计算机科学
COMPUTER SCIENCE

基于自适应门控信息融合的多模态情感分析

陈真, 普园媛, 赵征鹏, 徐丹, 钱文华

引用本文

陈真, 普园媛, 赵征鹏, 徐丹, 钱文华基于自适应门控信息融合的多模态情感分析[J]. 计算机科学, 2023, 50(3): 298-306.

CHEN Zhen, PU Yuanyuan, ZHAO Zhengpeng, XU Dan, QIAN Wenhua. Multimodal Sentiment Analysis Based on Adaptive Gated Information Fusion [J]. Computer Science, 2023, 50(3): 298-306.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100156

基于自适应门控信息融合的多模态情感分析

陈　真[1]　普园媛[1,2]　赵征鹏[1]　徐　丹[1]　钱文华[1]
1 云南大学信息学院　昆明 650504
2 云南省高校物联网技术及应用重点实验室　昆明 650504
(15837332933@163.com)

摘　要　多模态情感分析的目标是使用由多种模态提供的互补信息来实现可靠和稳健的情感分析。近年来,通过神经网络提取深层语义特征,在多模态情感分析任务中取得了显著的效果。而多模态信息的不同层次的特征融合也是决定情感分析效果的重要环节。因此,提出了一种基于自适应门控信息融合的多模态情感分析模型(AGIF)。首先,通过门控信息融合网络将Swin Transformer和ResNet提取的不同层次的视觉和色彩特征根据对情感分析的贡献进行有机融合。其次,由于情感的抽象性和复杂性,图像的情感往往由多个细微的局部区域体现,而迭代注意可以根据过去的信息精准定位这些情感判别区域。针对Word2Vec和GloVe无法解决一词多义的问题,采用了最新的ERNIE预训练模型。最后,利用自动融合网络"动态"融合各模态特征,解决了(拼接或TFN)确定性操作构建多模态联合表示所带来的信息冗余问题。在3个公开的真实数据集上进行了大量实验,证明了该模型的有效性。

关键词:多模态情感分析;门控信息融合网络;迭代注意;ERNIE;自动融合网络
中图法分类号　TP391

Multimodal Sentiment Analysis Based on Adaptive Gated Information Fusion

CHEN Zhen[1],PU Yuanyuan[1,2],ZHAO Zhengpeng[1],XU Dan[1] and QIAN Wenhua[1]
1 College of Information Science and Engineering,Yunnan University,Kunming 650504,China
2 University Key Laboratory of Internet of Things Technology and Application,Yunnan Province,Kunming 650504,China

Abstract　The goal of multimodal sentiment analysis is to achieve reliable and robust sentiment analysis by utilizing complementary information provided by multiple modalities. Recently,extracting deep semantic features by neural networks has achieved remarkable results in multimodal sentiment analysis. But the fusion of features at different levels of multimodal information is also an important part in determining the effectiveness of sentiment analysis. Thus,a multimodal sentiment analysis model based on adaptive gating information fusion(AGIF) is proposed. Firstly,the different levels of visual and color features extracted by swin transformer and ResNet are organically fused through a gated information fusion network based on their contribution to sentiment analysis. Secondly,the sentiment of an image is often expressed by multiple subtle local regions due to the abstraction and complexity of sentiment,and these sentiment discriminating regions can be located accurately by iterative attention based on past information. The latest ERNIE pre-training model is utilized to solve the problem of Word2Vec and GloVe's inability to handle the word polysemy. Finally,the auto-fusion network is utilized to "dynamically" fuse the features of each modality,solving the problem of information redundancy caused by the deterministic operation(concatenation or TFN) to construct multimodal joint representation. Extensive experiments on three publicly available real datasets demonstrate the effectiveness of the proposed model.

Keywords　Multimodal sentiment analysis,Gated information fusion networks,Iterative attention,ERNIE,Auto-fusion network

到稿日期:2022-01-16　返修日期:2022-09-20
基金项目:国家自然科学基金(62162068,61271361,61761046,62061049);云南省应用基础研究面上项目(2018FB100);云南省科技厅应用基础研究计划重点项目(202001BB050043,2019FA044)
This work was supported by the National Natural Science Foundation of China(62162068,61271361,61761046,62061049),Yunnan Science and Technology Department Project(2018FB100) and Key Program of the Applied Basic Research Programs of Yunnan(202001BB050043,2019FA044).
通信作者:赵征鹏(zhpzhao@ynu.edu.cn)

1　引言

随着社交网络和移动设备的普及，用户每天都有大量的图片和文字记录着自己生活中的各种活动，例如人们分享其旅行经历和对某些事件的看法等。这类带有用户丰富情感的多模态数据（图片和文本），在政治选举[1]、大盘走势[2,3]、票房预测[4]和情绪干预等方面都有着潜在的应用价值。

图1给出了来自ArtPhoto数据集的8种不同情绪的图文示例。其中，图1(a)的图片和文本都表达了“愉快”的情绪；图1(e)中的文本不同的人看到会有不同的感受，有的人觉得是生气，也有人觉得是一种警示，而所对应的图片表达出了非常强烈的愤怒；图1(b)中的图片对不同的人可能会有不同的情感，而文本明显表达出了“敬畏”之情。在图文多模态数据中，图像和文本包含的信息一般是相辅相成的。相比单模态（图片或文本）情感分析，多模态包含了更全面的信息，也可以更好地表达用户的真实情感。此外，随着社交媒体多样化、多元化的发展，单模态情感分析也无法处理多样性的社交媒体数据。目前的研究可以分为以下两类。

图1　ArtPhoto数据集图文示例

Fig. 1　Examples of images and text from ArtPhoto

第一种类型是分别对待不同来源的特征。例如，Xu等[5]提出了一种用于多模态情感分析的深度语义网络。首先识别目标和场景，并将其作为显著检测器提取图像的深层语义特征，再利用视觉特征引导的LSTM提取重要的词语。但只是利用显著性进行采样，忽略了人类的感知过程。Huang等[6]提出了一种深度多模态注意融合网络，利用视觉内容和语义信息之间的区别特征和内在联系，建立了一种混合融合框架进行情感分析。该方法虽取得了不错的效果，但仅利用了深层语义信息，忽略了其他各层有用信息对情感分析的贡献。此外，人在感知事物时，一般由多个区域的信息综合进行整体的判断和感受。如图1(g)所示，“老鼠”这块区域传达出了恐惧的心理，而“猫”这块区域更好地体现出了老鼠恐惧的原因，起到了很好的补充作用。而仅一次注意力往往无法完整地定位到一幅图像中的多个情感判别区域。Lin等[7]提出了一种基于注意力的多模态情感分析模型，通过张量融合网络(Tensor Fusion Network，TFN)融合各个模态的特征。一旦特征过长，就会引起参数爆炸，难以训练。

