文章编号:1007-5321(2019)01-0061-07 **DOI**:10.13190/j.jbupt.2018-040

基于多模态判别性嵌入空间的图像情感分析

吕光瑞, 蔡国永, 林煜明
(桂林电子科技大学 广西可信软件重点实验室, 桂林 541004)

摘要: 为了解决图像情感分析中存在的情感鸿沟和大的类内方差问题,提出了一种可以同时利用视觉模态和文本模态之间的深度潜在关联、视觉模态的深度线性判别和图像中层语义融合的弱监督方法. 利用多模态深度网络结构找到一个视觉模态和文本模态之间最大深度关联且视觉模态具有深度判别性的潜在嵌入空间,并在该潜在空间中将文本的语义映射特征迁移到图像的判别性视觉映射特征中;结合注意力机制,设计涵盖潜在空间中映射特征的注意力网络,用于情感分类. 在真实数据集上的实验结果表明,所提出的方法获得了更好的情感分类准确率.

关 键 词: 情感分析; 潜在关联; 线性判别; 多模态网络; 注意力机制
中图分类号: TP391 **文献标志码**: A

Image Sentiment Analysis with Multimodal Discriminative Embedding Space

LÜ Guang-rui, CAI Guo-yong, LIN Yu-ming
(Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: In order to alleviate affective gap and large intra-class variance existing in visual sentiment analysis, firstly a new method is proposed, which exploits simultaneously not only deep latent correlations between visual and textual modalities, but also deep linear discrimination of visual modality and weak supervision of mid-level semantic features of images. The method uses multimodal deep network architecture to find a latent embedding space in which deep correlations between visual and textual modalities are maximized, and at the same time there is a deep discrimination on visual modality. In the latent space, the extracted semantic feature of texts can be transferred to the extracted discriminant visual feature of images. Secondly based on the usfulness of attention mechanism, an attention network is presented, which accepts the extracted features in the latent space as input and is trained as a sentiment classifier. Results of experiments conducted on real datasets show that the proposed approach achieves better sentiment classification accuracy than those state-of-the-art approaches.

Key words: sentiment analysis; latent correlation; linear discrimination; multimodal network; attention mechanism

由于社交网络用户喜欢上传附带短文本或者没有文本的图像,研究者开始致力于从视觉和多模态内容中探测情感. 然而情感的主观性和图像特征与情感语义之间的情感鸿沟问题,使得视觉情感分析是一项极具挑战的任务. 虽然图像标注和搜索上的一些方法有助于缓解语义鸿沟,但在现实应用中获

收稿日期: 2018-03-20
基金项目: 国家自然科学基金项目(61763007,61562014); 广西自然科学基金项目(2017JJD160017); 广西可信软件重点实验室项目(kx201503)
作者简介: 吕光瑞(1989—), 男, 硕士生.
通信作者: 蔡国永(1971—), 男, 教授, 硕士生导师, E-mail:ccgycai@guet.edu.cn.

取大量高质量的有标记的图像代价极高. 因此有些研究尝试从其它辅助源信息中迁移知识到图像分类任务中[1]. 相比于有标签的图像数据,共现数据在社交网站上更容易获取,例如图像及其对应的描述可视为共现数据. 共现数据中的文本描述有益于语义理解,因此图像及其共现的文本协同使用可以帮助图像内容的识别.

然而,视觉情感分析可能涉及图像对象、场景、动作等情感上下文,相同积极/消极的情感可以呈现在不同的物体对象上,从而存在大的情感类内方差. 例如花和鸟在视觉上是不相似的,但是漂亮的花和漂亮的鸟却展示了同样积极的情感. 同样,相同物体对象也可能推断出不同的情感,因此视觉情感分析也需要在相同的对象类中探测细微的情感差别.

为此,本文首先提出了一种多模态深度单重判别性相关分析的方法来映射图像和与之共现文本的深度特征到潜在空间中,在潜在空间中迁移文本的语义信息到图像的判别性视觉特征中以形成多模态判别性嵌入空间;同时利用三分支网络来联合学习形容词名词对(ANP, adjective noun pair)[2-3]中的形容词、名词以及相对应图像,发掘相同形容词或名词下的图像共享特征,然后将判别性嵌入空间中的特征结合注意力机制网络来设计情感分类器.

