 u 和节点 v 之间边的权重, 其中

$$\alpha_{pq}(t, u) = \begin{cases} \frac{1}{p}, & d_{tu} = 0 \\ 1, & d_{tu} = 1 \\ \frac{1}{q}, & d_{tu} = 2 \end{cases} \quad (3)$$

式中: d_{tu} 为节点 t 和 u 之间的最短路径距离。

经过以上转移概率的预处理, 以及有偏置的随机游走和异步随机梯度下降优化, Node2vec 可以学习得到节点的向量表示。

1.5 模型训练

在构建了知识图谱以及完成图嵌入后,通过多级损失函数来训练视觉-语义嵌入模型。

图像特征由深度神经网络 ResNet50 的倒数第2层表示,情感标记通过 Node2vec 表示。全连接层 $T \in \mathbf{R}^{d \times e}$ 将图像特征投影至语义空间,使得其维数与情感语义节点维数相同,其中 d 为图像特征维数, e 为情感语义特征维数。

在视觉语义嵌入的研究中,点积相似度是最常用的衡量2个向量相似度的指标。为了在语义空间中将视觉特征与情感相匹配,通常的做法是将负点积相似度^[18]作为损失函数来优化模型,其形式为

$$L_{\text{dot}} = \sum_{i=1}^M (1 - a_i T \cdot \delta_s(y)) \quad (4)$$

式中: a_i 表示第 i 个训练图片的特征; $\delta_s(y)$ 表示与 a_i 相应的情感标记特征。

然而,只对2个向量之间的差异进行惩罚,这样的约束力较弱。因此设计一个三元组损失函数,实现视觉-情感相似度约束,在减小图像特征和情感之间差异的同时,使得图像特征远离其他不同类别的情感向量。损失函数为

$$L_1 = \sum_{i=1}^M [m_1 + D(a_i T, \delta_s(y)) - D(a_i T, \delta_d(y))]_+ \quad (5)$$

式中: $D(\cdot, \cdot)$ 表示2个向量之间的负点积相似度; $[\cdot]_+ = \max(0, \cdot)$ 。 $\delta_d(y)$ 表示与 x_i 不同的情感向量。 m_1 为固定间隔。

进一步,为了优化视觉向量之间的分布,使用另一个三元组损失函数进行视觉相似度约束

$$L_2 = \sum_{i=1}^M [m_2 + D(a_i T, p_i T) - D(a_i T, n_i T, 0)]_+ \quad (6)$$

式中: p_i 表示与 x_i 类别相同的正示例特征; n_i 表示与 x_i 类别不同的负示例特征。然而,情感图像的类内差异可能会很大,同时,类间差异会很小。在通常的三元组损失函数中,间隔对于任何样本对都是固定的,忽略了上述差异性,故而不能很好地适用于图像情感问题。因此,利用情感概念设计了一种自适应可变间隔,记为 $m_2 = B_{vi} \cdot B_{ni}$ 。 B_{vi} 表示图像 x_i 对应的检测概率最高的概念, B_{ni} 表示图像 n_i 对应的检测概率最高的概念。

最终的损失函数记为:

$$L = \lambda L_1 + (1 - \lambda) L_2 \quad (7)$$

式中: $0 < \lambda < 1$ 为 L_1 和 L_2 之间的平衡参数,将在第3节进行讨论。

给定一张测试图片 x ,通过计算投影视觉向量和每一个情感向量之间的相似度,并通过最邻近搜索法,预测测试图片所属的情感类别

$$y = \underset{y \in Y}{\operatorname{argmax}} (xT \cdot \delta(y)) \quad (8)$$

2 实验设置

2.1 数据集

本文基于2个情感图像数据集构建情感知识图谱,并且评估所提出方法的性能。

You 等^[19]在 Flickr 和 Instagram 上检索 Mikel^[20]所提出的8种情感对应的词语(即:兴奋、娱乐、满足、惊叹、愤怒、厌恶、悲伤、恐惧),建立了大规模的情感图像数据集(Flickr&Instagram,即 FI)。通过检索,得到了超过300万张带有弱标记的图片,其中的90 000张图片由 Amazon Mechanical Turk (AMT)进行进一步标注。为了得到更为可靠的图片标记,5名具有资质的志愿者对每张图片检索时使用的情感进行确认,给出“是”或“否”的选择。最终,23 308张得到了3个以上“是”的图片得以保留。FI数据集提供了每张图片的下载链接,由于一些链接失效,最终下载了22 097张图片(其中,娱乐4 770张,满足5 196张,兴奋2 753张,惊叹2 917张,厌恶1 643张,恐惧957张,悲伤2 672张,愤怒1 189张)。Peng 等^[21]基于 Ekman^[22]所提出的6种情感(惊奇、喜悦、厌恶、恐惧、悲伤、愤怒)及其近义词,在 Flickr 上进行图像检索,得到1 980张图片(每类情感330张)。Peng 等^[21]雇佣 AMT 的工作人员为每张图片的情感进行投票,通过计数,生成了分布式标记。本研究中,仅使用图片的原始类别标记。

2.2 图像检索评价指标

给定一张待检索图片,检索目标是返回与该图片情感相同的图片。

在本研究中,参考文献[23]中的指标进行检索,使用测试集中的图片作为被检索图片,返回结果来自于训练集。2张图片的情感相似性通过点积相似度来衡量。最邻近率(nearest neighbor rate, NN)定义为第一个返回的检索图像与被检索图像属于同一类的比例,用 R_{NN} 表示。

第一层级 (first tier, FT) 和第二层级 (second tier, ST) 表示前 τ 个和前 2τ 个检索结果的召回率 F_{FT} 和 F_{ST} 。

$$F_{\text{FT}} = n_{\tau} / \tau$$

(9)

$$F_{\text{ST}} = n_{2\tau} / 2\tau$$

(10)

式中, n_{τ} 和 $n_{2\tau}$ 分别表示前 τ 个和前 2τ 个检索结果中正确的数量。

$$P_{\text{mA}}$$
 为平均检索准确率在多个类别下的均值

$$P_{\text{mA}} = \frac{1}{C} \sum_{j=1}^C P_{\text{A}}(j)$$

(11)

式中, $P_{\text{A}}(j)$ 为第 j 类的平均精度。

折扣累计增益 (discounted cumulative gain, DCG) 是根据位置先后顺序, 对累计增益 (cumulative gain, CG) 的加权, 用 G_{DC} 表示。

平均归一化修正检索秩 (average normalized modified retrieval rank, ANMRR) 是一种基于正确图片在检索结果中的排序位置的度量指标, 其定义为^[24]:

$$R_{\text{ANMR}} = \frac{1}{Q} \sum_{q=1}^Q \frac{R_{\text{MR}}(q)}{K - 0.5(n_{\text{g}}(q) + 1)}$$

(12)

式中: R_{MR} (modified retrieval rank, MRR) 为修正检索秩; K 为常数; q 为查询图像编号; $n_{\text{g}}(q)$ 代表与查询图像类别相同的被检索图像集大小; R_{ANMR} 数值越小, 代表性能越好。所有检索指标的数值范围都在 0 和 1 之间。

2.3 实验细节

本研究中, 图片的尺寸均调整为 224×224 像素。图像的视觉特征从 ResNet50 提取, 同时也研究基于其他常用深度神经网络模型时的性能。深度模型的初始参数采用在 ImageNet^[25] 上预训练的参数, 并通过本研究提出的方法进行微调。

经过大量观察和研究, DeepSentibank 的检测结果中, 前 5 个响应概率最高的情感概念能够较好地描述图像内容。因此, 每张图片都对应 5 个情感概念。2 个情感图像数据集均随机分为 80% 训练, 5% 验证, 15% 测试。模型的优化采取随机梯度下降方法, mini-batch 的大小为 64, 学习率设置为 0.001。三元组的选取采用半困难 (semi-hard) 采样策略。固定间隔 m_1 设置为 0.1。Node2vec 节点表示向量为 128 维。