第二种类型是联合不同来源的特征。例如，Xu等[8]提出了一种新的共记忆网络，迭代地模拟视觉内容和文本词汇之间的交互作用。Xu等[9]提出了一种新的双向多层次注意模型，利用视觉特征和文本特征之间的互补性和综合性信息进行图像-文本情感联合分析。Yang等[10]提出了一种多视角注意模型。该模型利用不断更新的记忆网络来获取图像-文本的深层语义特征。该模型包括特征映射、交互学习和特征融合3个阶段。这些方法虽考虑了不同模态之间的内在联系，但大多数图像和文本之间的语义不相关或弱相关性，会导致错误的关联。此外，若想取得更好的效果，出色的融合机制是必不可少的。这些方法在特征融合上大多是暴力拼接，导致信息冗余，无法有效地利用各模态信息。

除了情感分析领域，视觉问答(VQA)作为多模态任务，也需要理解图片和文本问题，从而推理出答案。Anderson等[11]提出了一种自底向上与自顶向下相结合的注意力方法。自底向上的机制（基于Faster R-CNN）用于提取图像中的显著区域，自顶向下的机制利用语言输入学习区域特征对应的权重。之后，基于区域的特征超过了网格特征，成为了视觉问答的标准。Jiang等[12]则重新审视了VQA的网格特征，发现准确度不输区域特征，导致区域特征效果很好的主要原因是大规模标注和高分辨率。然而，在许多情况下，仅在图像和文本问题上进行推理难以得到正确的答案。Wen等[13]提出了一种多层次知识转移网络，从不同角度捕获外部语言知识来支持视觉常识推理。Engin等[14]将视频和对话（场景和片段）转化为本文描述，转换后的输入与问题和每个答案一起独立处理，为每个答案产生一个分数。

在多模态情感分析中，除了上述不足，还涉及其他一些问题。在艺术摄影中，摄影师通过调节图像的饱和度和色调来传达不同的情绪，给人不同的感受。如图1所示，我们通过横向和纵向对比发现，表达不同情感的图像在饱和度、色调和亮度上都存在差异，这是被忽略的。而且，现有的研究大多采用深层语义特征，忽略了多模态信息的不同层次特征的融合。此外，利用Word2Vec[15]或GloVe[16]作为预训练模型，不考虑上下文，无法处理一词多义的问题。

针对上述不足，我们提出了一种基于自适应门控信息融合的多模态情感分析模型。我们先通过门控信息融合网络将Swin Transformer[17]和ResNet[18]提取的视觉和色彩的不同层次特征进行加权融合[19]，充分利用不同模态不同层次的特征对情感分析的贡献。其次，由于情感的复杂性，我们通过迭代注意定位与情感相关的多个细微的局部区域，解决了复杂场景情感区域定位不完整的问题[20]。针对Word2Vec和GloVe的不足，我们利用ERNIE[21]预训练模型，该模型将词汇结构、语法结构和语义信息进行统一建模，极大地增强了通用语义表示能力。最后，利用自动融合网络解决了拼接或TFN无法有效利用多模态信息的问题，该网络通过最大化各模态之间的相关性来充分保留各个模态的有用信息。

本文的贡献总结如下：

(1)提出了一种不同模态不同层次的特征融合方法。该方法利用门控信息融合网络将视觉和色彩的不同层次特征根据对情感分析的贡献进行加权融合，充分利用了各层次特征

对情感分析的影响。

(2)由于图像的抽象性和复杂性,图像的情感往往由多个细微的局部特征所体现。我们通过迭代注意可以更加准确和完整地定位多个与情感相关的判别区域。

(3)提出了一个全新的多模态情感分析框架,在3个公开的真实数据集上进行了大量的实验,证明了该模型的有效性。

2 本文方法

本文提出的多模态情感分析框架如图2所示,该框架主要由两部分组成:视觉流和文本流。在视觉流中,我们先使用Swin Transformer提取RGB图像的图像特征,将最后一层的输出作为全局深度表示,再利用ResNet提取HLS图像的色彩特征,重点研究了饱和度、色调和亮度对情感分析的影响。然后,我们通过门控信息融合网络将不同层次的图像特征和颜色特征进行加权融合,并利用迭代注意来定位重要的局部区域特征。最后,通过门单元学习全局特征和局部特征对情感分析的不同贡献进行加权融合。在文本流中,我们先使用ERNIE预训练模型生成词向量,再通过BiLSTM和CNN分别提取与文本上下文相关的语义信息和局部特征,最后利用自动融合网络对各个模态的特征进行"动态"融合。