1 视觉情感分析相关工作

传统的视觉情感分析方法关注于构造人工设定特征来表示图像,但由于情感涉及高层抽象的事实,Borth 等[2]提出利用视觉实体和属性抽取中层视觉特征以克服低层视觉特征和高层情感语义之间的情感鸿沟,他们通过建模 ANP 这样的中层表示构建了视觉情感本体库(VSO, visual sentiment ontology),Jou 等[3]继续扩展这方面的研究,并构建了包含多种语言 ANP 的多语言视觉情感本体库(MVSO, multilingual VSO). 然而传统的方法很难处理大规模数据的伸缩性和泛化性问题,而卷积神经网络(CNN, convolutional neural network)能够自动地从大规模图像数据中学习稳健的特征且展示了优异的性能[4-7]. You 等[5]提出一个自定义的 CNN 结构用于视觉情感分析,并提出渐进式 CNN(PCNN, progressive CNN)的概率采样方法,来减少噪声对训练图像的影响. Campos 等[6]和 Islam 等[7]分别利用预训练权重微调或初始化的迁移学习方法进行图像情感分析. 尽管这些模型取得了较好的效果,然而仅从视觉模态分析情感,没有借助图像共现的其它模态数据来辅助视觉情感分析.

2 方法描述

本文方法的整体模型结构如图 1 所示,图 1(a)中通过 3 个子网络来提取视觉模态的特征:利用图 1(a-1)所示的深度卷积网络(VGG16, visual geometry group)提取图像的特征,图 1(a-2)所示的形容词特征提取网络(A-net)和名词特征提取网络(N-net)分别提取图像对应 ANP 中形容词的描述性特征和名词的客观性特征. 如果仅将图 1(b)中提取的文

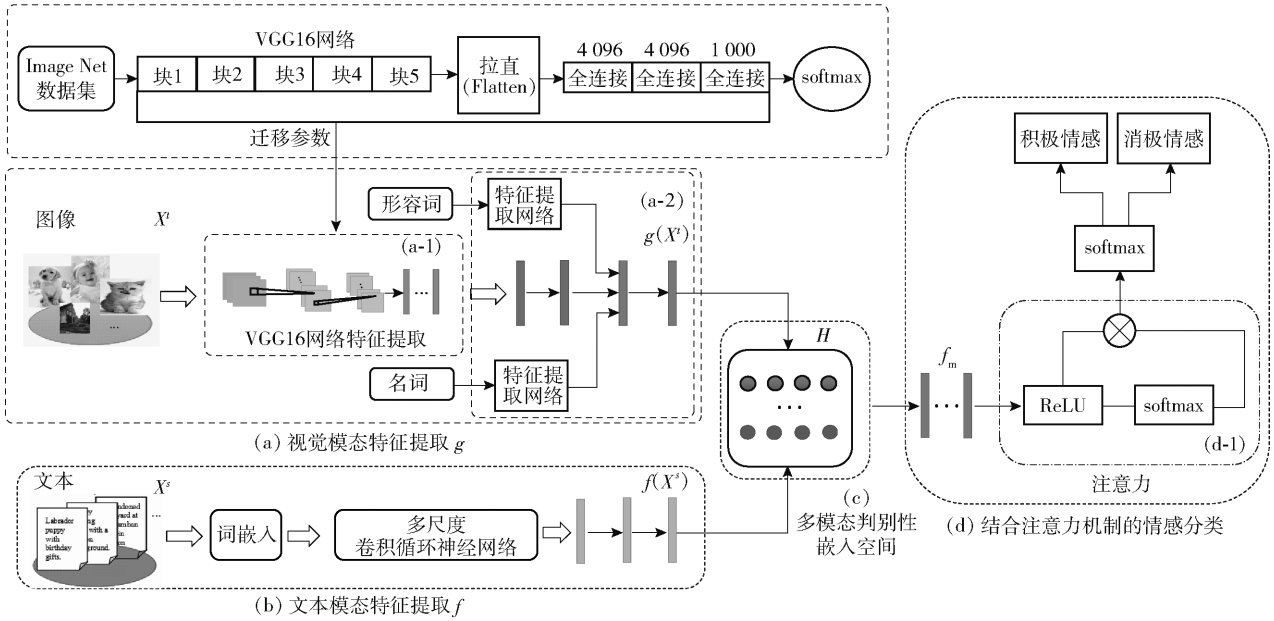

图 1 基于多模态判别性嵌入空间的图像情感分类框架图解

本语义特征在图 1(c)中迁移到图 1(a-1)中仅用 VGG16 网络提取的图像视觉特征中,则后文将其命名为 M1 模型;如将图 1(b)中提取的文本语义特征在图 1(c)中迁移到图 1(a)中 3 个子网络共同提取的图像视觉共享特征中,后文称其为 M2 模型,最后基于图 1(c)中的特征结合图 1(d)的注意力网络调节后输入分类器.

