计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190800154

基于视觉语义联合嵌入和注意力机制的情感预测

蓝亦伦　孟　敏　武继刚
广东工业大学计算机学院　广州 510006
(allentoretto@163.com)

摘　要　为了缓解图像视觉特征与情感语义特征之间存在的鸿沟,减弱图像中情感无关区域对情感分类的影响,提出了一种结合视觉语义联合嵌入和注意力模型的情感分类算法。首先利用自编码器学习图像的视觉特征和情感属性的语义特征的联合嵌入特征,缩小低层次的视觉特征与高层次的语义特征之间的差距;然后提取图像的一组显著区域特征,引入注意力模型建立显著区域与联合嵌入特征的关联,确定与情感相关的显著区域;最后基于这些显著区域特征构建情感分类器,实现图像的情感分类。实验结果表明,该算法有效地改进了现有的图像情感分类方法,显著提高了对测试样本的情感分类精度。

关键词:视觉情感预测;视觉语义联合嵌入;注意力机制;显著区域检测
中图法分类号　TP391.41

Visual Sentiment Prediction with Visual Semantic Embedding and Attention Mechanism

LAN Yi-lun, MENG Min and WU Ji-gang
Department of Computer Science, Guangdong University of Technology, Guangzhou 510006, China

Abstract　In order to bridge the semantic gap between visual features and sentiments and reduce the impact of sentiment irrelevant regions in the image, this paper presents a novel visual sentiment prediction method by integrating visual semantic embedding and attention mechanism. Firstly, the method employs the auto-encoder to learn joint embedding of image features and semantic features, so as to alleviate the difference between the low-level visual features and the high-level semantic features. Secondly, a set of salient region features are extracted as input to the attention model, in which the correlations between salient regions and joint embedding features can be established to discover sentiment relevant regions. Finally, the sentiment classifier is built on top of these regions for visual sentiment prediction. The experimental results show that, the proposed method significantly improves the classification performance on testing samples and outperforms the state-of-the-art algorithms on visual sentiment analysis.

Keywords　Visual sentiment prediction, Visual semantic embedding, Attention mechanism, Salient regions detection

1　引言

随着互联网技术的飞速发展,社交网络成为了现代社会人们日常生活中最为重要的信息交流平台。在大数据时代,海量的信息蕴含着人们的各种看法和观点,这使得互联网成为了丰富的情绪资源库。对人们在互联网上发布的多媒体信息(如文字、图片、视频等)进行情感计算和情感分析,在舆情监控、辅助人类决策、智能广告等应用上发挥着重要的作用[1]。不同于从文本语义分析得到情感信息的文本情感分析,图像情感分析能从图像内容中获得更丰富的情感线索。

图像情感分析主要研究的是人类对视觉刺激(如图像、视频)的情绪反应[2],这是一个高层次的抽象问题,它要求计算机从人类情感的角度理解一个抽象的概念。另外,由于人们受同一张图片的视觉刺激的反应可能存在差异,具有主观性,使得这个问题极具挑战。

图像情感分析所要解决的关键性问题是如何填补低层次的视觉特征与高层次的情感之间巨大的情感鸿沟[3]。直接建立图像视觉特征到情感类别的关联,效果往往不尽如人意。目前,解决这一问题的一种方法是在分类模型中增加辅助信息,使得图像能够从图像的特征空间转换到辅助信息的语义特征空间,实现关联的迁移,从而减少模型在分类过程中对训练样本的依懒性。常用的辅助信息有属性等情感信息。

从已有的图像情感分析方面的文献可知,研究主要从3个方面针对视觉特征进行改进:基于低层次视觉特征的方法、基于中层语义的方法以及基于深度视觉特征的方法[4]。图像情感分析的性能随着视觉特征的鲁棒性的增强而逐步提高。然而,大部分方法都是试图从整张图像来挖掘其表达的情感,人们对一张图片中局部区域表达情感的强弱程度,以及局部区域如何对图像情感分析产生影响的研究却很少。文献[3]的研究结果表明,借助辅助信息构建属性检测器,再结合注意