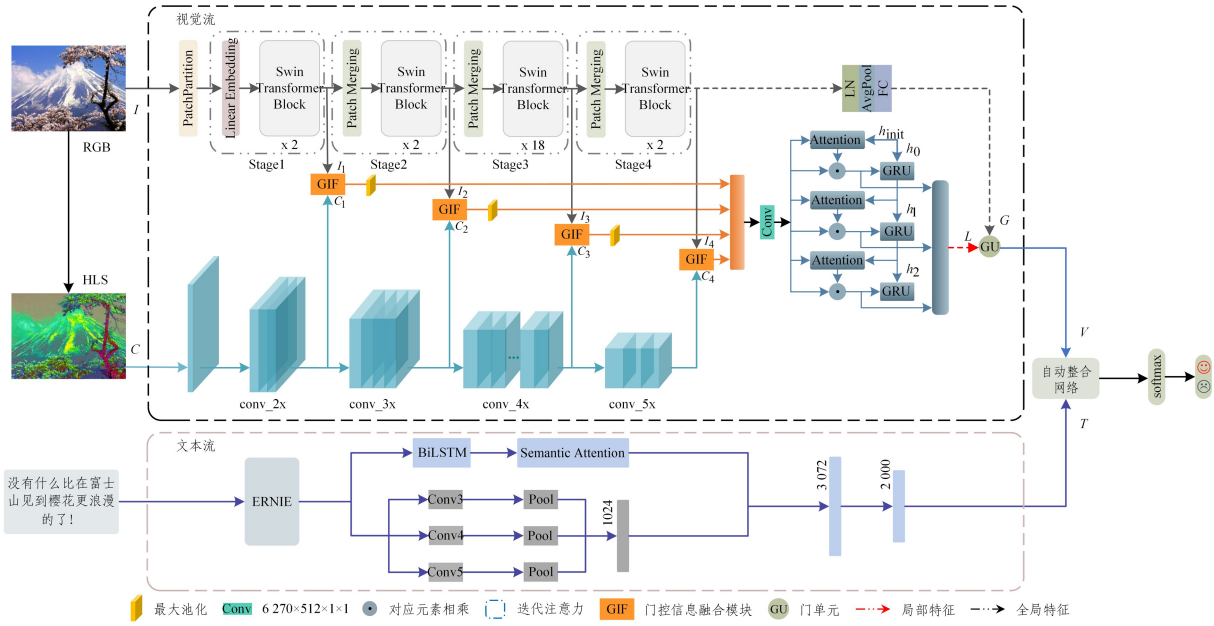


图2 网络框架

Fig. 2 Network framework

2.1 视觉流

2.1.1 全局特征

为了分析图像的情感,我们需要提取整幅图像的全局深度表示。设 $I=\{I_1,I_2,\cdots,I_i,\cdots,I_n\}$ 代表 n 个图像的集合。对于每一幅图像 I_i,我们采用在ImageNet上预训练的Swin Transformer网络来获取全局特征。Swin Transformer引入了CNN中常用的层次化构建方式和滑动窗操作,在性能上已经超过了CNN。我们将网络最后一层的输出作为全局特征表示 G。

2.1.2 局部特征

(1)多层次特征融合

随着网络的加深,特征具有非常大的感受野。感受野越大,特征表示也就越抽象。为了获得尽可能多的图像细节信息,特征需要有较小的区域感受野。一般来说,浅层网络对应较小的感受野。对此,我们利用不同层次特征的互补性来提升情感分析的性能。同时,不同模态的不同层次特征对情感分析的贡献也不相同,我们通过门控信息融合网络将视觉和色彩的不同层次特征进行自适应加权融合。

总体框架如图2所示。我们先使用Swin Transformer提取RGB图像中间层(stage1—stage4)视觉特征,再利用ResNet提取HLS图像中间层(conv2_x-conv5_x)色彩特征。然后,通过门控信息融合网络将stage1—stage4层的特征分别与conv2_x-conv5_x层的特征进行特征级信息融合,如图3所示。该网络采用SENet的思想[22],通过建模通道之间的关系进行加权。该过程主要由两部分组成:权重生成部分和特征融合部分。设 I_1 和 C_1 为两个输入模态对应的 $M\times N\times K$ 的特征图。在权重生成部分中,连接特征 I_1 和 C_1 之前,先进行全局平均池化,将全局空间信息压缩到一个通道描述中,然后输入到全连接层,以捕获通道方面的依赖关系,再通过Sigmoid激活函数去调节学习到的权值。最后,将加权后的特征连接在一起。门控信息融合模块的操作总结如下:

$$u_j=\frac{1}{M\times N}\sum_{m=1}^{M}\sum_{n=1}^{N}i_j(m,n) \tag{1}$$

$$s_j=\frac{1}{M\times N}\sum_{m=1}^{M}\sum_{n=1}^{N}c_j(m,n) \tag{2}$$

$$Z=U\oplus S \tag{3}$$

$$\boldsymbol{w}_1=\sigma(f_1(Z,W))=\sigma(W_2\delta(W_1Z)) \tag{4}$$

$$\boldsymbol{w}_2=\sigma(f_2(Z,W))=\sigma(W_4\delta(W_3Z)) \tag{5}$$

$$R_1=(I_1\odot\boldsymbol{w}_1)\oplus(C_1\odot\boldsymbol{w}_2) \tag{6}$$

其中,i_j 和 c_j 分别是特征图 I_1 和 C_1 第 j 通道的特征,u_j 和 s_j 是全局平均池化后的特征,R_1 是第一层融合特征;σ 是Sigmoid函数,δ 是RELU函数,$\oplus$ 表示特征拼接,$\odot$ 表示对应元素相乘,$\boldsymbol{w}_*$ 是权值矩阵,W_* 是全连接层参数。

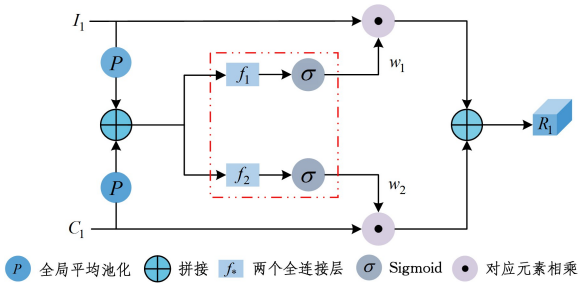


图 3　门控信息融合网络

Fig. 3　Gated information fusion network

(2)迭代注意

人类在情感分析过程中，并非将目光放在整张图像中，而是按照某种次序在图像上进行扫描，从一个区域转移到另一个区域，然后将这些区域的信息相结合，用于整体的判断和感受。GRU 也用类似的方法在每次迭代中生成注意图，结合这些注意表示产生视觉特征，从而进行判断。如图 4 所示，在每一步中，模型根据过去的信息选择下一个位置。

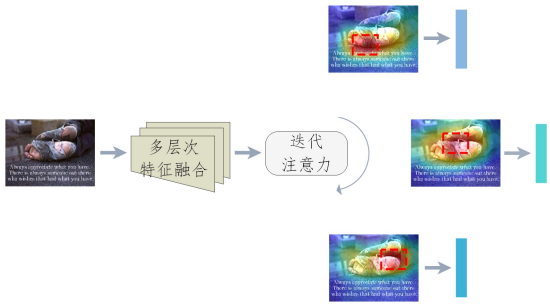


图 4　迭代注意结构

Fig. 4　Iterative attention structure

在获得多层次融合特征之后，我们聚焦于关键的情感区域来生成图像表示，如图 5 所示。迭代注意是通过 GRU[23] 网络计算的，它有一个门控可以进行遗忘和选择记忆。我们结合 GRU 网络隐藏层状态 h_{t-1} 和多层次融合特征 R 来模拟人类的感知过程，得到预测的注意图。

