计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180341
56(6): 1312-1324, 2019

基于层次化深度关联融合网络的社交媒体情感分类

蔡国永 吕光瑞 徐 智
(广西可信软件重点实验室(桂林电子科技大学) 广西桂林 541004)
(ccgycai@guet.edu.cn)

A Hierarchical Deep Correlative Fusion Network for Sentiment Classification in Social Media

Cai Guoyong, Lü Guangrui, and Xu Zhi
(*Guangxi Key Laboratory of Trusted Software* (*Guilin University of Electronic Technology*), *Guilin*, *Guangxi* 541004)

Abstract Most existing research of sentiment analysis are based on either textual or visual data and can not achieve satisfied results. As multi-modal data can provide richer information, multi-modal sentiment analysis is attracting more and more attentions and has become a hot research topic. Due to the strong semantic correlation between visual data and the co-occurrence textual data in social media, mixed data of texts and images provides a new view to learn better classifier for social media sentiment classification. A hierarchical deep correlative fusion network framework is proposed to jointly learn textual and visual sentiment representations from training samples for sentiment classification. In order to alleviate the problem of fine-grained semantic matching between image and text, both the middle level semantic features of images and the deep multi-modal discriminative correlation analysis are applied to learn the most relevant visual feature representation and semantic feature representation, meanwhile, keeping both the visual and semantic feature representations to be linear discriminable. Motivated by the successful use of attention mechanisms, we further propose a multi-modal attention fusion network by incorporating visual and semantic feature representations to train sentiment classifier. Experiments on the real-world datasets which come from social networks show that, the proposed method gets more accurate prediction on multi-media sentiment analysis by capturing the internal relations between text and image hierarchically.

Key words social media; sentiment analysis; deep correlation; discriminant correlation analysis; multi-modal attention fusion

摘 要 现有的多数情感分析研究都是基于单一文本或视觉数据,效果还不够理想,多模态数据由于能够提供更丰富的信息,因此多模态情感分析正受到越来越多的关注.社交媒体上视觉数据常常和与之共现的文本数据存在较强的语义关联,因此混合图文的多模态情感分类为社交媒体情感分析提供了新的视角.为了解决图文之间的精细语义配准问题,提出了一种基于层次化深度关联融合网络的多媒体数据情感分类模型.该模型不仅利用图像的中层语义特征,还利用多模态深度多重判别性相关分析来学习最大

收稿日期:2018-05-11;**修回日期**:2018-12-13
基金项目:国家自然科学基金项目(61763007,66162014);广西自然科学基金重点项目(2017JJD160017);广西可信软件重点实验室项目(201503)
This work was supported by the National Natural Science Foundation of China (61763007, 66162014), the Natural Science Foundation of Guangxi Province of China (2017JJD160017), and the Project of the Guangxi Key Laboratory of Trusted Software (201503).
通信作者:吕光瑞(lvguangrui_rio@163.com)

相关的图像视觉特征表示和文本语义特征表示,而且使形成的视觉特征表示和语义特征表示均具有线性判别性.在此基础上,提出合并图像视觉特征表示和文本语义特征表示的多模态注意力融合网络,以进一步改进情感分类器.最后,在来自于社交网络的真实数据集上的大量实验结果表明,通过层次化捕获视觉情感特征和文本情感特征之间的内部关联,可以更准确地实现对图文融合社交媒体的情感分类预测.

关键词 社交媒体;情感分析;深度关联;判别性相关分析;多模态注意力融合

中图法分类号 TP391

随着移动互联网和智能终端的快速发展,社交媒体上的用户生成内容变得越来越多样化,社交媒体数据已不再仅限于单一的文本形式,例如越来越多的社交用户倾向于使用图像和短文本这种多模态内容的形式来表达他们的观点和在社交媒体上相互交流.这些大量社交用户分享的多模态数据为人们提供了探索众多话题的情感和观点的宝库,因此多模态情感分析已经成为一个重要的研究热点[1-9],但是大规模多模态社交媒体数据的情感分析还是一个充满挑战的任务.

早期的情感研究较多关注单一的文本或图像,且采用传统的机器学习分类算法.近年来,鉴于深度学习技术的优异表现,越来越多的研究人员倾向于使用深度神经网络来学习文本的分布式和稳健的特征表示用于情感分类[10-13].与此同时,卷积神经网络(convolutional neural network, CNN)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,一些研究者开始探索基于 CNN 的图像情感分析[14-16].最近,在多模态情感分析的研究中[1-9],利用深度神经网络的方法在性能上也更优异.多模态情感分析是融合多种模态的信息进行统一的分类预测任务,其关键的问题是多模态样本特征的融合.由于不同模态的异质性,模态之间特征的融合是较困难的.尽管基于深度网络相关的模型已经取得了不错的进展,但是基于深度网络的融合模型仍需要进一步深入研究.

为了克服已有的图像-文本的多媒体情感分析研究中存在的异构模态的特征融合方式相对简单以及单一图像处理上仅从图像自身提取特征等不足,本文的主要贡献有 4 个方面:

1) 在图像的处理上利用迁移学习策略和图像中层语义特征相结合的方法来构建具有一定语义的视觉情感特征表示.

2) 结合深度典型相关分析(deep canonical correlation analysis, DCCA)[17]和深度线性判别分析(deep linear discriminant analysis, DeepLDA)[18]的思想提出多模态深度多重判别性相关分析的联合优化目标,通过优化生成最大相关的判别性视觉特征和判别性语义特征以构建图像和文本在特征层次上的语义相关,且使特征具有判别性的能力,从而提升语义配准.

3) 提出基于多模态协同注意力网络的融合方法,能进一步序列化地交互图像的视觉特征和文本的语义特征,从而更好地匹配融合多模态特征.

4) 在多个数据集上的对比实验表明,本文提出的层次化深度关联融合的网络模型在情感分类任务中能取得更好的分类效果.

1 相关工作

多模态情感分析的研究尚且处于初期阶段,大致可以分为 2 类.较早的研究以特征选择模型为主,最近开始基于深度神经网络模型展开研究.

Wang 等人[1]利用统一的跨媒体词袋模型来表示文本特征和图像特征,且利用机器学习的方法来预测融合后的情感,结果表明跨模态情感分类结果要略优于单模态的情感分类结果.Cao 等人[2]融合来自于形容词名词对(adjective noun pairs, ANPs)[19]的图像中层视觉特征的预测结果和由情感词、情感标签和句子结构规则组成的文本特征的预测结果,其中图像和文本的融合权重是通过参数来控制,最后用于微博的公共情感分析.Poria 等人[3]通过使用特征级的和决策级的融合方法合并来自于多模态的情感信息.Katsurai 等人[4]首先构筑视觉特征、文本特征和情感特征,然后利用映射矩阵映射视觉、文本、情感这 3 个模态的数据到一个共同的潜在嵌入空间中,认为潜在空间中的映射特征是来自于不同模态的互补信息从而被用于训练情感分类器.