到稿日期:2019-08-29　返修日期:2019-11-22　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61702114);广东省科技计划重点领域研发项目(2019B010121001)
This work was supported by the National Natural Science Foundation of China (61702114) and Guangdong Key R&D Project of China (2019B010121001).
通信作者:孟敏(minmeng@gdut.edu.cn)

力模型，图像情感的预测性能得到了有效的改善。但是，由于其属性检测器仅利用基本的卷积神经网络进行微调，泛化性能差，导致属性检测器的效果不够好，且其基于注意力模型的情感预测极度依赖于属性检测器的精度，进而导致情感预测的性能受限。

图像的显著区域能够不同程度地影响视觉情感[5]。文献[5]提出了图像的消极情感是由显著区域引起的，而积极情感同时受显著区域和背景信息的影响。文献[6]研究了视觉显著区域与其表达的情感信息之间的联系，提出了一种新的方法，用于量化在复杂场景下的情感反应与视觉注意力之间的关系。但是，不管图像的内容如何，以往的注意力机制的输入都是卷积神经网络的某一层特征或者多层特征的输出，这些特征就像是用均匀网格切分后得到的相同形状大小的神经感受域[7]。这类方法很少考虑如何确定图像的显著区域。

为了解决这些问题，本文提出了一个新的图像情感预测框架。本文框架由基于自编码器的视觉语义联合嵌入和基于注意力机制的情感分类两个模块组成。本文首先设计了一种基于自编码器的视觉语义联合嵌入，使中间的隐藏层有能力重构从卷积神经网络提取的图像特征，并能回归到用于情感属性表达的语义向量。这样设计的目的是尽可能消除低层次的视觉特征与高层次的情感属性之间的语义鸿沟。其次，受到文献[7]的启发，利用文献[7]提出的自底向上的显著区域检测网络，提取一组显著的图像区域，其中每个区域由一个卷积特征向量表示。我们将多个显著区域特征和自编码器训练得到的视觉语义联合嵌入层输入到注意力模型中，学习多个显著图像区域与情感语义的关联，进而进行情感分类。本文的贡献如下：

1）设计了一种基于自编码器的视觉语义联合嵌入的方法，将情感属性向量作为额外的监督，得到的视觉语义联合嵌入层能有效地缩小低层次的图像特征与高层次的情感属性之间的语义差距。

2）在得到训练好的视觉语义的联合嵌入特征的基础上，将图像的显著区域检测网络与注意力机制结合起来，在注意力机制中学习显著区域与带有情感语义的联合嵌入层之间的关联，基于这些区域的分类器在图像情感预测方面表现出了较好的性能。

2 相关工作

现今，人们在社交网络上使用大量的图片来表达观点，因此图像情感分析越来越受到人们的关注。

在图像情感分析工作的早期，研究主要针对传统的低层次视觉特征进行改进。文献[8]提出了3种模糊的视觉直方图来预测图像中的情绪。文献[9]提取了基于心理学和艺术理论的组合特征用于情感检测，这些特征包括构图、颜色变化、图像纹理等。文献[10]提出了一种根据艺术原理设计的视觉特征。这些人工改造的视觉特征在几个特定领域（如抽象绘画和艺术照片）的小数据集上被证明是有效的[9]。

基于低层次视觉特征的方法并不能有效弥补低层特征与高层情感之间的情感鸿沟，因此研究逐渐转向中层语义，即利用相关的属性来弥补情感鸿沟。文献[11]提出了中层概念，即借助形容词、名词对来检测图像，而不是直接检测图像表达的情感。文献[12]在图像情感分析中考虑了文本情感，提出计算描述图像的形容词名词对所包含的文本情感值的加权和。

现今，深度学习在视觉和语言相关的任务中取得了重大进展，在解决更高层次的语义理解任务中起着不可小觑的作用。近期的研究逐渐转向设计更为深入的神经网络架构来分析视觉情感。文献[13]利用从大型通用数据集[14]中学习到的模型参数，来对现有的卷积神经网络模型进行情绪预测的微调。文献[15]提出了一个CNN-RNN模型，通过不同层次的特征融合，利用它们之间的依赖关系来预测情绪。文献[3]提出了一种基于注意力模型的方法，通过关注局部区域的图像信息来学习图像特征与语义信息之间的映射。但是，文献[3]并未真正消除图像特征与情感语义之间的鸿沟，利用部分类别的样本所训练的属性检测器，来对图像进行属性分类，其泛化性能差，严重影响了最终情感分类的精度。