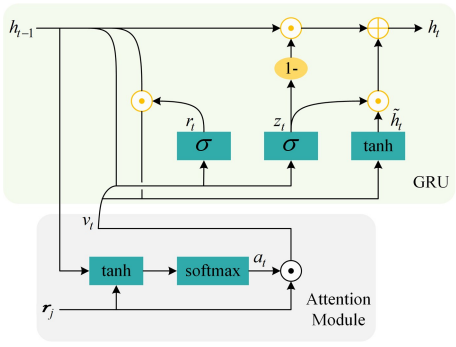


图 5　GRU 单元和注意力模块结构

Fig. 5　Structure of GRU unit and attention module

$$\boldsymbol{e}_{t,j}=N_j\tanh(\boldsymbol{W}_1\boldsymbol{h}_{t-1}+\boldsymbol{W}_2\boldsymbol{r}_j+\boldsymbol{b}) \tag{7}$$

$$\boldsymbol{a}_{t,j}=\frac{\exp(\boldsymbol{e}_{t,j})}{\sum_{k=1}^{L}\exp(\boldsymbol{e}_{t,k})} \tag{8}$$

其中，$\boldsymbol{r}_j$ 是 R 中的第 j 个特征向量，N_j，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是可训练参数，$\boldsymbol{a}_{t,j}$ 是注意权值。将注意权值与输入特征相乘，计算出新的视觉特征。

$$\boldsymbol{v}_t=\sum_{j=1}^{P}\boldsymbol{a}_{t,j}\boldsymbol{r}_j \tag{9}$$

$$\boldsymbol{a}_t=\{\boldsymbol{a}_{t,1},\boldsymbol{a}_{t,2},\cdots,\boldsymbol{a}_{t,P}\},t\in 1,\cdots,M \tag{10}$$

其中，P 是多层次融合特征的位置数量，$\boldsymbol{a}_t$ 是注意权值矩阵，M 是迭代注意的次数。局部特征是这些注意表征的拼接，即 $L=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_M]$。

2.1.3　门单元

全局特征和局部特征可能对情感分析有不同的贡献，本文采用一个门单元[24]来结合全局特征 G 和局部特征 L。门单元的结构如图 6 所示。池化层分别计算 G 和 L 的最大值、最小值、平均值和标准差。然后，将合并的特征输入到全连接层、Tanh 层、全连接层和 Sigmoid 层中，输出两个权重因子 α 和 β。将 α 和 β 分别乘以每个特征元素，合并得到最终的视觉特征 V。

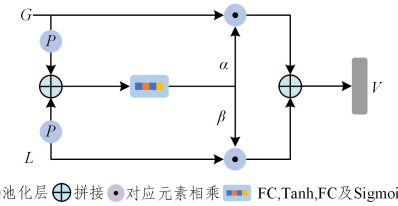


图 6　门单元

Fig. 6　Gate unit

2.2　文本流

文本处理模块如图 2 所示，我们提出了一个基于注意力的 BiLSTM 和 CNN 网络模型。该模型不仅能够关注到句子中对情感分类重要的词语，而且结合了 BiLSTM 网络提取与文本上下文相关的语义信息和 CNN 提取文本局部特征的优势，提高了模型的文本特征提取能力。

在 BiLSTM 模型中，我们利用 ERNIE 预训练模型将文本中的单词编码为一个 768 维的词向量，输入到双向 LSTM 网络中，将前向隐藏层和后向隐藏层的输出视为文本表示，然后利用语义注意力机制突出对情感分类重要的词语。在 CNN 模型中，通过几个不同大小的过滤器提取不同的特征，然后对这些特征进行最大池化操作，以捕获重要的特征。最后，将这两种模型所提取的特征融合得到文本特征 T。

2.3　自动融合网络

为了取得更好的效果，出色的融合机制是必不可少的。以往的大多数融合技术(如拼接和 TFN[25])都以确定性的操作构建多模态联合表示。由于没有特定的学习过程，模型无法有效地利用多模态信息。为了缓解现有融合方法的静态性，我们采用了一种新的融合技术——自动融合网络，如图 7 所示。先将特征向量连接在一起得到 $\boldsymbol{P}_i^k$，再通过全连接层 f 输出“自动融合”向量 $\boldsymbol{O}_i^t$。然后，将自动融合向量通过另一个全连接层 f^{-1} 得到重构向量 $\hat{\boldsymbol{P}}_i^k$。最后，对 $\hat{\boldsymbol{P}}_i^k$ 和 $\boldsymbol{P}_i^k$ 之间的损失进行优化。

自动融合网络采用均方误差(MSE)作为损失函数，这符合压缩多模态特征的动机：过滤掉无用信息。自动融合网络的 MSE 损失为：

$$L_{\mathrm{MSE}}=\|\hat{\boldsymbol{P}}_i^k-\boldsymbol{P}_i^k\|^2 \tag{11}$$

将“自动融合”向量 $\boldsymbol{O}_i^t$ 作为融合的多模态表示，送入全连接层进行情感分类。

$$Y_i = f(\boldsymbol{O}_i^t;\theta), Y_i \in \mathbb{R}^N \tag{12}$$

其中，t 是自动融合向量的特征维度，θ 是全连接层 f 的参数，N 表示类别数。对于融合向量，采用交叉熵损失函数计算损失。

$$L_{CROSS} = [y \log \hat{y} + (1-y)\log(1-\hat{y})] \tag{13}$$

其中，y 表示真实的情感标签，$\hat{y}$ 表示预测的情感标签。那么，总的损失函数为：

$$L = L_{CROSS} + L_{MSE} \tag{14}$$

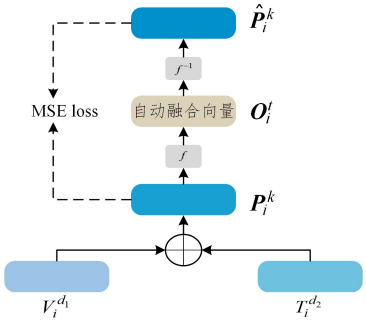


图 7 自动融合网络

Fig. 7 Auto-fusion network

3 实验结果和分析

3.1 数据集

本文基于 MVSA 和 ArtPhoto 数据集进行实验。MVSA 是一个公开的图片-英文数据集，数据集根据标注者人数的不同分为 MVSA-Single 和 MVSA_Multiple 两个子数据集。ArtPhoto 数据集是我们手工添加文本标注所建立的图片-中文数据集。