3 实验结果与分析

3.1 消融试验

3.1.1 知识图谱构建方法

知识图谱的构建对于准确反映其中节点之间的关系起到重要作用, 本小节讨论不同的知识图谱构建方法对分类和检索性能的影响。

本研究构建的知识图谱中存在 2 种节点: 情感节点及概念节点。首先, 边的权重通过节点之间的共现频次进行归一化来计算。这种方法对应表 1 中的方法 I。

表 1 不同知识图谱构建方法对检索与分类性能的影响

数据集	损失函数	方法	准确率/ %	检索性能					
				mAP ↑	FT ↑	ST ↑	NN ↑	DCG ↑	ANMRR ↓
FI	L_1	I	64.98	0.651 7	0.660 4	0.700 6	0.661 5	0.798 6	0.383 7
		II	65.65	0.686 2	0.686 7	0.733 0	0.663 6	0.803 2	0.288 3
		III	66.29	0.697 4	0.695 7	0.751 8	0.681 1	0.811 2	0.256 4
	L_{dot}	III	65.41	0.654 4	0.677 0	0.738 1	0.662 2	0.796 5	0.375 1
Emotion6	L_1	I	56.11	0.488 5	0.593 5	0.605 8	0.488 4	0.784 5	0.420 8
		III	56.60	0.506 1	0.606 8	0.622 5	0.493 1	0.789 0	0.379 6
	L_{dot}	III	55.26	0.487 2	0.584 4	0.607 6	0.476 3	0.779 4	0.510 4

注: 表 1~4 中黑体数字表示该列一个数据集下最好的性能, “↑”表示数值越大时性能越好, “↓”反之

然而, FI 数据集中不同类别的图片数量存在较大差别, 拥有较多样本的类别中, 情感与一些概念的关系会得到加强, 同时减弱其他类别和相应概念之

间的权重。为抑制这种不平衡效应, 对 FI 数据集中情感和概念之间边的权重乘以平衡系数 N_{M}/N_i , N_i ($i=1, 2, \dots, 8$) 表示 FI 数据集中第 i 类图像的数量,

N_M 是所有类别中最大的图片数量(对应表 1 中方法 II)。此外,在加入平衡系数的基础上,为减少概念检测噪声的影响,将权重值小于 0.05 的边置为 0(对应表 1 中方法 III)。在该步骤中,仅使用 L_1 对模型进行优化。最终,基于 FI 数据集构建的知识图谱有 2 065 个节点、10 812 个边,基于 Emotion6 数据集构建的知识图谱有 2 018 个节点和 6 045 个边。表 1 中的实验结果表明,相比于方法 I,方法 II 和 III 均使检索和分类性能得到了提升,并且方法 III 的作用最为明显。构建良好的知识图谱可以更准确地反映其中实体之间的关系,在视觉向量投影至语义空间后,会更接近所属情感以及相应概念的空间位置。同时,对负点积相似度损失函数(L_{dot})的性能进行了评估。实验结果表明,使用 L_{dot} 时的性能次于使用 L_1 损失函数时的性能,这证明了 L_1 对视觉向量和情

感向量之间施加的约束更为有效。

3.1.2 超参数分析

本节研究不同的 λ 值对模型性能的影响。在损失函数 L 中, λ 平衡 L_1 和 L_2 的关系, λ 的值越大, L_1 的贡献越大。因为三元组中的视觉投影特征向量依赖于投影层,为使 2 个损失函数同时发挥作用, λ 的值设为 0.1 ~ 0.9。如表 2 所示,当 λ 在 0.5 ~ 0.7 之间时,2 个数据集上大部分的指标都达到了最优,并且检索和分类的性能表现出了相同的变化趋势。这说明视觉相似度约束与视觉 - 情感相似度约束之间可以起到互补作用。但是因为 L_1 和 L_2 对于检索和分类 2 个任务的贡献程度存在差异,在个别情况下,使检索和分类性能达到最优的 λ 的值不完全一致。同时,过小的 λ 值,会降低整体损失函数的性能。