与文献[3]不同的是，本文提出基于自编码器的视觉语义联合嵌入方法来减小图像特征与情感语义之间的鸿沟，得到融合情感语义与图像特征的视觉语义的联合嵌入层，增强了模型的泛化能力。在得到视觉语义的联合嵌入层后，利用显著区域检测网络来提取多个显著区域的特征，将显著区域特征与视觉语义的联合嵌入特征一起输入到注意力模型中进行情感分类。与文献[7]不同的是，本文所采用的显著区域检测网络用于情感分析领域，而文献[7]的显著区域检测网络用于图像字幕生成和视觉问答。

3 本文方法

本文提出的框架主要由两个模块组成：基于自编码器的视觉语义联合嵌入模块和基于注意力模型的情感分类模块。

3.1 基于自编码器的视觉语义联合嵌入

自动编码器作为深度学习中的一种无监督学习方法，在自然语言处理领域取得了较好的效果[16-17]。自动编码器的基本思想是：把原始的高维特征向量转化成低维向量，在这个过程中学习原始数据中的潜在特征，剔除高维特征中的冗余部分，得到原始数据的精炼表达。本文所采用的方法针对图像特征和属性语义特征进行自编码学习，如图1所示。

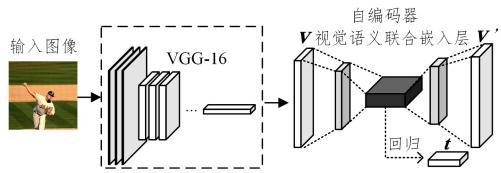

图1 基于自编码器的视觉语义联合嵌入框架

Fig. 1 Auto-encoder based visual semantic embedding framework

本文使用ImageNet分类挑战中预先训练好的VGG-16卷积网络，将图片输入后得到fc7全连接层的特征向量 $\boldsymbol{v}$。将 $\boldsymbol{v}$ 作为自编码器的输入，先通过编码器对其进行线性变化，然后通过激活函数得到一个编码后的视觉语义的联合嵌入特征 z。此处我们使用relu函数作为激活函数，具体如式(1)所示。随后把编码结果 z 输入解码器中，得到重构的向量 $\boldsymbol{v}'$，此处选用的激活函数为sigmoid函数，计算式如式(2)所示。为了让自编码器的嵌入层能够学习到情感属性的语义信息，我们将编码得到联合嵌入特征 z 回归到图像的属性向量 $\boldsymbol{t}$，计算式如式(3)所示。

$$z=f_\theta(v)=\text{relu}(W^{\text{in}}v+b^{\text{in}}) \tag{1}$$

$$v'=g_\alpha(z)=\text{sigmoid}(W^{\text{out}}z+b^{\text{out}}) \tag{2}$$

$$t=h_\beta(z)=g(W^h z+b^h) \tag{3}$$

其中，W^{in} 和 b^{in} 是编码器的参数，W^{out} 和 b^{out} 是解码器的参数，W^h 和 b^h 是回归的参数。$g(\cdot)$ 表示激活函数，3.3 节将对所选择的激活函数进行测试比较。本文采用交叉熵损失函数来定义视觉语义联合嵌入模型的损失，具体损失函数如下：

$$L_{\theta,\alpha}=-\sum_i v_i\log(v_i') \tag{4}$$

$$L_{\theta,\beta}=-\sum_i l_i\log(t_i) \tag{5}$$

$$L_{\theta,\alpha,\beta}=L_{\theta,\alpha}+L_{\theta,\beta} \tag{6}$$

其中，$L_{\theta,\alpha}$ 是图像特征的重构误差，$L_{\theta,\beta}$ 是属性向量的回归误差；l 是每张图片对应的标注属性值，将其作为回归得到的属性向量 t 的监督信息；θ 是自编码器中编码器的参数，α 是自编码器中解码器的参数，而 β 是属性回归的参数。