最近深度学习方法应用于多模态情感预测也备受关注.如 Cai 等人[5]利用 2 个单独的 CNN 结构

分别学习文本特征表示和图像特征表示,将其合并后输入另外的 CNN 结构以进行多媒体的情感分析. Yu 等人[6]也利用 2 个 CNN 结构分别提取文本和图像的特征表示,使用逻辑回归对文本的和图像的特征表示进行情感预测,最后使用平均策略和加权的方法融合概率结果.Baecchi 等人[7]提出基于连续词袋模型和降噪自动编码的多模态特征的学习模型以进行 Twitter 数据情感分析,当然该模型也可应用到其他的社交媒体数据上.You 等人[8]提出跨模态一致回归的方法用于结合视觉和文本的情感分析,该方法利用深度视觉的和文本的特征构建回归模型.而 Xu 等人[9]利用卷积网络的结构来提取图像和文本的特征表示,然后利用残差的模型来合并图像和文本的多模态特征用于情感分析.

尽管这些模型都是有效的,但是大多都独立地使用视觉和文本的信息,且在融合过程中往往忽略了图像和文本之间的内在关联.通常,组合不同模态数据的多模态融合方法可以分为早融合、后融合、混合融合[20].其中,后融合涉及为每种模态数据构建相应的分类器,然后结合这些决策进行预测;而早融合需要将不同模态的特征融合到单个分类器中.本文的研究仍属于特征层的融合,但是不同于已有的研究方法,本文工作的关注点有 2 个方面:1)同时处理图像和与之共现的文本信息;2)在多模态深度网络的结构中,利用层次化深度关联融合的方法来探究图像和文本之间的语义关联.首先,本文整合 DCCA[17]和 DeepLDA[18]到一个统一的联合多模态优化目标中,以此构建图像和与之共现的文本在特征层次上的语义关联,且使各自生成的特征具有较好的判别性.此外,最近注意力模块已经成为应用于各种任务的现代神经系统的组成部分,比如机器翻译[21]、图像问答任务[22]和图像标题生成[23]等,然而很少的研究工作已经利用注意力机制进行融合,本文提出基于协同注意力(co-attention)机制的多模态融合策略,用于训练情感分类器.

2 方法描述

本节介绍提出的用于多模态情感分析任务的层次化深度关联融合的网络模型,整体结构如图 1 所示,总共由 5 个部分构成:①视觉模态特征提取网络;②文本模态特征提取网络;③多模态深度多重判别性相关分析;④co-attention 网络的多模态注意力融合模型;⑤分类网络.

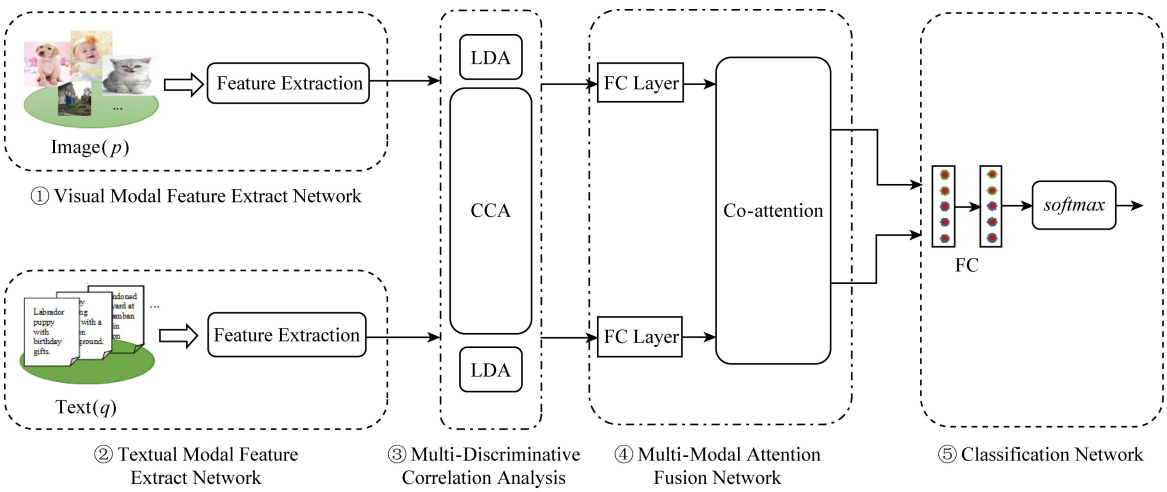

Fig. 1 Framework of hierarchical deep correlative fusion network for multi-modal sentiment classification

图 1 基于层次化深度关联融合网络的社交媒体多模态情感分类框架图

基于层次化深度关联融合网络的多模态情感分类模型首先利用图 1 中①②的多模态特征提取网络逐层提取视觉模态和文本模态的特征,得到相对应的顶层特征表示,然后通过图 1 中③进一步生成最大相关的判别性特征表示,最后使用图 1 中④的 co-attention 网络来交互合并这 2 种特征表示并传递到图 1 中⑤的全连接神经网络(fully connected neural network, FCNN)中进一步深层融合后再用于训练情感分类器.下面阐述模型的细节.

2.1 视觉模态特征提取网络

尽管已有学者在情感分析相关研究上探测过图像视觉特征[14-16,24]或者图像中层语义特征[19,25-26],但是仅从单一视觉特征或中层语义特征的角度来构筑视觉情感特征,并不能构筑完整的且易于理解的

图像视觉特征.本文同时从图像特征提取和图像中层语义特征提取的角度来学习高层次的视觉情感表示,如图 2 中①所示.

图像的特征提取是基于 VGG[27] 展开的,其由 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.本文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.首先,VGG16 模型在 ImageNet 的数据集上训练好,然后迁移已经学习好的参数到情感分析的目标中.在提出的模型中,修改最后用于目标分类的全连接层为特征映射层,然后提取该全连接层的特征输出,如图 2 中①(a-1)所示.

为了提取更全面的图像中层语义特征,首先划分每一个图像对应的中层语义特征(ANP)为形容词和名词,然后通过 CNN 来分别提取图像的形容词描述性特征和名词客观性特征.针对形容词和名词的特征提取网络,CNN 采用的是二维卷积,每一个形容词或名词的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 2 中①(a-2)中 A-net 和 N-net,其分别由同样的卷积层和全连接层组成.