(1)MVSA 数据集[26]。为了确保数据集的质量，我们对数据集进行了预处理。首先删除了同时包含消极和积极的图像-文本对；其次，一个标签为中性，另一个标签为积极(或消极)，我们选择积极(或消极)作为样本标签。对于 MVSA_Multiple 数据集，当 3 个标注者中至少两个给出了相同的标记时，该样本才被认为是有效的。最后，将包含 4 511 个图文对的 MVSA-Single 数据集和包含 17 024 个图文对的 MVSA_Multiple 数据集作为实验数据。

(2)ArtPhoto 数据集[27]。该数据集是从艺术网站抓取的 806 张照片，被分为 8 个情感类别，即“高兴”“兴奋”“满足”“敬畏”“厌恶”“愤怒”“恐惧”和“悲伤”。拍摄者利用位置、光线和色彩等来唤醒观察者的某种情感。我们通过 5 位研究员对数据集进行文本标注。

3.2 实验设置

在训练阶段，我们将每张图像的大小调整为 260×260，然后随机裁剪为 224×224 的子图像。同时，通过随机水平翻转和归一化来扩展数据集的规模。在测试阶段，将每张图像缩放到 260×260，从中心裁剪到 224×224，再执行归一化。

在文本内容方面，利用最新的 ERNIE 预训练模型将单词转换为 768 维的词向量。其中，文本的最大输入长度为 32，不足 32 词的数据采用 0 填充，超过的则截断。

我们采用 AdamW 优化器来优化参数。其初始学习率设置为 0.001，每迭代 10 次降低为原来的 1/10。矩估计的指数衰减速率分别为 0.9 和 0.99。在 MVSA 和 ArtPhoto 数据集上批量大小分别为 16 和 32，权重衰减设置为 0.01。对于每个数据集，随机选择 80% 作为训练集，20% 作为测试集。我们的网络框架是在 V100s 上完成的，并选用 Windows，CUDA10.1，cudnn7.5 及 PyTorch1.6 来搭建我们的实验平台。

3.3 对比方法

本文将所提方法与现有的情感分析方法进行了比较。基于文本情感分析的方法包括 Att-BiLSTM(Word2Vec)，CNN+Att-BiLSTM(Word2-Vec)和 CNN+Att-BiLSTM(BERT)；基于图像情感分析的方法包括 VGG19[28]，ResNet101[18]，SentiNet-A[29]和 DA-MLCNN[30]；基于多模态情感分析的方法包括 MultiSentiNet[5]，Hu[31]，Co-Memory-M[8]，DMAF[6]，ANNM[7]，MVAN-M[10]和 MLSA[32]。

3.4 实验结果

3.4.1 在 MVSA 数据集上的实验结果

表 1 列出了不同方法在 MVSA 数据集上的实验结果。F1 和准确率两个评价指标的结果一致表明，本文方法在单模态和多模态情感分析方面均优于其他先进的情感分析方法。我们以 MVSA_Single 数据集为例进行实验分析。

在文本情感分析中，CNN+Att-BiLSTM-(Word2Vec)相比 Att-BiLSTM(Word2Vec)准确率提高了 0.93%，这说明结合 CNN 所提取的局部特征有效地提高了文本特征的提取能力。本文采用 CNN+Att-BiLSTM(ERNIE)，准确率为 70.02%，相比 CNN+Att-BiLSTM(BERT)和 CNN+Att-BiLSTM(Word2Vec)分别提高了 1.27%和 3.93%。ERNIE 通过对词、实体等语义单元 mask，结合词法结构、语法结构和语义信息进行统一建模，增强了通用语义表示能力。

在图像情感分析中，本文的 AGIF-V 的准确率为 68.17%，比 DA-MLCNN 提高了 3.97%。相比此方法，我们利用 Swin Transformer 网络来提取视觉特征，它引入了 CNN 的层次化构建方式和滑动窗操作，在性能上已经超越了 CNN。同时，我们考虑了视觉和色彩的不同层次的特征融合，并通过迭代注意能更准确和完整地定位到多个细微的情感区域。

在多模态情感分析中，我们提出的 AGIF(Auto-Fusion)方法也优于其他的情感分析方法。AGIF(Auto-Fusion)的准确率为 74.88%，与 MultiSentiNet，Hu，Co-Memory-M，DMAF，ANNM，MVAN-M 和 MLSA 相比分别提高了 5.04%，6.56%，4.37%，7.41%，5.7%，1.9%和 4.99%。相比其他多模态情感分析方法，我们考虑了视觉和色彩的各个层次特征对情感分析的影响，并通过门控信息融合网络对不同层次特征根据对情感分析的贡献进行加权融合。此外，这些方法的融合技术(如拼接和 TFN)都以确定性的操作来构建多模态联合表示，不能有效地利用多模态信息。我们利用自动融合网络对各模态的特征进行融合，通过最大化各模态之间的相关性来提取多模态特征，这一过程确保学习到的自动融合向量充分保留了各个模态的有用信息。结果表明，AGIF(Auto-Fusion)比 AGIF(Concat)的准确率提高了 0.69%。