本文所构建的基于自编码器的视觉语义联合嵌入网络，既确保了联合嵌入特征 z 能够学习到有效的图像特征，又能通过标注属性的监督学习到情感属性的语义特征，尽可能弥补低层次的图像特征与高层次的语义特征之间的鸿沟。

3.2 基于注意力模型的情感分类

图 2 给出了基于注意力模型的情感分类结构，该模型由显著区域初始化模块、视觉注意模块和情感分类模块 3 个部分组成。首先，把图片输入显著区域检测模块，利用 Faster-RCNN 与 ResNet101 相结合的模型[7]，得到一组图像显著区域特征。然后，将这组特征与视觉语义的联合嵌入特征一起，作为视觉注意模块的输入，计算出每一个显著区域的注意权重。将注意权重与对应显著区域进行加权，得到注意特征。最后，将所有的注意特征输入情绪分类模块，实现最终的情感预测。

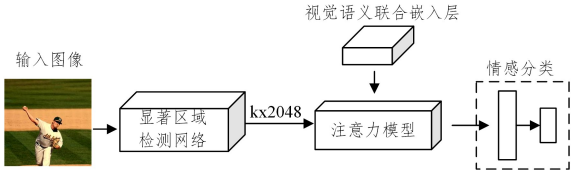

图 2 基于注意力模型的情感分类框架

Fig. 2 Visual sentiment prediction framework based on attention mechanism

3.2.1 显著区域初始化

为了检测图片中的显著区域，本文使用文献[7]中预先训练好的 Faster-RCNN 模型与 ResNet-101 卷积神经网络相结合的模型对图片的显著区域进行检测模型。对于卷积神经网络得到的某一层或多层特征而言，这些特征就像是从用均匀网格切分后得到的相同尺寸的神经感受域中提取的。这样所得到的特征没有突出显著区域，对于后续任务的效果有限[7]。而文献[7]提出的显著区域检测网络，能够得到图像中物体和其他显著区域的特征，能够促进后续任务的效果。

本文将图片输入显著区域检测模型后，得到一组包含 k 个显著图像的区域，其中每个区域由一个 D 维的卷积特征向量 $\boldsymbol{X}_i$ 表示。

3.2.2 视觉注意

视觉注意机制的存在，使我们能够捕捉到最突出的区域。它在研究中已被证明是一种有效的理解图像的方法[18]。视觉注意机制在情感分析中也很重要。本文在显著区域特征和视觉语义的联合嵌入特征的基础上，对视觉注意力集中的区域进行建模，以弥补不同模态下的数据之间的差距。

给定一组包含 k 个显著图像区域特征 $\boldsymbol{X}_i$，用 $\boldsymbol{X}_i=(x_{i,1},x_{i,2},\cdots,x_{i,k})$ 表示，其中 $x_{i,j}$ 表示第 i 张图片的第 j 个显著区域。在注意力模型中，$s_{i,j}$ 表示第 i 张图片的第 j 个显著区域与视觉语义的联合嵌入特征 z_i 之间的相关性，本文将其作为衡量图像的每个显著区域的权重分数。

$$s_{i,j}=softmax(z_i\boldsymbol{U}x_{i,j}) \tag{7}$$

其中，$\boldsymbol{U}$ 是要学习的权重矩阵。通过式(7)计算得到图像的不同显著区域与联合嵌入特征之间的关联度。将每个显著区域对应的权重分数 $s_{i,j}$ 加权到对应的图像特征 $x_{i,k}$ 中，得到所有显著区域特征的加权求和。