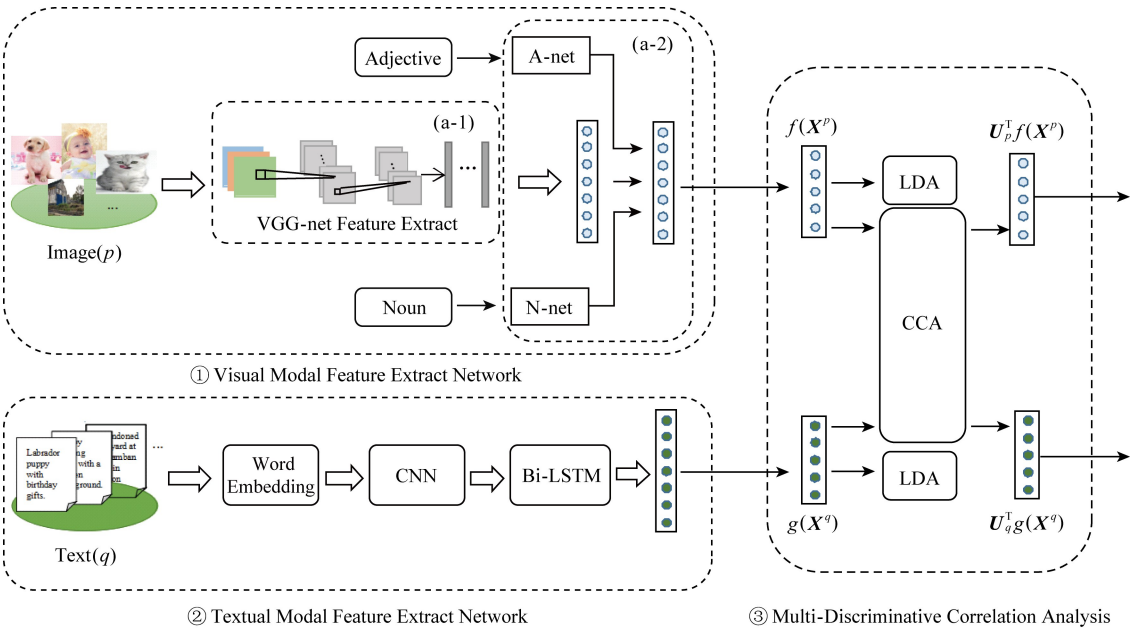

Fig. 2 Schematic sketch of deep multi-modal multi-discriminative correlation analysis to learn the visual and textual content

图 2 视觉和文本的多模态深度多重判别性相关分析图解

总之,在视觉模态特征提取上,本文提出联合学习图像 ANP 的形容词和名词以及图像特征以构筑具有一定语义的视觉情感特征表示,以此缓解图像视觉特征和文本语义特征之间的语义鸿沟.后文中将称视觉模态特征提取网络为 f.

2.2 文本模态特征提取网络

文本模态特征提取网络是由词向量输入层、卷积层、双向长短时记忆网络(Bi-LSTM)层和全连接层组成,如图 2 中②所示.

假设 $\boldsymbol{x}_i \in \mathbb{R}^k$ 是句子中第 i 个词对应的 k 维词向量,则一个长度为 n 的句子表示为

$$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \cdots \oplus \boldsymbol{x}_n, \tag{1}$$

其中,$\oplus$ 表示连接操作,在句子矩阵 $\boldsymbol{x}_{1:n}$ 上利用一个单层的 CNN[28],它的卷积层包含高度分别为 h_1,h_2,h_3 的 3 个滤波器 $\boldsymbol{F}_1 \in \mathbb{R}^{h_1 \times k}$,$\boldsymbol{F}_2 \in \mathbb{R}^{h_2 \times k}$,$\boldsymbol{F}_3 \in \mathbb{R}^{h_3 \times k}$.每个滤波器 $\boldsymbol{F}_i$ 在输入的句子序列上进行滑动,当 $\boldsymbol{F}_i$ 应用到整个句子矩阵中每一个可能的 h_i 窗口的词上时,就会产生一个特征映射 $\boldsymbol{c}_i \in \mathbb{R}^{n-h_i+1}$,其中某一项窗口的词的特征映射 $\boldsymbol{c}_{i,j}$ 为

$$\boldsymbol{c}_{i,j} = \delta(\boldsymbol{F}_i * \boldsymbol{x}_{[j:j+h_i-1]} + b_i), \tag{2}$$

这里 $*$ 是卷积操作,$j = 1, 2, \cdots, n-h_i+1$,$b_i \in \mathbb{R}$ 是一个偏置项,$\delta(\cdot)$ 是一个非线性激活函数.每一个滤波器 $\boldsymbol{F}_i$ 能够生成 M 个这样的特征映射,因此总共获得了 $3M$ 个特征映射.然后,在滤波器 $\boldsymbol{F}_i$ 的 M 个特征映射向量的每一个长度上应用最大池化操作,则结果输出向量为 $\boldsymbol{o}_i \in \mathbb{R}^M$,具体表示为

$$\boldsymbol{o}_i=(\max(\boldsymbol{c}_i^1),\max(\boldsymbol{c}_i^2),\cdots,\max(\boldsymbol{c}_i^M)). \quad (3)$$

通过⊕连接每一个 $\boldsymbol{o}_i$ 得到 $\boldsymbol{o}=(\boldsymbol{o}_1\oplus\boldsymbol{o}_2\oplus\boldsymbol{o}_3)\in\mathbb{R}^{3M}$.然后将 $\boldsymbol{o}$ 输入 Bi-LSTM 网络,从正向和反向的角度来使用已提取的特征从而更好地学习输入的文本序列.最后,经过对文本的序列建模后,将 Bi-LSTM 的输出传递给全连接的神经网络以更好地融合时序特征以形成更容易被区分的高层特征表示.后文中称文本模态特征提取网络为 g.

2.3 多模态深度多重判别性相关分析

本文提出的多模态深度多重判别性相关分析是基于典型相关分析(canonical correlation analysis, CCA)和线性判别分析(linear discriminant analysis, LDA)展开的.两者都来自经典的多元统计,都依赖于各自输入特征分布的协方差结构.不同之处在于,CCA 是一种适用于多模态数据的分析方法,但是它既没有考虑标签信息也不能对各自模态的内部信息进行分析;而 LDA 是一种利用标签信息的且适用于单模态数据的分析方法,但是它不能直接地应用到多模态数据分析上,因此可将两者结合起来,以充分发掘各自模型的优势,从而形成一个在多模态学习过程中既探究不同模态之间的最大相关性又兼顾各自模态最大判别性的多模态数据处理方法.

多模态的样本数据往往来自于异构特征空间,不同模态数据的特征分布差异较大,此时如果将异构特征融合后再进行 LDA,较难取得好的效果.例如,那些来自于社交网站的图像和文本,如果直接将图像和文本的特征融合后再用于 LDA,这既没有考虑图像和文本的对应关联也没有考虑图像和文本各自特征分布的差异.因此本文将在考虑不同模态之间相关性的同时,也尽量考虑不同模态之间的特征分布的差异,即在寻求视觉模态和文本模态最大相关性的同时,兼顾视觉模态和文本模态各自的线性判别性.多模态深度多重判别性相关分析方法包含 2 部分:相关性分析部分和判别性分析部分.