2.1　视觉模态特征提取

提出联合学习图像对应 ANP 中的形容词描述和名词描述以及图像特征的 3 个平行子网络来共同地构建稳健的视觉情感表示,如图 1(a-2)所示. 即视觉模态特征提取网络 g 共包含 ANP 特征提取、图像特征提取以及网络的深度融合,最后形成视觉模态网络的顶层特征表示 $g(\boldsymbol{X}^t)$.

ANP 特征提取　首先,划分每一个图像的 ANP 标签为形容词和名词,利用形容词和名词这两种类型的语义标签作为弱监督学习的图像语义. 针对形容词和名词,CNN 的卷积层用的是二维卷积,每一个形容词或名词样本像单通道图像一样被调整为 50×50 的大小,利用 2 个平行的子网络,即图 1(a-2)所示的 A-net 和 N-net,它们由同样的卷积层和全连接层组成,分别抽取形容词和名词的语义特征.

图像特征提取 利用预训练的 16 层 VGG 网络来提取图像特征映射,如图 1(a-1)所示. 基于 VGG 的图像 CNN 由 5 个卷积块和 3 个全连接层组成,且已经在 1 000 个目标分类的 ImageNet 数据集上表现出了极好的性能. 利用迁移学习的策略来克服 ImageNet 数据集和图像情感数据集的不同差异. 即 VGG16 模型在 ImageNet 的数据集上训练好,然后迁移学好的参数到情感分析的目标中.

2.2　文本模态特征提取

文本模态特征提取网络 f 由多尺度卷积循环神经网络以及全连接神经网络组成,如图 1(b)所示. 其中多尺度卷积循环神经网络由一维卷积和双向长短时记忆网络(Bi-LSTM, bidirectional long short term memory)组成. 一维卷积被用于接收预训练的词向量的输入,卷积层的输出被池化成较小的维度后输入到 Bi-LSTM. 其中卷积层用于提取文本的局部语义特征,Bi-LSTM 从正向和反向的角度来使用已提取的特征. 最后,经过对文本序列建模后,将 Bi-LSTM 的输出传递给全连接神经网络以更好地融合时序特征,形成更容易被区分的高层特征表示. 具体来讲,在一维卷积层分别用了 3 个不同的卷积核(3、4、5)来提取不同语义层次的特征,且对每个卷积核使用了 20 个滤波器. 在句子矩阵上滤波器执行卷积并生成可变长度的特征映射. 在每一个映射上执行滑动长度为 2 的最大池化操作以形成维度较低的序列特征. 然后按顺序合并池化的特征后输入 Bi-LSTM,最后通过全连接层形成高层次的语义特征表示 $f(\boldsymbol{X}^s)$.

2.3　多模态判别性嵌入空间

源领域文本和目标领域图像通过相对应的非线性特征提取网络 f 和 g 生成的顶层特征分别表示为 $f(\boldsymbol{X}^s)\in\mathbf{R}^{N\times L}$ 和 $g(\boldsymbol{X}^t)\in\mathbf{R}^{N\times L}$,设 f 和 g 的学习参数 $(\boldsymbol{W}_l^s;\boldsymbol{b}_l^s)$ 和 $(\boldsymbol{W}_l^t;\boldsymbol{b}_l^t)$ 的集合分别表示为 θ_s 和 θ_t,且设定 $f(\boldsymbol{X}^s)$ 和 $g(\boldsymbol{X}^t)$ 的维度是相同的,记为 L.

该部分融合深度典型相关分析(DCCA, deep canonical correlation analysis)[8] 和深度线性判别分析(DeepLDA, deep linear discriminant analysis)[9] 的做法. DCCA 是典型相关分析(CCA, canonical correlation analysis)的深度网络版,DeepLDA 是将线性判别分析(LDA, linear discriminant analysis)放在深度网络的顶层以学习可以最大化不同类别之间间距的潜在表示.

在 CCA 中,首先通过预处理操作,分别使 $f(\boldsymbol{X}^s)$ 和 $g(\boldsymbol{X}^t)$ 变成中心数据矩阵:

$$\bar{f}(\boldsymbol{X}^s)=f(\boldsymbol{X}^s)-\frac{1}{N}f(\boldsymbol{X}^s)\mathbf{1} \tag{1}$$

$$\bar{g}(\boldsymbol{X}^t)=g(\boldsymbol{X}^t)-\frac{1}{N}g(\boldsymbol{X}^t)\mathbf{1} \tag{2}$$

其中 N 表示数据的总数,$\mathbf{1}\in\mathbf{R}^{N\times N}$ 表示全 1 的矩阵.