$$\hat{x}_i=\sum_{j=1}^{k}s_{i,j}(\boldsymbol{U}x_{i,j}) \tag{8}$$

在得到所有显著区域的加权特征 $\hat{x}_i$ 之后，把它输入情感分类模块中。

3.2.3 情感分类

将显著区域的加权特征输入用于情绪分类的 $softmax$ 分类器，在图像对应的情感标签的监督下，利用负对数似然法计算注意力模型的损失，计算式如下：

$$f(\hat{x}_i)=softmax(W_2\tanh(W_1\hat{x}_i)) \tag{9}$$

$$L(\hat{x}_i)=-\log(f(\hat{x}_i,p_i)) \tag{10}$$

式(9)中，W_1 和 W_2 是情感分类器的参数；式(10)中，p_i 为第 i 个图像的情感标签。整个网络框架通过反向传播进行训练。

4 实验及结果分析

4.1 实验数据集

本文提出的模型在一个基准数据集 Visual Realism Dataset[19]（简称为 VRD）上进行实验和评估。VRD 数据集共有 2520 张图像，每张图像对应着 38 个标注，其中既包含带有人类情感信息的标注，也包含带有其他语义信息的标注，每个标注都有一个属性值与其对应，取值范围为 0～1。由于 VRD 数据集不提供人工标注的二进制情感标签，本文参照文献[5]中的方法对数据集进行了处理。首先选取了 VRD 中与情感信息密切相关的 3 个标注属性“make happy”“attractive”和“colorful”，通过计算这 3 个标注属性的均值得到每张图片的情感分数，其取值范围为 0～1。为了得到一个相对均衡的数据集，我们选取了情绪评分表上得分较高的前 20%作为图像情感正标签和得分较低的前 15%作为情感负标签，正负标签的图像数量分别为 504 和 378。

4.2 实验细节

实验开发环境为 windows 10，python 3.5，开发工具为 JetBrains PyCharm 2017。本文所有方法在单个 NVIDIA GeForce GTX Titan Xp GPU 环境下用 tensorflow 框架训练。

本文算法需要在训练自动编码器和训练注意力模型之前进行图像特征提取。在训练自动编码器前，视觉特征是利用在 ImageNet 上预先训练好的 VGG-16 模型进行提取的，输入整张图像到 VGG-16 模型中，提取 fc7 全连接层的 4 096 维特

征。为了获得更好的回归效果，本文将每张图片对应的 38 个连续的标注属性值作为监督信息，用于自编码器中视觉语义的联合嵌入层的回归。利用小批量的随机梯度下降法，设置学习率为 0.001，每批次输入 32 张图片，迭代次数为 5，对基于自编码器的视觉语义的联合嵌入模型进行训练，训练时间大概为 10～15 min。

在训练注意力模型之前，我们利用文献[7]预训练好的 Faster-RCNN 与 ResNet101 相结合的显著区域检测模型，从整个图像中提取一组包含 k 个 D 维的显著区域图像特征，将这组特征和训练得到的联合嵌入特征一起输入注意力模型中。在本文实验中，$k=36$，$D=2048$。我们使用小批量的随机梯度下降法训练注意力模型，设置学习速度为 0.001，每批次输入 32 张图片，迭代次数为 5 次，训练时间大概为 20～25 min。

4.3 实验结果分析

为了评估基于自编码器的视觉语义联合嵌入模块和基于注意力机制模块对情感分类的影响，本文设置了 3 组实验与本文算法进行对比。

实验 1(表 1 将实验 1 中的模型称为原始模型)中，我们将图片输入预训练的 VGG-16 模型得到 fc7 层图像特征，基于该特征搭建一个两层的情感分类器，直接对特征进行情感预测。这样设置的目的是将实验 1 的结果作为基准，体现 VGG-16 在情感分类上的能力，并与本文方法进行对比。

实验 2(表 1 将实验 2 中的模型称为去联合嵌入模型)中，为了体现视觉语义联合嵌入模块对情感分类的效果，我们将本文方法中视觉语义联合嵌入模块替换成 VGG-16 的属性回归检测模块，其余模块保留不变。VGG-16 的属性回归检测模块，是把 VGG-16 中提取的 fc7 层特征输入到两层的回归网络中，得到一个回归向量，把它和显著特征作为注意力机制的输入进行情感分类。

实验 3(表 1 将实验 3 中的模型称为去显著检测模型)中，为了体现显著区域检测网络在本文方法中所起的作用，我们将本文方法中的显著性检测网络所提取的一组显著特征替换成预训练 VGG-16 模型的 Conv5-3 特征，其余模块不变。具体实验结果如表 1 所列。