在描述方法之前,首先给出相关变量的数学定义形式.令 $\{(\boldsymbol{x}_i^p,\boldsymbol{x}_i^q)\}_{i=1}^N$ 表示一系列 N 对的图像-文本观察值,其中 $\{\boldsymbol{x}_i^p\}_{i=1}^N=\boldsymbol{X}^p\in\mathbb{R}^{N\times d_p}$,$\{\boldsymbol{x}_i^q\}_{i=1}^N=\boldsymbol{X}^q\in\mathbb{R}^{N\times d_q}$,且属于 C 个不同的类 $c\in\{k_1,k_2,\cdots,k_C\}$,这里上标 p 和 q 分别表示图像和文本,即视觉特征向量 $\boldsymbol{x}_i^p$ 表示第 i 个对中的图像,同理 $\boldsymbol{x}_i^q$ 表示对应于图像 $\boldsymbol{x}_i^p$ 的文本特征向量.此外,当在视觉模态上同时利用图像特征和图像的中层语义特征共同学习时,视觉模态上是 3 个输入,这里不再形式化表示,可依照图像-文本对的形式化表示类推.利用 2.1 节设计的视觉模态特征提取网络 f 和 2.2 节设计的文本模态特征提取网络 g 作为非线性特征映射来处理输入的数据 $\{(\boldsymbol{x}_i^p,\boldsymbol{x}_i^q)\}_{i=1}^N$.图像和文本分别通过 f 和 g 这 2 个不同的神经网络生成各自的顶层特征表示 $f(\boldsymbol{X}^p)\in\mathbb{R}^{N\times L}$ 和 $g(\boldsymbol{X}^q)\in\mathbb{R}^{N\times L}$.设 f 和 g 这 2 个不同神经网络的学习参数 $(\boldsymbol{W}^p;\boldsymbol{b}^p)$ 和 $(\boldsymbol{W}^q;\boldsymbol{b}^q)$ 的集合分别表示为 $\boldsymbol{\theta}^p$ 和 $\boldsymbol{\theta}^q$,且固定 $f(\boldsymbol{X}^p)$ 和 $g(\boldsymbol{X}^q)$ 的维度是相同的,记为 L.则多模态深度多重判别性相关分析的模型框架简单描述为

$$J(f(\boldsymbol{X}^p),g(\boldsymbol{X}^q))=C(f(\boldsymbol{X}^p),g(\boldsymbol{X}^q))+[D(f(\boldsymbol{X}^p))+D(g(\boldsymbol{X}^q))], \quad (4)$$

其中,$J(f(\boldsymbol{X}^p),g(\boldsymbol{X}^q))$ 表示 p,q 模态间的多重判别性相关分析的目标函数,$C(f(\boldsymbol{X}^p),g(\boldsymbol{X}^q))$ 表示两者模态间的相关性分析项,$D(f(\boldsymbol{X}^p))$ 和 $D(g(\boldsymbol{X}^q))$ 分别表示各自模态内部的判别性分析项.

本文以式(4)为基准来设计模型,即从不同模态之间来考虑多重判别性的相关分析,下面分别对模型中的各项进行阐述.

2.3.1 多模态深度相关性分析

Andrew 等人[17]提出基于 CCA 的端到端的深度神经网络的解释方法 DCCA,其优化目标是推动多模态网络学习高度关联的特征表示.受到 DCCA 方法的启发,本文在自定义的多模态深度网络结构 f 和 g 下来学习视觉模态和文本模态间的相关性,称为 Multi-DCCA.

在 CCA 中,首先通过预处理的操作,分别使 $f(\boldsymbol{X}^p)$ 和 $g(\boldsymbol{X}^q)$ 变成中心数据矩阵,表示为

$$\bar{f}(\boldsymbol{X}^p)=f(\boldsymbol{X}^p)-\frac{1}{N}f(\boldsymbol{X}^p)\boldsymbol{1}, \quad (5)$$

$$\bar{g}(\boldsymbol{X}^q)=g(\boldsymbol{X}^q)-\frac{1}{N}g(\boldsymbol{X}^q)\boldsymbol{1}, \quad (6)$$

其中,N 表示数据的总数,$\boldsymbol{1}\in\mathbb{R}^{N\times N}$ 表示元素全为 1 的矩阵.

视觉模态和文本模态的顶层特征表示的正则化自协方差矩阵,分别表示为

$$\boldsymbol{\Sigma}_{pp}=\frac{1}{N-1}\bar{f}(\boldsymbol{X}^p)\bar{f}(\boldsymbol{X}^p)^{\mathrm{T}}+r_p\boldsymbol{I}, \quad (7)$$

$$\boldsymbol{\Sigma}_{qq}=\frac{1}{N-1}\bar{g}(\boldsymbol{X}^q)\bar{g}(\boldsymbol{X}^q)^{\mathrm{T}}+r_q\boldsymbol{I}, \quad (8)$$

其中,r_p,r_q 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.

除了领域自身的方差外,不同领域学习到的特征表示的交叉协方差矩阵为

$$\boldsymbol{\Sigma}_{pq}=\frac{1}{N-1}\bar{f}(\boldsymbol{X}^p)\bar{g}(\boldsymbol{X}^q)^{\mathrm{T}}. \quad (9)$$

基于 CCA 中介绍的协方差矩阵 $\boldsymbol{\Sigma}_{pp}$,$\boldsymbol{\Sigma}_{pq}$,$\boldsymbol{\Sigma}_{qq}$,定义矩阵 $\boldsymbol{T}=\boldsymbol{\Sigma}_{pp}^{-1/2}\boldsymbol{\Sigma}_{pq}\boldsymbol{\Sigma}_{qq}^{-1/2}$.然后 $f(\boldsymbol{X}^p)$和 $g(\boldsymbol{X}^q)$的总体关联通过相对应的奇异值问题 $\boldsymbol{T}=\boldsymbol{U}_p\boldsymbol{\Lambda}\boldsymbol{U}_q$ 和 $\boldsymbol{\Lambda}=\mathrm{diag}(d)$中的奇异值 d 的求和来计算.$\boldsymbol{U}_p$ 和 $\boldsymbol{U}_q$ 分别是转化视觉模态和文本模态到线性 CCA 子空间的映射矩阵.故 Multi-DCCA 的相关分析是在 f 和 g 相对应的网络参数 $\boldsymbol{\theta}^p$ 和 $\boldsymbol{\theta}^q$ 下最大化奇异值 d 的和,即:

$$C(f(\boldsymbol{X}^p),g(\boldsymbol{X}^q))=\arg\max_{\boldsymbol{\theta}^p,\boldsymbol{\theta}^q}\sum_{i=1}^{L}d_i. \quad (10)$$

在网络 f 和 g 有相同的特征维度 L 时,也可以通过最大化矩阵迹的范数 $\|\boldsymbol{T}\|_{\mathrm{tr}}=\mathrm{tr}((\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T})^{1/2})$来优化典型关联.