表 1　对比方法和本文方法在 MVSA 测试集上的实验结果

Table 1　Results of compared methods and our approaches on MVSA testing dataset

（单位：%）

类型	模型	MVSA_Single		MVSA_Multiple	
		*F*1	*Accuracy*	*F*1	*Accuracy*
文本	Att-BiLSTM(Word2vec)	62.69	65.16	65.08	67.22
	CNN＋Att-BiLSTM (Word2vec)	64.58	66.09	65.68	68.08
	CNN＋Att-BiLSTM (BERT)	67.10	68.75	69.50	70.02
	CNN＋Att-BiLSTM (ERNIE)	**69.04**	**70.02**	**69.51**	**70.40**
图像	Vgg19	60.70	62.76	61.89	64.06
	Resnet101	63.53	64.06	64.25	64.76
	Swin Transformer	64.69	65.39	64.86	65.57
	DA-MLCNN	63.10	64.20	63.75	64.03
	AGIF-V	**66.96**	**68.17**	**66.13**	**67.77**
图像＋文本	MultiSentiNet	69.63	69.84	68.11	68.86
	Hu	66.58	68.32	65.16	68.43
	Co-Memory-M	70.01	70.51	69.83	69.92
	DMAF	65.94	67.47	65.82	67.88
	ANNM	68.64	69.18	65.57	68.95
	MVAN-M	72.98	72.98	72.30	72.36
	MLSA	68.84	69.89	68.12	69.07
	AGIF(Concat)	**73.64**	**74.19**	**71.43**	**74.22**
	AGIF(Auto-Fusion)	**74.40**	**74.88**	**71.16**	**74.65**

3.4.2　在 ArtPhoto 数据集上的实验结果

我们在 ArtPhoto 数据集上取得了不错的结果（见表 2），均优于其他的情感分析方法。

在文本情感分析方面，本文使用的 CNN＋Att-BiLSTM (ERNIE)模型相比 CNN＋Att-BiLSTM(BERT)，CNN＋Att-BiLSTM(Word2Vec)和 Att-BiLSTM(Word2Vec)分别提高了 0.63%，5.63%和 6.26%。在图像情感分析方面，与 SentiNet-A 和 DA-MLCNN 相比，AGIF-V 的准确率提高了 18.35%和 11.25%。在多模态情感分析中，AGIF(Auto-Fusion)比 MultiSentiNet，Hu，DMAF 和 ANNM 的准确率分别提高了 8.12%，6.87%，7.49%和 6.25%。

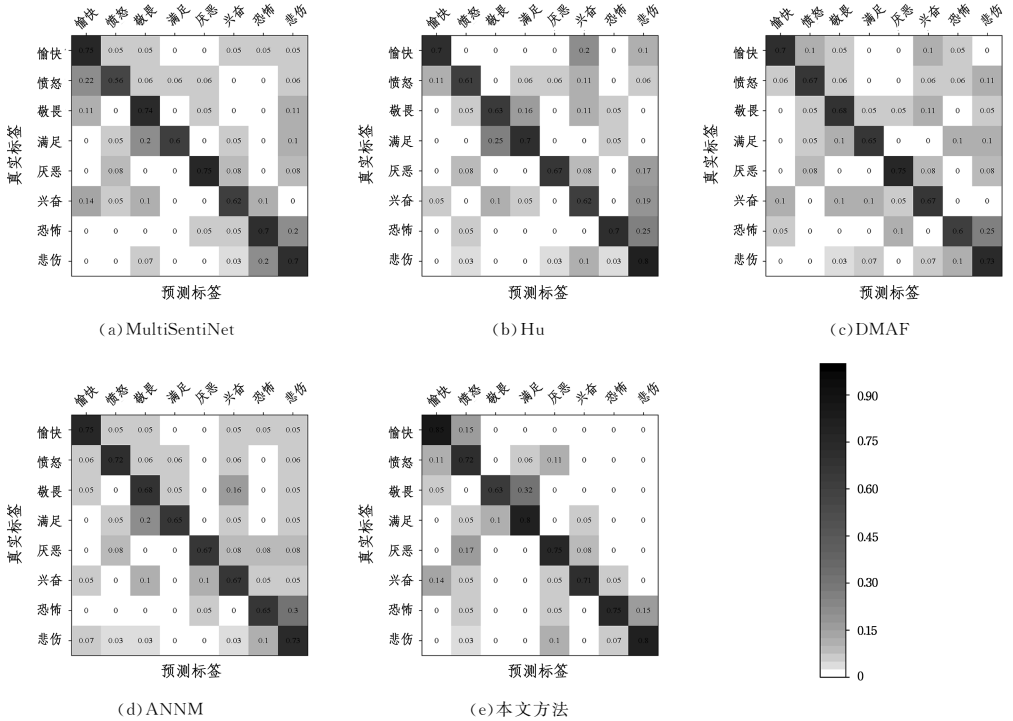
表 2　对比方法和本文方法在 ArtPhoto 测试集上的实验结果

Table 2　Results of compared methods and our approaches on ArtPhoto testing dataset

（单位：%）

类型	模型	*F*1	*Accuracy*
文本	Att-BiLSTM(Word2vec)	65.48	65.62
	CNN＋Att-BiLSTM(Word2vec)	66.30	66.25
	CNN＋Att-BiLSTM(BERT)	71.07	71.25
	CNN＋Att-BiLSTM(ERNIE)	**71.37**	**71.88**
图像	Vgg19	40.99	41.88
	Resnet101	41.32	42.50
	Swin Transformer	46.87	46.88
	SentiNet-A	—	35.40
	DA-MLCNN	41.84	42.50
	AGIF-V	**52.88**	**53.75**
图像＋文本	MultiSentiNet	67.59	67.50
	Hu	69.00	68.75
	DMAF	68.13	68.13
	ANNM	69.54	69.37
	AGIF(Concat)	**74.84**	**75.00**
	AGIF(Auto-Fusion)	**75.54**	**75.62**

图 8 给出了 ArtPhoto 数据集上各种方法的混淆矩阵。与其他方法相比，除了“敬畏”情感的预测，其他类型的情感预测均取得了最好的结果。此外，其他方法的预测标签非常散乱，而本文方法则比较集中，这充分说明了该模型的优越性和鲁棒性。



(a) MultiSentiNet　(b) Hu　(c) DMAF

(d) ANNM　(e) 本文方法

图 8　不同方法在 ArtPhoto 数据集上的混淆矩阵

Fig. 8　Confusion matrix of different methods on ArtPhoto dataset

为了更具说服力，我们将对比方法的图像处理部分和文本处理部分中可以替换的网络进行了替换，换成了相同的 Swin Transformer 网络和 ERNIE 预训练模型，在 ArtPhoto 数据集上进行了重新训练，如表 3 所列。可以发现，在统一的网络和预训练模型下，本文方法仍取得了最好的效果，这充分证明了模型的有效性和优越性。除了更强的视觉网络和预训练模型对情感分析的提升外，我们的各种信息融合方法对情感分析的性能提升也具有一定的贡献，消融实验也证实了这一点。