表 1 对比实验结果

Table 1 Results of contrast experiment

对比实验	分类精度
原始模型	0.72
去联合嵌入模型	0.79
去显著检测模型	0.69
本文方法	0.92

由表 1 可以得到如下结论：1)原始模型得到了 72.9%的精度，表明预训练好的 VGG-16 模型对 VRD 数据集具有一定的情感分类能力；2)去联合嵌入模型得到了 78.7%的精度，对比本文方法可知，基于自编码器的视觉语义联合嵌入模块通过有效融合图片视觉特征和图片属性的语义特征，弥补了低层次的图像特征与高层次的语义特征之间的鸿沟，为后续的注意力模型提供了较为准确的情感语义信息，从而有效提高了图像情感分类的精度；3)去显著检测模型得到了 69.3%的精度，表明显著区域检测网络能够确定图像重要区域的位置，其显著区域特征对于后续的情感分类起到了重要的作用，有效提升了情感分类的精度。

表 2 不同方法的实验结果对比

Table 2 Comparison of performance of different algorithms

Algorithm	Accuracy
DeepSentiBank	0.82
PCNN	0.81
FTCNN	0.75
NUSFocal	0.86
文献[3]方法	0.77
本文方法	0.92

将本文方法与 5 种最先进的图像情感分类方法进行比较，结果如表 2 所列。5 种方法分别为：1)文献[3]中的方法；2)DeepSentiBank[20]，一种利用 CaffeNet[21] 进行视觉情感概念分类的模型；3)PCNN[22]，一种利用逐步训练和域迁移的情感分类模型；4)FTCNN[23]，一种采用 AlexNet 风格搭建的深度神经网络情感分类模型；5)NUSFocal[5]，一种将聚焦对象 mask 或显著性特征作为图像输入的第四通道的情感分类模型。以上 5 种方法均在 VRD 数据集上进行训练测试。我们主要对文献[3]中的算法进行了复现，对于其余算法本文直接采用文献[5]所提供的结果。由表 2 可知，本文方法在 VRD 数据集上的分类精度明显高于 5 种对比方法的分类精度，得到了较好的情感分类结果。对比文献[3]在 VRD 数据集上的结果可以看出，本文方法中图像的显著区域和视觉语义联合嵌入模块在一定程度上提升了图像情感分类的精度。

为了进一步验证本文方法的可靠性，测试自编码器中视觉语义联合嵌入层的不同维度 N 对算法性能的影响以及测试自编码器中视觉语义联合嵌入层回归到属性向量所选择的激活函数 $g(\cdot)$对算法性能的影响，我们对自编码器中的嵌入层维度 N 和回归的激活函数 $g(\cdot)$进行测试。由于属性值的取值为 0～1，我们设置两种激活函数 *softmax* 和 *sigmoid*；对于联合嵌入层维度 N，我们设置 $N=100,200,300,400,500$。

实验结果如图 3 所示，当嵌入层的激活函数选取 *softmax* 时，在不同嵌入层维度 N 上的实验效果比激活函数选取 *sigmoid* 时的效果要好，其中当 N 取 300 维时，所得到的视觉语义联合嵌入层对后续的情感分类能得到最优的分类结果。

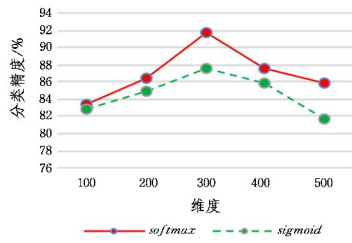

图 3 嵌入层的维度与激活函数对精度的影响

Fig. 3 Impact of dimensionality of embedding layer and activation function on accuracy

4.4 实验方法的局限性

本文方法在 VRD 数据集上取得了较好的效果，但存在着一些局限性。首先，基于自编码器的视觉语义联合嵌入层能够有效弥补低层次的图像特征与高层次的语义特征之间的差距，但所采用的自编码器结构能捕捉到的图像信息和属性语义信息比较有限，未来将尝试对自编码器结构进行改进或