2.3.2 多模态深度判别性分析

Dorfer 等人[18]提出基于 LDA 的端到端的深度神经网络的解释方法 DeepLDA,其优化目标是推动网络在顶层表示上学习线性可分的潜在空间.受到 DeepLDA 的启发,本文在视觉模态特征提取网络 f 的顶层和文本模态特征提取网络 g 的顶层同时学习可以最大化 C 个不同的多模态数据类别之间区分的潜在表示,称为 Multi-DeepLDA.

对于 LDA 而言,$\boldsymbol{\Sigma}_{pp}$ 可作为视觉模态的总体离散度矩阵,同理 $\boldsymbol{\Sigma}_{qq}$ 可作为文本模态的总体离散度矩阵.此外,由于图像-文本对的标签属于 C 个不同的类 $c\in\{k_1,k_2,\cdots,k_C\}$,则 LDA 还需要 C 个不同类别中每个类别的视觉模态和文本模态的协方差矩阵 $\boldsymbol{\Sigma}_{pc}$,$\boldsymbol{\Sigma}_{qc}$,以及视觉模态和文本模态中所有不同类协方差矩阵的均值 $\boldsymbol{\Sigma}_{p\mathrm{w}}$,$\boldsymbol{\Sigma}_{q\mathrm{w}}$,即类内离散度矩阵,分别表示为

$$\boldsymbol{\Sigma}_{pc}=\frac{1}{N_c-1}\bar{f}(\boldsymbol{X}_c^p)\bar{f}\ (\boldsymbol{X}_c^p)^{\mathrm{T}}+r\boldsymbol{I}, \quad (11)$$

$$\boldsymbol{\Sigma}_{qc}=\frac{1}{N_c-1}\bar{g}(\boldsymbol{X}_c^q)\bar{g}\ (\boldsymbol{X}_c^q)^{\mathrm{T}}+r\boldsymbol{I}, \quad (12)$$

$$\boldsymbol{\Sigma}_{p\mathrm{w}}=\frac{1}{C}\sum_c\boldsymbol{\Sigma}_{pc}, \quad (13)$$

$$\boldsymbol{\Sigma}_{q\mathrm{w}}=\frac{1}{C}\sum_c\boldsymbol{\Sigma}_{qc}, \quad (14)$$

其中,r 是正则化参数,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 $\boldsymbol{\Sigma}_{pp}$,$\boldsymbol{\Sigma}_{qq}$ 和类内离散度矩阵 $\boldsymbol{\Sigma}_{p\mathrm{w}}$,$\boldsymbol{\Sigma}_{q\mathrm{w}}$来定义视觉模态和文本模态的各自类间离散度矩阵 $\boldsymbol{\Sigma}_{p\mathrm{b}}$,$\boldsymbol{\Sigma}_{q\mathrm{b}}$:

$$\boldsymbol{\Sigma}_{p\mathrm{b}}=\boldsymbol{\Sigma}_{pp}-\boldsymbol{\Sigma}_{p\mathrm{w}};\ \boldsymbol{\Sigma}_{q\mathrm{b}}=\boldsymbol{\Sigma}_{qq}-\boldsymbol{\Sigma}_{q\mathrm{w}}, \quad (15)$$

则 Multi-DeepLDA 是通过找到视觉模态和文本模态内部的映射矩阵 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$,使得在相同标签下各自模态内的类间离散度矩阵和类内离散度矩阵的比值最大化,具体表述为

$$D(f(\boldsymbol{X}^p))=\arg\max_{\boldsymbol{A}_1}\frac{|\boldsymbol{A}_1\boldsymbol{\Sigma}_{p\mathrm{b}}\boldsymbol{A}_1^{\mathrm{T}}|}{|\boldsymbol{A}_1\boldsymbol{\Sigma}_{p\mathrm{w}}\boldsymbol{A}_1^{\mathrm{T}}|}, \quad (16)$$

$$D(g(\boldsymbol{X}^q))=\arg\max_{\boldsymbol{A}_2}\frac{|\boldsymbol{A}_2\boldsymbol{\Sigma}_{q\mathrm{b}}\boldsymbol{A}_2^{\mathrm{T}}|}{|\boldsymbol{A}_2\boldsymbol{\Sigma}_{q\mathrm{w}}\boldsymbol{A}_2^{\mathrm{T}}|}, \quad (17)$$

其中,映射矩阵 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$ 分别转化各自模态的数据到一个 $C-1$ 维的空间中,在各自空间中的映射特征变得线性可区分.

下面仅对 $\boldsymbol{A}_1$ 的求解做简要阐述,$\boldsymbol{A}_2$ 的求解以此类推.具体来讲,最大化类别间隔的线性组合 $\boldsymbol{A}_1$ 是通过求解 $\boldsymbol{\Sigma}_{p\mathrm{b}}\boldsymbol{e}_i=v_i(\boldsymbol{\Sigma}_{p\mathrm{w}}+\lambda\boldsymbol{I})\boldsymbol{e}_i$ 确定的,而映射矩阵 $\boldsymbol{A}_1$ 是与特征值问题有关的特征向量 $\boldsymbol{e}$ 的集合.最后得到的特征值 v_i 量化了在对应特征向量方向 $\boldsymbol{e}_i$ 上的判别性方差,最终的优化目标关注的是最大化 k_p 个最小的特征值$\{v_1^p,v_2^p,\cdots,v_k^p\}$,即:

$$\arg\max_{\boldsymbol{\theta}^p}\frac{1}{k_p}\sum_{i=1}^{k_p}v_i^p, \quad (18)$$

其中,$\{v_1^p,v_2^p,\cdots,v_k^p\}=\{v_j^p\mid v_j^p<\min\{v_1^p,v_2^p,\cdots,v_{C-1}^p\}+\varepsilon\}$,该目标函数的设计是为了推动网络能学习判别性的能力到特征空间中所有可用的维度中[18].

2.3.3 相关分析与判别分析的融合

综合 2.3.1 节和 2.3.2 节可看出,Multi-DCCA 和 Multi-DeepLDA 都是基于相对应的特征值问题的特征结构优化的.其中,Multi-DCCA 的优化是把最大化视觉模态特征提取网络 f 和文本模态特征提取网络 g 的隐层输出的相关性作为目标来求解矩阵 $\boldsymbol{T}$ 的奇异值;而 Multi-DeepLDA 的优化是在相同的多模态类别下最大化视觉的和文本的各自模态内类别的区分,其由相对应的广义特征值问题的特征值大小进行量化.尽管两者的优化有差异,但是这 2 种方法有相同之处,即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.

故多模态深度多重判别性相关分析是同时使用 Multi-DCCA 和 Multi-DeepLDA 的模型和优化理论,即同时优化 2 个不同模态之间隐层表示的相关性以及使各自模态学到表示具有判别性能力的联合优化目标的形式化表示为

$$\arg\max_{\boldsymbol{\theta}^p,\boldsymbol{\theta}^q}\frac{1}{L}\sum_{i=1}^{L}d_i+\frac{1}{k_p}\sum_{i=1}^{k_p}v_i^p+\frac{1}{k_q}\sum_{i=1}^{k_q}v_i^q, \quad (19)$$

其中,第 1 项是为了优化视觉模态和文本模态之间的相关性,其中用 L 来泛化典型相关;而第 2 项和第 3 项分别是为了优化视觉模态和文本模态的判别性.