表 3 不同方法在统一网络和预训练模型下的实验结果

Table 3 Experimental results of different methods with unified network and pre-trained models

（单位：%）

模型	*F*1	*Accuracy*
MultiSentiNet	72.39	72.50
Hu	74.23	74.38
DMAF	72.82	73.13
ANNM	73.56	73.75
Our Method	**75.54**	**75.62**

表 2 对比方法和本文方法在 ArtPhoto 测试集上的实验结果

Table 2 Results of compared methods and our approaches on ArtPhoto testing dataset

（单位：%）

类型	模型	*F*1	*Accuracy*
文本	Att-BiLSTM(Word2vec)	65.48	65.62
	CNN+Att-BiLSTM(Word2vec)	66.30	66.25
	CNN+Att-BiLSTM(BERT)	71.07	71.25
	CNN+Att-BiLSTM(ERNIE)	**71.37**	**71.88**
图像	Vgg19	40.99	41.88
	Resnet101	41.32	42.50
	Swin Transformer	46.87	46.88
	SentiNet-A	—	35.40
	DA-MLCNN	41.84	42.50
	AGIF-V	**52.88**	**53.75**
图像+文本	MultiSentiNet	67.59	67.50
	Hu	69.00	68.75
	DMAF	68.13	68.13
	ANNM	69.54	69.37
	AGIF(Concat)	**74.84**	**75.00**
	AGIF(Auto-Fusion)	**75.54**	**75.62**

3.5 参数分析

3.5.1 特征融合位置数量分析

在特征融合位置数量实验中，我们剔除了其他一些模块，只保留了 Swin Transformer 网络和 ResNet 网络提取视觉特征和色彩特征这两部分。

本文通过门控信息融合网络将 stage1－stage4 层的视觉特征分别与 conv2_x-conv5_x 层的色彩特征进行了自适应融合。为了说明综合各个位置的特征融合是最佳的，我们进行了组合实验，如表 4 所列。可以发现，只采用第四个位置的融合特征准确率就达到了 66.32%，这说明最深层次的特征对情感分析的贡献是最大的。再加入其他位置的融合特征时，准确率也都有所提升，因此我们综合考虑了各个位置的融合特征。

表 4 不同位置的融合特征在 MVSA_Single 测试集上的实验结果

Table 4 Experimental results of fusion features at different locations on MVSA_Single testing dataset

（单位：%）

组合	1	2	3	4	*F*1	*Accuracy*
1	√				47.08	57.52
2		√			57.05	60.76
3			√		59.88	62.15
4				√	65.39	66.32
5	√	√			57.17	60.76
6	√		√		59.55	62.27
7	√			√	65.52	66.32
8		√	√		59.83	62.04
9		√		√	65.38	66.44
10			√	√	65.29	66.44
11	√	√	√		60.78	63.19
12	√	√		√	65.24	66.55
13	√		√	√	65.18	66.67
14		√	√	√	65.20	66.67
15	√	√	√	√	**65.56**	**66.78**

3.5.2 迭代次数

在迭代注意力的次数分析中，我们剔除了其他一些模块，只保留了融合的视觉特征和色彩特征以及迭代注意力这部分。

迭代次数 T 表示注意力执行的次数，这是一个重要的参数，因为它控制着模型识别的准确度和复杂度。本文在 MVSA_Single 数据集上尝试了 1－5 这 5 个参数，如图 9 所示。从图中可以发现，随着迭代次数的增加，情感识别的准确率也随之增加。当 $T=3$ 时，准确率达到了最大值。随着 T 的继续增加，准确率开始下降。因为随着迭代次数的增加，关注的区域越来越多，就可能定位到一些无关的情感区域，导致信息冗余问题，从而引起识别准确率的下降。

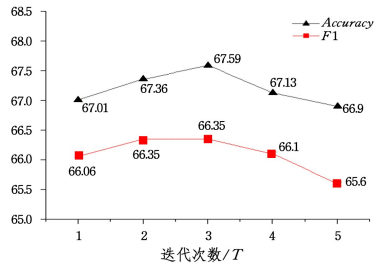


图 9 不同迭代次数在 MVSA_Single 测试集上的实验结果

Fig. 9 Experimental results on MVSA_Single testing dataset with different number of iterations

3.6 消融实验

本节通过消融实验验证了我们引入的色彩特征(CF)、多层次特征融合(MFF)、局部特征(LF)、全局特征(GF)、门单元(GU)、文本特征(TF)和自动融合网络(AFN)的有效性。具体来说，我们逐个添加上述模块来进行重新训练。

MVSA_Single 数据集上的消融实验的结果如表 5 所列。我们的基础模型(BM)是利用 Swin Transformer 提取特征，通过全连接层进行分类，分类准确率为 65.39%。加入色彩特征之后，准确率提高了 0.7%，说明了饱和度和色调对情感分析的有效性。引入多层次特征融合之后，准确率提高了 0.69%。通过自适应融合视觉和色彩的不同层次特征，不仅可以捕捉图像的细节信息，还可以更好地对不同模态的特征

进行加权融合。加入局部视觉特征之后，准确率提高了0.81%。由于情感的复杂性，图像的情感一般由多个细微的局部特征来体现。通过迭代注意准确地定位这些情感判别区域，可以更好地进行情感分析。加入全局视觉特征之后，准确率提高了0.23%，说明图像的情感不仅体现在局部特征上，也要从全局视角去把握。加入门单元之后，准确率提高了0.35%，我们通过学习全局特征和局部特征对情感表达的不同贡献来进行加权融合，更有利于情感分析。加入文本特征以实现多模态情感分析，与图像和文本情感分析相比，准确率提高了6.02%和4.17%。与单模态情感分析相比，多模态情感分析利用视觉内容和文本描述之间的互补信息，提高了情感分析的准确性。最后，我们加入了一个自动融合网络来实现各模态的"动态"融合，与拼接融合相比，本文方法的准确率提高了0.69%。