源领域文本和目标领域图像的顶层特征表示的正则化自协方差矩阵分别为

$$\boldsymbol{M}_{ss}=\frac{1}{N-1}\bar{f}(\boldsymbol{X}^s)\bar{f}(\boldsymbol{X}^s)^{\mathrm{T}}+r_s\boldsymbol{I} \tag{3}$$

$$\boldsymbol{M}_{tt}=\frac{1}{N-1}\bar{g}(\boldsymbol{X}^t)\bar{g}(\boldsymbol{X}^t)^{\mathrm{T}}+r_t\boldsymbol{I} \tag{4}$$

其中:r_s、r_t 是正则化参数,是为了确保协方差有积极的定义,$\boldsymbol{I}$ 是单位矩阵.

除了领域自身的协方差外,不同领域学习到的特征表示的交叉协方差矩阵表示为

$$\boldsymbol{M}_{st}=\frac{1}{N-1}\bar{f}(\boldsymbol{X}^s)\bar{g}(\boldsymbol{X}^t)^{\mathrm{T}} \tag{5}$$

基于 CCA 中介绍的协方差矩阵 $\boldsymbol{M}_{ss}$、$\boldsymbol{M}_{tt}$ 和 $\boldsymbol{M}_{st}$,定义矩阵 $\boldsymbol{T}=\boldsymbol{M}_{ss}^{-1/2}\boldsymbol{M}_{st}\boldsymbol{M}_{tt}^{-1/2}$. 然后 $f(\boldsymbol{X}^s)$ 和 $g(\boldsymbol{X}^t)$ 的总体关联是通过相对应的奇异值问题 $\boldsymbol{T}=\boldsymbol{U}_s\boldsymbol{\Lambda}\boldsymbol{U}_t$ 和 $\boldsymbol{\Lambda}=\mathrm{diag}(d)$ 中的奇异值 d 的求和来计算. 其中

$\boldsymbol{U}_s$ 和 $\boldsymbol{U}_t$ 是转化文本模态和视觉模态到线性 CCA 子空间的映射矩阵. DCCA 的总体关联是在相对应的网络参数 θ_s 和 θ_t 下最大化奇异值 d 的和:

$$\arg\max_{\theta_s,\theta_t}\sum_{i=1}^{L} d_i \tag{6}$$

设 LDA 中图像的标签属于 C 个不同的类 $c\in\{1,\cdots,C\}$,对于 LDA,$\boldsymbol{M}_{ss}$ 和 $\boldsymbol{M}_{tt}$ 也分别表示总体离散度矩阵. 笔者只考虑 $\boldsymbol{M}_{tt}$ 作为目标领域图像的总体离散度矩阵. 此外,LDA 还需要 C 个不同类别中每个类别的协方差矩阵 $\boldsymbol{M}_{tc}$,以及所有不同类协方差矩阵的均值 $\boldsymbol{M}_{tw}$,即类内离散度矩阵:

$$\boldsymbol{M}_{tc}=\frac{1}{N_c-1}\bar{g}(\boldsymbol{X}_c^t)\bar{g}(\boldsymbol{X}_c^t)^{\mathrm{T}}+r_{tc}\boldsymbol{I} \tag{7}$$

$$\boldsymbol{M}_{tw}=\frac{1}{C}\sum_{c}\boldsymbol{M}_{tc} \tag{8}$$

其中 r_{tc} 是正则化参数,引入它是为了确保协方差有积极的定义.

最后,通过总体离散度矩阵 $\boldsymbol{M}_{tt}$ 和类内离散度矩阵 $\boldsymbol{M}_{tw}$ 来定义类间离散度矩阵 $\boldsymbol{M}_{tg}$:

$$\boldsymbol{M}_{tg}=\boldsymbol{M}_{tt}-\boldsymbol{M}_{tw} \tag{9}$$

通常,LDA 通过找到一个映射矩阵 $\boldsymbol{A}$ 来最大化类间离散度矩阵 $\boldsymbol{M}_{tg}$ 和类内离散度矩阵 $\boldsymbol{M}_{tw}$ 的比值:

$$\arg\max_{\boldsymbol{A}}\frac{|\boldsymbol{A}\boldsymbol{M}_{tg}\boldsymbol{A}^{\mathrm{T}}|}{|\boldsymbol{A}\boldsymbol{M}_{tw}\boldsymbol{A}^{\mathrm{T}}|} \tag{10}$$

映射矩阵 $\boldsymbol{A}$ 转化数据到一个 $C-1$ 维的子空间中,在这个空间中的映射特征变得线性可区分.