多模态深度多重判别性的优化目标式(19)是个端到端的优化过程,首先需要计算相关性的优化目标分别对 $f(\boldsymbol{X}^p)$ 和 $g(\boldsymbol{X}^q)$ 的梯度,以及各自判别性的优化目标对 $f(\boldsymbol{X}^p)$ 和 $g(\boldsymbol{X}^q)$ 的梯度,然后沿着多模态网络的 2 个分支并通过标准的反向传播的方法计算针对 $\boldsymbol{\theta}^p$ 和 $\boldsymbol{\theta}^q$ 的梯度.

经过式(19)这种多模态深度多重判别性相关分析的优化,最后经 CCA 的特征映射可将图 2 中①②这 2 个网络的顶层输出 $f(\boldsymbol{X}^p)$ 和 $g(\boldsymbol{X}^q)$ 转化成最大相关的且各自具有判别性的特征表示 $\boldsymbol{U}_p^{\mathrm{T}}f(\boldsymbol{X}^p)$ 和 $\boldsymbol{U}_q^{\mathrm{T}}g(\boldsymbol{X}^q)$,如图 2 中③所示.如果使用图 2 中②提取的文本语义特征和仅用图 2 中①的 VGG-net 提取图像视觉特征,通过图 2 中③生成新的特征表示,后文称该模型为 DDC;若使用图 2 中②提取的文本语义特征和图 2 中①的 3 个子网络共同提取含有一定语义的图像视觉特征通过图 2 中③生成新的特征表示,则后文称该模型为 DANDC.式(19)的表述是从优化求解的角度展开的,如果从理论层次上来形式化表述整个过程,则表示为

$$\begin{aligned}&\min_{\boldsymbol{\theta}^p,\boldsymbol{\theta}^q,\boldsymbol{U}^p,\boldsymbol{U}^q}\sum_{i\in\mathbf{N}_+}\left\|\boldsymbol{U}_p^{\mathrm{T}}f(\boldsymbol{x}_i^p)-\boldsymbol{U}_q^{\mathrm{T}}g(\boldsymbol{x}_i^q)\right\|_{\mathrm{F}}^2+\\&\sum_{i\in\mathbf{N}_+,y_i\in c}L(y_i,f(\boldsymbol{x}_i^p),g(\boldsymbol{x}_i^q)),\\&\text{s.t. }\boldsymbol{U}_p^{\mathrm{T}}\boldsymbol{\Sigma}_{pp}\boldsymbol{U}_p=\boldsymbol{I},\boldsymbol{U}_q^{\mathrm{T}}\boldsymbol{\Sigma}_{qq}\boldsymbol{U}_q=\boldsymbol{I},\\&\boldsymbol{u}_{p_i}^{\mathrm{T}}\boldsymbol{\Sigma}_{pq}\boldsymbol{u}_{q_j}^{\mathrm{T}}=0,\forall i\neq j,\end{aligned}\tag{20}$$

其中,式(20)中的第 1 项是在无监督的情况下,致力于使 2 个不同模态之间具有最大相关性,即两者的距离最小;而第 2 项是在相同标签的有监督情况下,致力于使 2 个模态能够各自产生具有可区分性的特征表示.

那些来自于社交网站上的图像-文本的共现数据,在人类概念理解层面上两者之间是存在语义相关性的,但是在特征层面上两者之间并没有关系,且属于异构模态特征,存在较大的语义鸿沟.经过上述系列操作,将存在语义相关的成对的图像-文本数据转化成在具体特征形式上的最大相关,即在特征层次上将图像数据和对应的文本数据建立起关联,从而使两者之间差异更小,如式(20)所示,这一定程度上缓解了异构模态特征之间的鸿沟,且使各个模态具有优异的判别能力.

2.4 多模态注意力融合网络的情感分类

受人类视觉注意力启发的注意力模块提供了一种机制来推断局部特征对于整体特征的相对重要性.鉴于它能够提供完整的可微性和可解释性来发掘网络关注的重点,目前已经在许多神经网络的应用中作为默认的组成部分.注意力模块可以是只关注整体特征中某一特定部分的硬性注意力机制,也可以是通过重要性的概率分布来分配给所有特征的软性注意力机制.本文主要选择软性注意力机制来展开后续的研究.

尽管 2.3 节中获得的判别性视觉特征表示 $\boldsymbol{U}_p^{\mathrm{T}}f(\boldsymbol{X}^p)$ 和判别性语义特征表示 $\boldsymbol{U}_q^{\mathrm{T}}g(\boldsymbol{X}^q)$ 两者在整体特征层面上是最大相关的且具有判别性的,即两者之间在特征层次上的语义差距最小,而在多模态情感分析任务中图像视觉特征和文本语义特征的贡献度不是同等重要的,因此,需要进一步探测视觉特征和语义特征之间更深层次的内部联系.于是提出一种先后序列化生成语义注意力和视觉注意力的 co-attention 网络,进一步发掘最大相关的判别性视觉特征和判别性语义特征之间更深层次的内部情感关系.由于视觉特征和人类高层情感概念理解之间存在情感鸿沟,为了形成更好的易于理解的视觉特征表示,本文首先关注基于语义的视觉注意力.

为了便于形式化表示,分别简写 $\boldsymbol{U}_p^{\mathrm{T}}f(\boldsymbol{X}^p)$ 和 $\boldsymbol{U}_q^{\mathrm{T}}g(\boldsymbol{X}^q)$ 为 $\boldsymbol{v}_{\mathrm{I}}$ 和 $\boldsymbol{v}_{\mathrm{S}}$.如图 3 中①所示,首先同时输入 $\boldsymbol{v}_{\mathrm{I}}$ 和 $\boldsymbol{v}_{\mathrm{S}}$ 到相同构造的 2 个全连接神经网络 f_{I} 和 f_{S} 中进一步提取成对的视觉和语义特征表示,其中 f_{I} 为学习视觉特征的网络,而 f_{S} 为学习语义特征的网络.在 $\boldsymbol{v}_{\mathrm{I}}$ 和 $\boldsymbol{v}_{\mathrm{S}}$ 共同逐层学习的过程中,通过使用一个单层的神经网络来结合视觉特征和语义特征,然后使用 $softmax$ 层来生成一个视觉注意力分布,相关操作为

$$\boldsymbol{h}_{\mathrm{I}}=\tanh(\boldsymbol{W}_{v_{\mathrm{I}}}f_{\mathrm{I}}(\boldsymbol{v}_{\mathrm{I}})\odot\boldsymbol{W}_{v_{\mathrm{S}}}f_{\mathrm{S}}(\boldsymbol{v}_{\mathrm{S}})),\tag{21}$$

$$\boldsymbol{\alpha}=softmax(\boldsymbol{W}_{h_{\mathrm{I}}}\boldsymbol{h}_{\mathrm{I}}+\boldsymbol{b}_{h_{\mathrm{I}}}),\tag{22}$$