表 2 不同的 λ 值下的检索和分类性能

数据集	λ	准确率/ %	检索性能					
			mAP $\uparrow$	FT $\uparrow$	ST $\uparrow$	NN $\uparrow$	DCG $\uparrow$	ANMRR $\downarrow$
FI	0.1	65.86	0.636 8	0.664 3	0.733 7	0.651 9	0.787 6	0.281 3
	0.3	67.08	0.726 4	0.703 4	0.752 8	0.677 0	0.805 0	0.226 9
	0.5	67.24	0.713 5	0.711 4	0.767 2	0.672 6	0.822 4	0.255 4
	0.7	68.07	0.746 7	0.712 1	0.806 2	0.682 4	0.828 7	0.207 4
	0.9	66.38	0.711 3	0.704 5	0.774 9	0.682 7	0.817 5	0.234 0
Emotion6	0.1	55.41	0.474 8	0.583 9	0.595 3	0.510 7	0.758 7	0.393 2
	0.3	56.67	0.543 8	0.630 5	0.629 1	0.551 8	0.801 9	0.306 5
	0.5	57.29	0.569 4	0.631 7	0.649 5	0.559 6	0.813 2	0.317 9
	0.7	57.35	0.531 7	0.627 3	0.633 3	0.553 5	0.797 9	0.338 6
	0.9	56.86	0.505 2	0.616 6	0.611 9	0.500 8	0.802 3	0.324 7

3.1.3 采样策略和间隔的影响

采样策略对于三元组的使用有重要影响,本节探讨不同采样策略对性能的影响。首先采用 Schroff 等^[26]提出的半困难采样(semi-hard sampling)策略,其中负样本与锚点之间的距离大于正样本与锚点之间的距离,由于间隔的存在,损失函数此时并不为零。作为对比,同时也使用困难采样(hard-sampling)策略进行评估。此时,负样本和锚点之间的距离小于正样本与锚点之间的距离。研究表明,半困难采样策略相比较其他方式能够使模型收敛得更快,不容易陷入局部极小值。同时,为了验证可变

间隔的作用,将 L_2 中的可变间隔设置为 0.1 进行对比。实验结果如表 3 所示,表中“H”、“SH”、“FM”、“VM”分别表示困难采样策略、半困难采样策略、固定间隔、自适应可变间隔。从表 3 可以看出,半困难采样策略的分类和检索性能均有所提升。同时,使用自适应间隔时,实验结果优于使用固定间隔。因为自适应间隔考虑到了不同正负样本的差异性,能够对难以区分的样本对施加更大的惩罚,从而更加有效地进行优化,而固定间隔忽略了不同样本的差异。

表 3 采样策略和间隔对性能的影响

数据集	H	SH	FM	VM	准确率/ %	检索性能					
						mAP ↑	FT ↑	ST ↑	NN ↑	DCG ↑	ANMRR ↓
FI	✓		✓		66.65	0.684 2	0.667 1	0.759 7	0.616 9	0.779 3	0.328 5
	✓			✓	67.22	0.735 2	0.692 9	0.771 5	0.674 4	0.812 7	0.228 1
		✓	✓		67.51	0.714 8	0.696 2	0.772 8	0.662 1	0.812 4	0.261 4
			✓	✓	68.07	0.746 7	0.712 1	0.806 2	0.682 4	0.828 7	0.207 4
Emotion6	✓		✓		55.89	0.537 0	0.514 6	0.495 8	0.513 1	0.756 1	0.547 3
	✓			✓	56.48	0.548 5	0.604 1	0.624 9	0.540 7	0.795 4	0.452 8
		✓	✓		56.78	0.540 9	0.588 4	0.642 4	0.539 0	0.788 5	0.431 6
			✓	✓	57.29	0.569 4	0.631 7	0.649 5	0.559 6	0.813 2	0.317 9

3.2 与相关研究的对比

为验证本文提出方法的有效性,在表 4 中与当前性能较好的优秀方法进行比较。基于情感标记的层级关系,Yang 等^[9]将普通的三元组损失函数扩展为一个具有层级结构的三原组损失函数。该损失函数包含两部分:第一部分使得情感相同的图片之间的距离小于不同情感类别而又属于相同情感极性的图片;第二部分则使得相同情感极性的图片之间