DeepLDA 是在深度学习的优化目标中充分利用 LDA 的优异特性,且再形式化表示特征值问题为 $\boldsymbol{M}_{tg}\boldsymbol{e}=v(\boldsymbol{M}_{tw}+\lambda\boldsymbol{I})\boldsymbol{e}$,最终的优化目标关注最大化 k 个最小的特征值 $\{v_1,\cdots,v_k\}$,如

$$\arg\max_{\theta_t}\frac{1}{k}\sum_{i=1}^{k} v_i \tag{11}$$

其中,$\{v_1,\cdots,v_k\}=\{v_j\mid v_j<\min\{v_1,\cdots,v_{C-1}\}+\varepsilon\}$.

综上可知,DCCA 和 DeepLDA 都是基于相对应的特征值问题的特征结构优化的. 其中,DCCA 的优化是以最大化 2 个不同神经网络的隐层输出的相关性为目标来求解矩阵 $\boldsymbol{T}$ 的奇异值;而 DeepLDA 的优化是最大化类别的区分,其由相对应的广义特征值问题的特征值的大小进行量化. 尽管两者的优化有差异,但是它们都反向传播一个由特征值问题引起的误差来调整深度神经网络的参数. 则在多模态学习中可以同时使用 DCCA 和 DeepLDA 的概念,故一个可以同时学习 2 个不同模态之间隐层表示的相关性以及使学到的视觉模态的表示具有判别能力的联合优化目标函数的形式化表示为

$$\arg\max_{\theta_s,\theta_t}\frac{1}{L}\sum_{i=1}^{L} d_i+\frac{1}{k}\sum_{i=1}^{k} v_i \tag{12}$$

经过式(12)这种多模态深度单重判别性相关分析的优化,最后分别通过映射矩阵 $\boldsymbol{U}_s$ 和 $\boldsymbol{U}_t$ 将 $f(\boldsymbol{X}^s)$ 和 $g(\boldsymbol{X}^t)$ 映射到一个共同的潜在空间(如图 1(c)所示),其中映射特征 $\boldsymbol{U}_s^{\mathrm{T}}f(\boldsymbol{X}^s)$ 和 $\boldsymbol{U}_t^{\mathrm{T}}g(\boldsymbol{X}^t)$ 是最大关联的且 $\boldsymbol{U}_t^{\mathrm{T}}g(\boldsymbol{X}^t)$ 是具有判别性的. 此时,源领域文本的语义特征 $\boldsymbol{U}_s^{\mathrm{T}}f(\boldsymbol{X}^s)$ 和目标领域图像的判别性视觉特征 $\boldsymbol{U}_t^{\mathrm{T}}g(\boldsymbol{X}^t)$ 变得接近,则可在潜在空间中将源领域文本的语义信息迁移到目标领域图像中形成多模态判别性嵌入空间.

2.4 结合注意力机制的情感分类

为了获得更好的情感分类效果,利用注意力机制计算注意力概率,注意力概率可以突出特定的特征对整体特征的重要程度. 基于形成的多模态判别性嵌入空间,将空间中的语义增强的判别性视觉特征表示 $\boldsymbol{H}$ 输入到多层全连接神经网络 f_{m} 以进一步提取特征 $f_{\mathrm{m}}(\boldsymbol{H})$,然后将 $f_{\mathrm{m}}(\boldsymbol{H})$ 通过注意力层得到特征表示 $\widetilde{\boldsymbol{h}}_v$,如图 1(d-1)所示注意力层操作的等式为:

$$\boldsymbol{h}_v=\mathrm{ReLU}(\boldsymbol{W}_v f_m(\boldsymbol{H})) \tag{13}$$

$$\boldsymbol{\alpha}=\mathrm{softmax}(\boldsymbol{W}_{p_v}\boldsymbol{h}_v) \tag{14}$$

$$\widetilde{\boldsymbol{h}}_v=f_m(\boldsymbol{H})\boldsymbol{\alpha}^{\mathrm{T}} \tag{15}$$

在获得了注意力层的输出 $\widetilde{\boldsymbol{h}}_v$ 后,通过 softmax 层将 $\widetilde{\boldsymbol{h}}_v$ 分类到输出类别中,整个过程是个端到端的过程,称该过程为 self-attention,如图 1(d)所示. 为了衡量 self-attention 网络的损失,本文使用交叉熵.