者增加辅助信息，以增强嵌入层在重构和回归上的能力；其次，本文所采用的注意力模型以显著区域的特征为输入，未来将尝试引入多层注意力机制，以增强模型对显著性区域所蕴含的信息的捕捉。最后，本文所采用的数据集规模相对较小，之后会尝试在相对较大的数据集上进行测试。

结束语 本文提出了一种基于自编码器的视觉语义联合嵌入和注意力机制的图像情感分类方法。该方法一方面通过自编码器学习图像视觉特征和属性语义特征的联合特征，来尽可能弥补低层次的图像特征与高层次的语义特征之间的语义鸿沟；另一方面考虑到图像中冗余区域对图像情感分类的影响，提出利用显著检测网络提取图像的一组显著区域，通过注意力机制建立显著区域与带有情感语义的联合嵌入特征之间的非线性关系，从而高效地进行图像情感预测。本文在一个基准数据集上对提出的算法进行验证，并与最先进的方法进行了比较，结果表明，本文方法从总体上增强了情感分类算法的性能，提高了对测试样本的情感预测精度。

参 考 文 献

[1] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends® in Information Retrieval, 2008, 2(1/2): 1-135.

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

[3] YOU Q, JIN H, LUO J. Visual sentiment analysis by attending on local image regions[C]// Thirty-First AAAI Conference on Artificial Intelligence. 2017: 231-237.

[4] SONG K, YAO T, LING Q, et al. Boosting image sentiment analysis with visual attention[J]. Neurocomputing, 2018, 312: 218-228.

[5] FAN S, JIANG M, SHEN Z, et al. The Role of Visual Attention in Sentiment Prediction[C]// Proceedings of the 25th ACM International Conference on Multimedia. ACM, 2017: 217-225.

[6] FAN S, SHEN Z, JIANG M, et al. Emotional attention: A study of image sentiment and visual attention[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7521-7531.

[7] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.

[8] WEI-NING W, YING-LIN Y, SHENG-MING J. Image retrieval by emotional semantics: A study of emotional space and feature extraction[C]// 2006 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2006, 4: 3534-3539.

[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. ACM, 2010: 83-92.

[10] ZHAO S, GAO Y, JIANG X, et al. Exploring principles-of-art features for image emotion recognition[C]// Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 47-56.

[11] 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. ACM, 2013: 223-232.

[12] LI Z, FAN Y, LIU W, et al. Image sentiment prediction based on textual descriptions with adjective noun pairs[J]. Multimedia Tools and Applications, 2018, 77(1): 1115-1132.

[13] CAMPOS V, JOU B, GIRO-I-NIETO X. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction[J]. Image and Vision Computing, 2017, 65: 15-22.

[14] 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. IEEE, 2009: 248-255.

[15] ZHU X, LI L, ZHANG W, et al. Dependency Exploitation : A Unified CNN-RNN Approach for Visual Emotion Recognition [C]// IJCAI. 2017: 3595-3601.

[16] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy layer-wise training of deep networks[C]// Advances in Neural Information Processing Systems. 2007: 153-160.

[17] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(Dec): 3371-3408.

[18] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// International Conference on Machine Learning. 2015: 2048-2057.

[19] FAN S, NG T T, HERBERG J S, et al. An automated estimator of image visual realism based on human cognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 4201-4208.

[20] CHEN T, BORTH D, DARRELL T, et al. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks[J]. arXiv: 1410.8586, 2014.

[21] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convol-utional neural networks[C]// Advances in Neural Information Processing Systems. 2012: 1097-1105.

[22] YOU Q, CAO L, JIN H, et al. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks[C]// Proceedings of the 24th ACM International Conference on Multimedia. ACM, 2016: 1008-1017.

[23] CAMPOS V, JOU B, GIRO-I-NIETO X. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction[J]. Image and Vision Computing, 2017, 65: 15-22.

LAN Yi-lun, born in 1995, postgraduate. His main research interests include visual sentiment prediction and image classification

MENG Min, born in 1985, Ph.D, associate professor, postgraduate supervisor, is a member of China Computer Federation. Her main research interests include image processing and machine learning.