的距离小于不同情感极性之间图片的距离。Yao 等^[10]在 Yang^[9]提出方法的基础上,提出了可变间隔以替换固定间隔。该方法通过 softmax 层计算样本在不同类别上的置信概率,然后对在其他类别上置信度较高的样本给予更大的惩罚。此外,Yao 等^[11]基于情感标记的层级关系,提出了一种改进的 N-pair 损失函数,并基于困难程度自适应地选取负样本^[12]。

表 4 与当前优秀算法的比较以及基于不同深度模型的实验结果

数据集	方法	准确率/ %	检索性能					
			mAP ↑	FT ↑	ST ↑	NN ↑	DCG ↑	ANMRR ↓
FI	Yang ^[9]	67.64	0.488 5	0.483 4	0.697 8	0.602 3	0.780 2	0.313 5
	Yao ^[10]	68.37	0.731 9	0.696 6	0.789 2	0.672 7	0.819 7	0.210 0
	Yao ^[11]		0.734 4	0.698 5	0.781 7	0.661 3	0.811 4	0.220 1
	Yao ^[12]		0.743 3	0.707 5	0.799 4	0.675 5	0.825 0	0.210 6
	本文(AlexNet)	61.74	0.424 7	0.421 4	0.587 9	0.587 1	0.575 2	0.526 8
	本文(VGGNet)	67.27	0.711 4	0.511 0	0.792 4	0.599 1	0.787 2	0.229 8
	本文(ResNet)	68.07	0.746 7	0.712 1	0.806 2	0.682 4	0.828 7	0.207 4
Emotion6	本文(AlexNet)	52.75	0.479 2	0.603 2	0.596 5	0.515 7	0.688 4	0.507 4
	本文(VGGNet)	56.58	0.541 1	0.615 3	0.641 8	0.528 6	0.722 8	0.431 9
	本文(ResNet)	57.29	0.569 4	0.631 7	0.649 5	0.559 6	0.813 2	0.317 9

从表 4 中的检索结果可以看出,本文提出的方法超过了当前多种优秀算法。文献[9-12]的方法利用情感标记层级结构对标记空间进行约束,得到了相对普通三元组较为细致的标记空间。然而这种约束仅能增强不同情感极性之间的可区分性,却难以区分相同情感极性的差异。在情感概念的帮助下,本文提出的方法构建了一个高度结构化的标记空间,无论是类间关系还是类内关系,都通过知识图谱

得到了学习。另一方面, L_1 中三元组中的锚点属于标记层级,位置是固定的,更有助于图像向量的聚类。因此,本文方法在图像检索指标上超过了现有方法,在图像分类上同样表现良好。

表 4 中也展示了基于其他常用深度模型,AlexNet^[27]和 VGGNet^[28]的实验结果。从结果可以看出,由于先进的深度模型可以更好表示图像中的语义信息,因此分类和检索性能得到了提升。

最后,本文提出的模型经过训练,包含相同概念

的图像在投影后,在语义空间中会较为接近。而分类和检索任务都需要通过语义空间中的比较来实现,因此,作为互逆问题,检索与分类性能在大部分实验中保持了一致性。为了更直观地展示模型对图像的区分能力,在图3中给出了基于t-SNE^[29]的模型输出向量的可视化结果。每一点代表一个测试图片的视觉特征(共3316个点),不同颜色代表不同

的情感类别。如图3所示,可视化图按照从左到右的顺序,可区分度程度依次增加,经过多级损失函数优化的模型,输出向量聚类性更好。可视化结果直观地说明了本文方法无论对于相同极性的情感还是不同极性的情感的图像,模型的输出特征均具有较高程度的可区分性。