表5　在MVSA_Single数据集上的消融实验

Table 5　Accuracy of ablation experiment on MVSA_Single

（单位：%）

Model	*F*1	*Accuracy*
BM	64.69	65.39
BM+CF	64.95	66.09
BM+CF +MFF	65.56	66.78
BM+CF +MFF+LF	66.35	67.59
BM+CF +MFF+LF+GF	66.64	67.82
BM+CF +MFF+LF+GF+GU	66.96	68.17
BM+CF +MFF+LF+GF+GU+TF	73.64	74.19
BM+CF+ MFF+LF+GF+GU+TF+AFN	74.40	74.88

结束语　本文提出了一种基于自适应门控信息融合网络的多模态情感分析。首先，通过门控信息融合网络将提取的不同层次的视觉和色彩特征进行加权融合，充分利用不同模态不同层次的特征信息对情感分析的影响。其次，由于情感的复杂性，我们通过迭代注意来定位与情感相关的多个细微的判别区域，解决了对复杂场景定位不准确和不完整的问题。最后，通过一个门单元依据对情感分析的贡献来结合全局特征和多区域局部特征。针对Word2Vec和GloVe的不足，我们采用ERNIE预训练模型，将词汇结构、语法结构和语义信息进行统一建模，极大地增强了通用语义表示能力。然后，利用基于注意力的BiLSTM和CNN提取与文本上下文相关的语义信息和局部特征。最后，利用自动融合网络解决"静态"融合带来的信息冗余问题。实验结果表明，该方法在3个真实数据集上均优于其他情感分析方法。在未来工作中，我们将考虑多尺度特征对情感分析的影响。

参 考 文 献

[1] KAGAN V，STEVENS A，SUBRAHMANIAN V S. Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election [J]. IEEE Intelligent Systems，2015，30(1)：2-5.

[2] BOLLEN J，MAO H N，ZENG S J. Twitter mood predicts the stock market [J]. Journal of Computational Science，2011，2(1)：1-8.

[3] LI X D，XIE H R，CHEN L，et al. News impact on stock price return via sentiment analysis [J]. Knowledge-Based Systems，2014，69(15)：14-23.

[4] HUR M，KANG P，CHO S. Box-office forecasting based on sentiments of movie reviews and Independent subspace method [J]. Information Sciences，2016：608-624.

[5] XU N，MAO W J. MultiSentiNet：A Deep Semantic Network for Multimodal Sentiment Analysis[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017：2399-2402.

[6] HUANG F R，ZHANG X M，ZHAO Z H，et al. Image-text sentiment analysis via deep multimodal attentive fusion [J]. Knowledge-Based Systems，2019：167：26-37.

[7] LIN M H，MENG Z Q. Multimodal Sentiment Analysis Based on Attention Neural Network [J]. Computer Science，2020，47(S2)：508-514，548.

[8] XU N，MAO W J，CHEN G D. A co-memory network for multimodal sentiment analysis[C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018：929-932.

[9] XU J，HUANG F R，ZHANG X M，et al. Visual-textual sentiment classification with bi-directional multi-level attention networks [J]. Knowledge Based Systems，2019，178(AUG. 15)：61-73.

[10] YANG X C，FENG S，WAND D L，et al. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network [J]. IEEE Transactions on Multimedia，2021，23(1)：4014-4026.

[11] ANDERSON P，HE X D，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.

[12] JIANG H，MISRA I，ROHRBACH M，et al. In defense of grid features for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：10264-10273.

[13] WEN Z，PENG Y. Multi-level knowledge injecting for visual commonsense reasoning [J]. IEEE Transactions on Circuits and Systems for Video Technology，2020，31(3)：1042-1054.

[14] ENGIN D，SCHNITZLER F，DUONG N Q K，et al. On the hidden treasure of dialog in video question answering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：2064-2073.

[15] MIKOLOV T，CORRADO G，KAI C，et al. Efficient Estimation of Word Representations in Vector Space [J]. Advances in Neural Information Processing Systems，2013，26(1)：3111-3119.

[16] PENNINGTON J，SOCHER R，MANNING C D. Glove：Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP，2014：1532-1543.

[17] LIU Z，LIN Y T，CAO Y，et al. Swin transformer：Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE，2021：10012-10022.

[18] 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. IEEE，2016：770-778.

[19] ZHAND X D, GAO X B, LU W, et al. A Gated Peripheral-Foveal Convolutional Neural Network for Unified Image Aesthetic Prediction [J]. IEEE Transactions on Multimedia, 2019, 21(11): 2815-2826.

[20] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention [C]// Proceedings of the Neural Information Processing Systems. 2014: 2204-2212.

[21] SUN Y, WANG S, LI Y, et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 8968-8975.

[22] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-Excitation Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8) 2011-2023.

[23] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [J]. arXiv: : 1412.3555, 2014.

[24] ZHAO L, SHANG M, GAO F, et al. Representation learning of image composition for aesthetic prediction [J]. Computer Vision and Image Understanding, 2020, 199(9): 103024.

[25] ZADEH A, CHEN M, PORIA S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen Denmark, 2017: 1103-1114.

[26] TENG N, ZHU S, LEI P, et al. Sentiment analysis on multi-view social data [C]// International Conference on Multimedia Modeling. 2016: 15-27.

[27] 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, NY, USA, 2010: 83-92.

[28] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. arXiv: 1409.1556, 2014.

[29] SONG K K, YAO T, LING Q, et al. Boosting Image Sentiment Analysis with Visual Attention [J]. Neurocomputing, 2018, 312(27): 218-228.

[30] CAI G Y, CHU Y Y. Visual SentimentAnalysis Based on Multi-level Features Fusion of Dual Attention [J]. Computer Engineering, 2021, 47(9): 227-234.

[31] HU A, FLAXMAN S. Multimodal Sentiment Analysis To Explore the Structure of Emotions [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 350-358.

[32] GUO K X, ZHANG Y X. Visual-textual sentiment analysis method with multi-level spatial attention [J]. Journal of Computer Applications, 2021, 41(10): 2835-2841.

CHEN Zhen, born in 1994, postgraduate. His main research interests include multimodal sentiment analysis and image processing.

ZHAO Zhengpeng, born in 1973, master, associate professor. His main research interests include digital image processing and speech signal processing.

(责任编辑:喻藜)