最后的 softmax 层解释特征表示 $\widetilde{\boldsymbol{h}}_{v_i}$ 到输出的类别中且分配一个相对应的概率分数 $\boldsymbol{p}_i$. 如果输出的情感类别的数量定义为 m,则

$$\boldsymbol{p}_i=\frac{\exp(\widetilde{\boldsymbol{h}}_{v_i})}{\sum_i\exp(\widetilde{\boldsymbol{h}}_{v_i})},\ i=1,2,\cdots,m \tag{16}$$

$$L=-\sum_i \boldsymbol{t}_i\log(\boldsymbol{p}_i) \tag{17}$$

其中:L 是网络的交叉熵损失,通过反向传播计算网络的梯度. 如果图像的真实标签定义为 $\boldsymbol{t}_i$,则

$$\frac{\partial L}{\partial\widetilde{\boldsymbol{h}}_{v_i}}=\boldsymbol{p}_i-\boldsymbol{t}_i \tag{18}$$

3　实验分析

3.1　数据集与对比方法

实验中总共用到了 5 个数据集，其中 3 个是根据 ANP[2-3] 从社交网络上爬取的，另外 2 个来自于公开数据集[10]. 数据集简介如下：

利用 VSO 中的 3 244 个 ANP[2] 作为关键词从视觉中国（VCG，visual china group）网站上的 Getty 专区爬取 38 363 条数据，称其为 VCG Ⅰ 数据集；同时从 3 244 个 ANP[2] 中随机选出 300 个 ANP 作为关键词从相同网站上爬取 37 158 条数据，称其为 VCG Ⅱ 数据集.

利用 MVSO[3] 中提供的英文语言 ANP，即英文的视觉情感关键词（E-VSK，english-visual sentiment keyword）选取其中情感分数绝对值大于 1 的 ANP 作为关键词从社交网站 Flickr 上爬取 75 516 条图像与其相对应的标题、标签、描述，称其为 E-VSK 数据集.

利用文献[10]中公布的带有积极、中性、消极标注的图像 ID 从社交网站 Flickr 上爬取 6 万余张图像以及相对应的标题、标签、描述，称其为 Flickr 数据集.

对于 VCG 的 2 个数据集，删除那些文本描述是中文的且删除英文描述少于 20 个字符的图像数据；而对于 E-VSK 数据集和 Flickr 数据集，选择那些标签和描述至少有 1 个存在的数据，将筛选过后的数据集中存在的标签、描述、标题组合成文本信息（这里并不是所有的数据均是 3 者都有，但至少有 1 个）. 由于来自于 Flickr 网站的文本信息中含有一些非词汇的内容，则利用 wordnet 删除文本信息中不在 wordnet 中的词汇以生成最终的文本.

VCG 数据集和 E-VSK 数据集的图像情感极性标签来自于 ANP 的情感分数，而 Flickr 数据集的标签来自于人工标注，将至少 2 个人标注为积极的图像的极性标签认为是积极，至少 2 个人标注为中性的图像的极性标签认为是中性，至少 2 个人标注为消极的图像的极性标签认为是消极. 此外，处理后的 Flickr 数据集有 3 万多张积极标签的图像，明显高于消极的和中性的数量. 为了人工构造一个较平衡的数据集，从积极的图像中随机取样一些与消极或中性大致数量相等的数据. 因此得到了实验中要使用的 5 个数据集，其具体信息如表 1 所示.

表 1　最后数据集的统计情况

数据集	积极	中性	消极	总计
VCG Ⅰ	18 847	0	15 837	34 684
VCG Ⅱ	18 134	0	16 184	34 318
E-VSK	35 295	0	24 363	59 658
Flickr-2	12 773	0	10 070	22 843
Flickr-3	12 773	13 518	10 070	36 361

实验中对比了如下几种方法：

1）CNN：具有 2 个卷积层和 4 个全连接层的方法[5].

2）PCNN：逐步概率采样的 CNN[5].

3）VGG-transfer：利用 Islam 等[7] 提出的基于迁移学习的视觉情感分析方法，不同的是实验中利用 VGG16 网络模型.