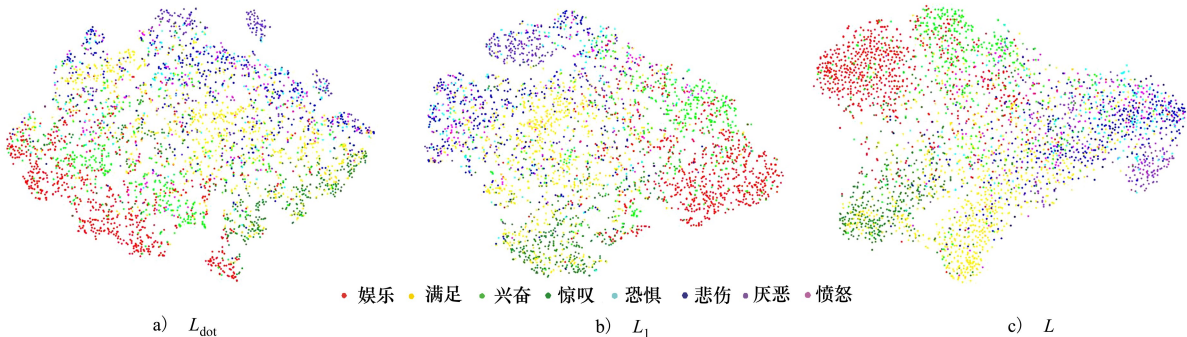


图3 模型输出特征可视化

图4中给出了一些检索结果的示例。对于被检索图像,所提出方法可以明显地得到类别相同的检索结果。并且具有一定的语义相关性。图中的最后2行展示了一些错误的检索示例,比如,代表娱乐的

图片被识别为兴奋。由于情感的模糊性,一些场景可能会引发人的多种情感,这时,单一的情感可能不足以充分代表图像的情感。

4 结 论

受到心理学中情感感知机理的启发,本研究提出利用情感概念作为媒介来解决图像情感分析中的主观性和模糊性问题。首先利用知识图谱建立情感和概念之间的联系,通过图表示方法,将知识图谱中的节点嵌入到低维的语义空间中。采用视觉语义嵌入的框架,在语义空间中对图像表示和情感进行匹配,从而学习图像特征和情感之间的关系。此此,提出了一种多级损失函数,从标记层面以及示例层面同时对模型进行优化。通过在多个检索指标以及分类指标上进行评估,本文方法表现良好,在不同的网络结构下具有鲁棒性。本研究使用的图片均来自于社交媒体,因此具有具象内容,可以通过语义概念描述。然而,一些艺术作品比如抽象画同样可以引发人的情感,却不包含明显的物体或场景。因此,对于这类图片,本研究使用的概念检测器不再适用。下一步的研究工作是使得模型能够分析抽象图片。



图4 基于本研究提出方法的图像检索示例

参考文献:

- [1] YOU Q, LUO J, JIN H, et al. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia[C]//Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, 2016: 13-22
- [2] ZHANG Y, LAI G, ZHANG M, et al. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis[C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, 2014: 83-92
- [3] ZHAO S, YAO X, YANG J, et al. Affective image content analysis: two decades review and new perspectives[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6729-6751
- [4] OLKIEWICZ K A, MARKOWSKA-KACZMAR U. Emotion-based image retrieval-an artificial neural network approach[C]//Proceedings of the International Multi Conference on Computer Science and Information Technology, 2010: 89-96
- [5] ZHANG H, YANG Z, GÖNEN M, et al. Affective abstract image classification and retrieval using multiple kernel learning[C]//International Conference on Neural Information Processing, 2013: 166-175
- [6] ZHAO S, YAO H, YANG Y, et al. Affective image retrieval via multi-graph learning[C]//Proceedings of the 22nd ACM International Conference on Multimedia, 2014: 1025-1028
- [7] YOU Q, LUO J, JIN H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015
- [8] YANG J, SHE D, SUN M. Joint image emotion classification and distribution learning via deep convolutional neural network[C]//International Joint Conference on Artificial Intelligence, 2017: 3266-3272
- [9] YANG J, SHE D, LAI Y K, et al. Retrieving and classifying affective images via deep metric learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018
- [10] YAO X, SHE D, ZHANG H, et al. Adaptive deep metric learning for affective image retrieval and classification[J]. IEEE Trans on Multimedia, 2020, 23: 1640-1653
- [11] YAO X, SHE D, ZHAO S, et al. Attention-aware polarity sensitive embedding for affective image retrieval[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1140-1150
- [12] YAO X, ZHAO S, LAI Y K, et al. APSE: Attention-aware polarity-sensitive embedding for emotion-based image retrieval[J]. IEEE Trans on Multimedia, 2020, 23: 4469-4482
- [13] GOI M T, KALIDAS V, YUNUS N. Mediating roles of emotion and experience in the stimulus-organism-response framework in higher education institutions[J]. Journal of Marketing for Higher Education, 2018, 28(1): 90-112
- [14] CHEN T, BORTH D, DARRELL T, et al. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks[J/OL]. (2014-10-30) [2022-08-15]. <https://arxiv.org/abs/1410.8586>
- [15] BORTH D, JI 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, 2013: 223-232
- [16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778
- [17] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 855-864
- [18] XU H, QI G, LI J, et al. Fine-grained image classification by visual-semantic embedding[C]//International Joint Conference on Artificial Intelligence, 2018: 1043-1049
- [19] YOU Q, LUO J, JIN H, et al. Building a large scale dataset for image emotion recognition: the fine print and the benchmark[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2016
- [20] MIKELS J A, FREDRICKSON B L, LARKIN G R, et al. Emotional category data on images from the International Affective Picture System[J]. Behavior research methods, 2005, 37(4): 626-630
- [21] 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, 2015: 860-868
- [22] EKMAN P. What emotion categories or dimensions can observers judge from facial behavior? [J/OL]. (2013-11-17) [2022-08-

15]. <https://www.sciencedirect.com/science/article/abs/pii/B9780080166438500252>

[23] GAO Y, WANG M, TAO D, et al. 3-D object retrieval and recognition with hypergraph analysis[J]. IEEE Trans on Image Processing, 2012, 21(9): 4290-4303

[24] MANJUNATH B S, OHM J R, VASUDEVAN V V, et al. Color and texture descriptors[J]. IEEE Trans on Circuits and Systems for Video Technology, 2001, 11(6): 703-715

[25] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C] // 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255

[26] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 815-823

[27] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 83-90

[28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C] // 3rd International Conference on Learning Representations, San Diego, 2015

[29] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605

Image emotion analysis based on semantic concepts

YANG Hansen¹, FAN Yangyu¹, LYU Guoyun¹, LIU Shiya², GUO Zhe¹

(1.School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China;
2.Content Production Center of Virtual Reality, Beijing 101318, China)

Abstract: With the increasing number of users express their emotions via images on social media, image emotion analysis attracts much attention of researchers. For the ambiguity and subjectivity of emotion, image emotion analysis is more challenging than other computer vision tasks. Previous methods merely learn a direct mapping between image feature and emotion. However, in emotion perception theory of psychology, it is demonstrated that human beings perceive emotion in a stepwise way. Therefore, we propose a novel image emotion analysis framework that makes use of emotional concepts as middle-level feature to bridge image and emotion. Firstly, the relationship between the concept and the emotion is organized in the form of knowledge graph. The relation between the image and the emotion in the semantic embedding space is explored where the knowledge is encoded into. On the other hand, a multi-level deep metric learning method to optimize the model from both label level and instance level is proposed. Extensive experimental results on two image emotion datasets, demonstrate that the present approach performs favorably against the state-of-the-art methods on both affective image retrieval and classification tasks.

Keywords: image emotion analysis;knowledge graph;visual-semantic embedding;deep metric learning

引用格式:杨瀚森, 樊养余, 吕国云, 等. 基于语义概念的图像情感分析[J]. 西北工业大学学报, 2023, 41(4): 784-793
YANG Hansen, FAN Yangyu, LYU Guoyun, et al. Image emotion analysis based on semantic concepts[J]. Journal of Northwestern Polytechnical University, 2023, 41(4): 784-793 (in Chinese)