其中,$\boldsymbol{W}_{v_{\mathrm{I}}},\boldsymbol{W}_{v_{\mathrm{S}}},\boldsymbol{W}_{h_{\mathrm{I}}},\boldsymbol{b}_{h_{\mathrm{I}}}$ 是参数,使用 $\odot$ 表示视觉特征表示和语义特征表示的结合,其中视觉特征 $f_{\mathrm{I}}(\boldsymbol{v}_{\mathrm{I}})\in\mathbb{R}^d$ 和语义特征 $f_{\mathrm{S}}(\boldsymbol{v}_{\mathrm{S}})\in\mathbb{R}^d$ 具有相同的特征维度 d,通过对应交互视觉特征 $f_{\mathrm{I}}(\boldsymbol{v}_{\mathrm{I}})$ 和语义特征 $f_{\mathrm{S}}(\boldsymbol{v}_{\mathrm{S}})$ 从而形成视觉语义特征 $f_{\mathrm{IS}}(\boldsymbol{v})$,为了更加频繁地深入交互特征元素,继续学习 $f_{\mathrm{IS}}(\boldsymbol{v})$ 使其特征元素全部关联到 d 维特征空间中,从而形成具有特征之间内部关联的新的视觉语义特征 $\boldsymbol{h}_{\mathrm{I}}$,因此可得对应于 $\boldsymbol{h}_{\mathrm{I}}$ 中特征的注意力概率 $\boldsymbol{\alpha}\in\mathbb{R}^d$,是一个 d 维向量.

基于每一个特征 i 的视觉注意力概率 α_i,新的判别性视觉特征表示通过视觉特征的权重和来构造,即:

$$\tilde{\boldsymbol{v}}_{\mathrm{I}}=\sum_i\alpha_i f_{\mathrm{I}}(\boldsymbol{v}_i).\tag{23}$$

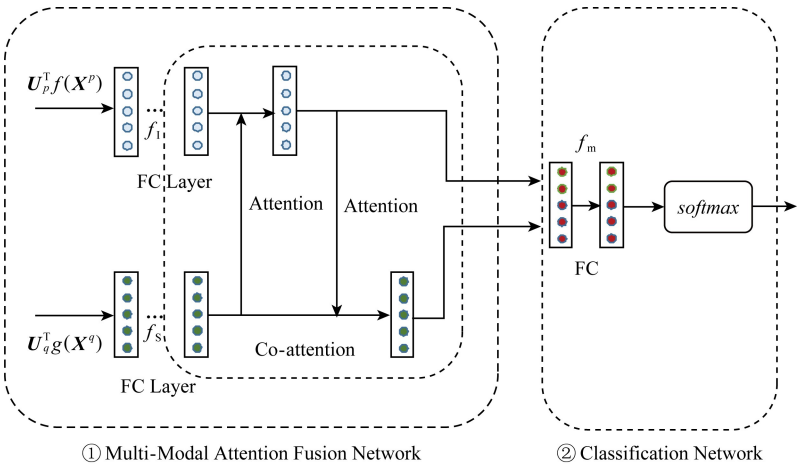

Fig. 3 Schematic sketch of multi-modal attention fusion network for sentiment classification

图 3 多模态注意力融合网络的情感分类图解

然后使用新的判别性视觉特征表示 $\tilde{\boldsymbol{v}}_{\mathrm{I}}$ 来形成基于视觉的语义注意力.

为了构建更好的信息表示,除了用文本语义特征来引导图像视觉特征的注意力,还进一步探讨基于视觉的语义注意力来研究视觉和语义的相互影响.基于视觉的语义注意力与基于语义的视觉注意力操作过程相似.为了获得语义的注意力分布,首先利用 $\tilde{\boldsymbol{v}}_{\mathrm{I}}$ 结合相对应的语义特征,然后基于概率分布形成新的判别性语义特征表示 $\tilde{\boldsymbol{v}}_{\mathrm{S}}$,详细的操作为

$$\boldsymbol{h}_{\mathrm{S}}=\tanh(\boldsymbol{W}_{\tilde{\boldsymbol{v}}_{\mathrm{I}}}\tilde{\boldsymbol{v}}_{\mathrm{I}}\odot\boldsymbol{W}_{\boldsymbol{v}_{\mathrm{S}}}f_{\mathrm{S}}(\boldsymbol{v}_{\mathrm{S}})),\tag{24}$$

$$\boldsymbol{\beta}=softmax(\boldsymbol{W}_{\boldsymbol{h}_{\mathrm{S}}}\boldsymbol{h}_{\mathrm{S}}+\boldsymbol{b}_{\boldsymbol{h}_{\mathrm{S}}}),\tag{25}$$

$$\tilde{\boldsymbol{v}}_{\mathrm{S}}=\sum_i\beta_i f_{\mathrm{S}}(\boldsymbol{v}_i).\tag{26}$$

同理,式(24)~(26)中的参数设置与基于语义的视觉注意力的等式设置相同.

总之,基于语义的视觉注意力和基于视觉的语义注意力是一个交互影响的过程,通过交互来形成更好的有利于图像和文本进行深层融合的特征表示.为了探索图像和文本之间更深层次的内部关联,可以尝试多次序列化地迭代交互视觉特征和语义特征,即形成嵌套的 co-attention 网络.

通过合并 $\tilde{\boldsymbol{v}}_{\mathrm{I}}$ 和 $\tilde{\boldsymbol{v}}_{\mathrm{S}}$ 生成图像视觉和文本语义的融合表示 $\boldsymbol{v}_{\mathrm{m}}$,即:

$$\boldsymbol{v}_{\mathrm{m}}=\tilde{\boldsymbol{v}}_{\mathrm{I}}\oplus\tilde{\boldsymbol{v}}_{\mathrm{S}},\tag{27}$$

其中,$\oplus$是连接操作.在网络学习的过程中,隐藏层可以自动地结合视觉的和文本的情感表示.

在获得了融合特征 $\boldsymbol{v}_{\mathrm{m}}$ 之后,通过 2 层全连接神经网络 f_{m} 进一步捕获更深层次的内部关联,将最后一个全连接层的输出通过 $softmax$ 层产生分类标签的分布,如图 3 中②所示,该过程简要描述为

$$\hat{\boldsymbol{y}}=softmax((\boldsymbol{W}_{f_{\mathrm{m}}})f_{\mathrm{m}}(\boldsymbol{v}_{\mathrm{m}})+\boldsymbol{b}_{f_{\mathrm{m}}}),\tag{28}$$

其中,$\boldsymbol{W}_{f_{\mathrm{m}}}\in\mathbb{R}^{C\times d}$ 和 $\boldsymbol{b}_{f_{\mathrm{m}}}\in\mathbb{R}^{C}$ 是参数,C 是标签的数量,在多模态注意力融合网络模型的设置中,$\boldsymbol{v}_{\mathrm{I}}$ 和 $\boldsymbol{v}_{\mathrm{S}}$ 的输入到最后的分类是一个端到端的过程,该模型使用分类交叉熵计算基于反向传播的训练的批量损失.

$$Loss=\frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}y_{i,j}\,\mathrm{lb}(\hat{y}_{i,j}),\tag{29}$$

其中,N 是一批中的总共样本数量,C 是情感类别的数量.$\boldsymbol{y}_i$ 表示来自于训练批次中第 i 个样本的独热的真实向量,$\hat{y}_{i,j}$ 表示对于相同样本中类别 j 的预测概率.