4）DCCA：利用 Andrew 等[8] 提出的深度典型关联分析方法，不同的是实验中利用所提出的视觉模态和文本模态的网络结构从迁移的角度将文本语义特征嵌入到图像中以生成语义增强视觉特征表示.

5）early-self-attention：2.4 节中 self-attention 模型的变体. 将多模态判别性嵌入空间中的特征表示 $\boldsymbol{H}$ 通过注意力层生成加权的特征表示 $\widetilde{\boldsymbol{H}}$，再将 $\widetilde{\boldsymbol{H}}$ 通过全连接神经网络学习后进行情感分类.

3.2　实验设置

VCG 数据集中图像的文本描述相对正式和简洁，但由于其文本长度普遍较短且长短不一，则选取所用训练集中最长的文本长度为最大长度，不足最大长度的文本用零向量填充. 而 E-VSK 数据集和 Flickr 数据集均来自社交网站 Flickr，不同是获取数据的方式以及图像标签（label）的方法不同. 由于不是所有的图像共现的文本信息中都含有标签（tags）、描述和标题，且文本长度长短不一，故截取最大文本长度为 300，不足最大长度的文本以零向量填充. 每一个词的维度设置为 300，在训练过程中微调词向量来适应本文获取的情感数据集. 在实验中 2 个端到端的过程均使用小批量的 RMSprop 方法来优化网络. 为了防止过拟合，实验中使用 0.5 概率的 dropout 值和 early-stopping 策略. 在 2 个端到端的过程中均使用 ReLU 作为网络层的激活函数.

3.3　实验结果

实验主要评估提出的方法在二分类（积极、消极）和三分类（积极、中性、消极）目标的适用情况.

本文共设计5组实验,每个实验均从各自数据集中随机选取80%用于训练,20%用于测试. 前4组实验分别采用准确率(Accuracy)、召回率、F1值3个评价方法衡量各个方法在VCGⅠ、VCGⅡ、E-VSK、Flickr-2这4个数据集上的情感二分类效果. 第5组实验采用Accuracy的评价方法衡量各个方法在Flickr-3数据集上的情感三分类效果.

所提出的方法分2个阶段进行,第1个阶段是为了形成多模态的嵌入空间,在实验中涉及到DCCA、M1以及M2;第2个阶段是利用2.4节提出的self-attention来学习嵌入空间中的特征以训练情感分类器. 为了评估self-attention方法的合理性,比较其与early-self-attention的性能差异,在5个数据集上的实验均显示self-attention相比于early-self-attention取得了更好的情感分类效果.

表2和表3展示了本文方法和对比方法在VCG这2个数据集上的比较结果. 传统的仅利用图像的CNN和PCNN的方法在VCG的2个数据集上效果普遍偏低,而利用VGG-transfer的思想处理图像情感分析,效果得到了很大的提升. 本文同时结合权重迁移和异构特征迁移融合的方法DCCA和M1相比VGG-transfer已经得到了提升,其中M1相比DCCA展示了更好的性能. 此外,利用形容词和名词弱监督的M2方法在性能上得到了进一步的提升. 由于提出的方法M1及其变体M2在VCG的2个数据集上相比其他对比方法均展示出更好的性能,说明提出的方法在相同领域不同背景的数据集下具有领域适应能力.

表4和表5分别展示了本文方法和对比方法在E-VSK数据集和Flickr-2数据集上的实验结果. 针对E-VSK数据集的实验评估采取与VCG数据集同样的对比方式,且方法M1及其变体M2都展示了优异的性能,尤其是M2效果更好. 由于Flickr-2数据集是公开数据集,其标签来自于人工标注,故没有图像的ANP信息,则在该数据集上仅评估提出的M1方法的性能.

表2 不同方法在VCGⅠ数据集上的情感分类效果

方法	准确率	召回率	F1值
CNN	0.552 3	0.559 2	0.565 7
PCNN	0.557 8	0.561 2	0.569 8
VGG-transfer	0.658 2	0.672 5	0.675 8
DCCA + early-self-attention	0.718 8	0.718 3	0.728 4
DCCA + self-attention	0.723 8	0.726 1	0.734 8
M1 + early-self-attention	0.723 9	0.730 2	0.740 6
M1 + self-attention	0.729 4	0.736 8	0.747 9
M2 + early-self-attention	0.725 8	0.729 3	0.742 3
M2 + self-attention	0.733 4	0.741 3	0.752 2

表3 不同方法在VCGⅡ数据集上的情感分类效果

方法	准确率	召回率	F1值
CNN	0.545 1	0.539 9	0.547 5
PCNN	0.562 1	0.575 3	0.574 8
VGG-transfer	0.757 1	0.773 4	0.771 3
DCCA + early-self-attention	0.776 6	0.799 6	0.793 2
DCCA + self-attention	0.811 7	0.836 7	0.828 1
M1 + early-self-attention	0.782 4	0.790 8	0.792 5
M1 + self-attention	0.821 3	0.881 2	0.837 6
M2 + early-self-attention	0.813 2	0.824 1	0.822 6
M2 + self-attention	0.846 2	0.874 2	0.858 5

表4 不同方法在E-VSK数据集上的情感分类效果

方法	准确率	召回率	F1值
CNN	0.604 7	0.615 2	0.614 9
PCNN	0.622 7	0.632 2	0.629 9
VGG-transfer	0.660 6	0.697 4	0.688 7
DCCA + early-self-attention	0.684 3	0.770 5	0.741 2
DCCA + self-attention	0.719 2	0.791 2	0.765 8
M1 + early-self-attention	0.710 4	0.772 4	0.758 4
M1 + self-attention	0.756 4	0.810 6	0.795 7
M2 + early-self-attention	0.725 1	0.780 2	0.769 2
M2 + self-attention	0.773 1	0.816 8	0.805 9

表5 不同方法在Flickr-2数据集上的情感分类效果

方法	准确率	召回率	F1值
CNN	0.668 1	0.691 2	0.689 7
PCNN	0.689 3	0.723 4	0.712 4
VGG-transfer	0.793 6	0.793 8	0.801 9
DCCA + early-self-attention	0.804 9	0.805 2	0.816 2
DCCA + self-attention	0.832 1	0.862 3	0.852 5
M1 + early-self-attention	0.813 1	0.825 2	0.830 3
M1 + self-attention	0.852 6	0.870 6	0.867 4

为了证明本文方法同样适用于情感三分类,表6给出了在Flickr-3数据集上的结果,同样显示了

表6 不同方法在Flickr-3数据集上的情感分类准确率

方法	准确率
CNN	0.521 3
PCNN	0.536 2
VGG-transfer	0.591 7
DCCA + early-self-attention	0.564 4
DCCA + self-attention	0.630 9
M1 + early-self-attention	0.579 6
M1 + self-attention	0.639 5

本文方法效果更好.

4　结束语

提出了一种基于两阶段深度网络结构的视觉情感分析方法. 该方法首先依赖提出的多模态深度单重判别性相关分析模型来映射图像和与之共现文本的深度特征到潜在空间中,在该潜在空间中迁移文本的语义特征到图像的判别性视觉特征中. 然后,进一步引入注意力网络来学习潜在空间中生成的语义增强的判别性视觉特征从而用于情感分类. 已经在 5 个真实数据集上评估了模型的有效性,且实验结果表明提出的方法优于其它仅利用视觉模态的方法和迁移学习的方法. 在未来的工作中将考虑设计更合理的注意力网络以及研究更好的特征迁移融合策略以进一步提高异构多模态特征融合的效果.

参考文献:

[1] Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1): 9.

[2] Borth D, Ji R, Chen T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs [C] // ACM International Conference on Multimedia. New York: ACM, 2013: 223-232.

[3] Jou B, Chen T, Pappas N, et al. Visual affect around the world: A large-scale multilingual visual sentiment ontology[C] // ACM International Conference on Multimedia. New York: ACM, 2015: 159-168.

[4] 李钊, 卢苇, 邢薇薇, 等. CNN 视觉特征的图像检索[J]. 北京邮电大学学报, 2015, 38(s1): 103-106.
Li Zhao, Lu Wei, Xing Weiwei, et al. Image retrieval based on CNN visual features[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(s1): 103-106.

[5] You Q, Yang J, Yang J, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C] //29th AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2015: 381-388.

[6] 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.

[7] Islam J, Zhang Y. Visual sentiment analysis for social images using transfer learning approach [C] // IEEE International Conferences on Big Data and Cloud Computing. Piscataway: IEEE, 2016: 124-130.

[8] Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis [C] // International Conference on Machine Learning. Atlanta: ICML, 2013: 1247-1255.

[9] Dorfer M, Kelz R, Widmer G, et al. Deep linear discriminant analysis [C] // International Conference on Learning Representations. San Juan: ICLR, 2016: 1-13.

[10] Katsurai M, Satoh S. Image sentiment analysis using latent correlations among visual, textual, and sentiment views[C] // IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2016: 2837-2841.