3 实验分析

本节首先介绍实验中要用到的 5 个数据集,其中 3 个是根据 ANP[19,25]从不同的社交网络上爬取的,另外 2 个是来自于公开的数据集[4];然后介绍了本文实验中的一些设置;最后通过实验来评估本文提出方法的性能,大致包括 2 部分内容:1)从整体情感分类性能的角度来比较本文提出方法和其他对比方法的实验结果的差异;2)从局部模型设置合理性的角度来确定整体模型中的 2 个关键部分对情感分类结果的影响.

3.1 数据集

在目前的多模态情感分析中,由于存在一些可以构建的具有英文描述的图像-文本对的多模态情感数据集,而缺乏公开的具有中文描述的多模态情感数据集,故在本文后续的实验中主要讨论英文

描述的图文多媒体情感数据集.但是本文提出的模型同样也适用于具有中文描述的多模态情感数据集,这是因为本文提出的模型主要关注的是构建视觉语义和文本语义之间的深层关联交互,与文本语言的表现形式关系不大.语言形式对模型的影响将在今后进一步的工作中验证.

由此,首先利用不同的情感关键词查询视觉中国官网的搜索引擎来构筑数据集.具体而言,利用视觉情感本体库(VSO)中 3 244 个 ANP[19] 作为情感关键词从视觉中国网站上的 Getty 专区爬取 38 363 条图像-文本对,称其为 VCGI 数据集;此外,从 3 244 个 ANP[19] 中随机选出 300 个 ANP 作为情感关键词,又从相同的网站上爬取 37 158 条图像-文本对,称其为 VCGII 数据集.

此外,多语言视觉情感本体库 MVSO 是由来自于 12 种语言(例如中文、英文等)的 15 600 个概念构成,这些概念和图像中表达的情感和情绪密切相关.类似于 VSO 数据集,这些概念也以 ANP 的形式定义.与 VCG 数据获取的方式相同,利用 MVSO[25] 中提供的英文语言 ANP,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 MVSO-EN 数据集.

文献[4]中公开了带有 3 个标注(积极、中性、消极)的 Flickr 图像 ID,幸运的是 Flickr 提供了 API,其能通过提供的唯一 ID 获得 1 张图像的元数据(描述、上传日期、标签(tags)等),因此利用公开的所有 ID 从 Flickr 网站上爬取了 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.

对于来自于 Getty 图像的 2 个数据集,由于存在极少量中文描述的数据集,则删除那些描述是中文的图像-文本对,同时为了获得更丰富的文本语义信息,则删除那些英文描述少于 20 个字符的图像-文本对;对于 MVSO-EN 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都是三者都有,但至少有 1 个).由于文本中存在一些不是词汇的内容,而是以链接、符号等明显不含语义信息的内容形式展示,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 MVSO-EN 数据集中图像的情感极性标签来自于 ANP 的情感分数值,而 Flickr 数据集中图像的情感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.因此得到了本文在实验中使用的 5 个数据集,分别为 VCGI,VCGII,MVSO-EN,Flickr-2,Flickr-3,其具体信息统计如表 1 所示:

Table 1 Statistic of The Datasets

表 1 实验使用数据集统计

Dataset	Positive	Neutral	Negative	Total
VCGI	18 847	0	15 837	34 684
VCGII	18 134	0	16 184	34 318
MVSO-EN	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

3.2 实验设置

VCG 数据来自于视觉中国网站的 Getty 专区,其图像的文本描述相对正式和简洁.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.MVSO-EN 数据集和 Flickr 数据集均来自社交网站 Flickr,不同的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,则文本长度长短不一且差别较大,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.

每一个词向量的维度设置为 300,在训练过程中微调词向量来适应本文使用的情感数据集.文本模态特征提取网络的卷积核在实验中使用了 3 个不同的卷积核尺寸,分别为 3,4,5,且针对每一个卷积核尺寸采用了 20 个滤波器.此外,针对所有的图像都调整其为相同的大小 224×224.在实验中总共有 2 个端到端的优化过程:1)多模态深度多重判别性相关分析的优化,除了在最后关联层上采用线性(linear)激活函数,其他网络层的输出均连接到 ReLU 激活函数;2)多模态注意力融合网络的分类交叉熵的优化,每一个全连接层(除最后一个)的输出均连接到 ReLU 激活函数,最后一个全连接层的

输出采用 *softmax* 进行分类.但是这 2 个优化的过程均使用小批量的 RMSprop 方法[29]来优化网络.为了防止过拟合,实验中整体模型上均采用 Dropout 策略,具体设定 Dropout 的值为 0.5.

本文实验主要评估提出的方法在二分类(积极、消极)目标和三分类(积极、消极、中性)目标上的效果.针对情感分类准确性评估和局部模型效用评估的所有实验中,每个实验均从各自对应数据集中随机选取 80%用于训练,20%用于测试.

3.3 实验 1:情感分类准确性评估

3.3.1 对比方法

为了证明提出方法的有效性,首先比较其与仅用图像和仅用文本进行情感分析的方法,然后进一步比较其与其他相关的图文融合情感分类方法的性能.对比方法说明有 4 种:

1) S-Visual. 利用文献[30]中提出的基于迁移学习的视觉情感分析方法,不同的是本文实验利用 VGG-16net 网络模型.

2) S-Text. 利用本文提出的文本模态特征提取网络,并通过 *softmax* 层对文本进行情感分类.

3) CNN-Multi. 由 3 个 CNN 组成.预训练的文本 CNN 和图像 CNN 分别抽取文本和图像的特征表示,然后拼接 2 个特征向量输入到另一个仅有 4 个全连接层的 multi-CNN 结构.文本 CNN 中的卷积层用的是二维卷积,每一个文本样本的维度像单通道图像一样被调整为 50×50 的大小[5].

4) DNN-Multi. 方法同 CNN-Multi,不同的是利用本文提出的视觉模态特征提取网络和文本模态特征提取网络分别抽取图像和文本的特征表示,然后拼接 2 个特征向量输入到另一个有 4 个全连接层的结构中.

3.3.2 结果与讨论

表 2 展示了本文方法和对比方法在 2 个 VCG 数据集上的比较结果.如表 2 所示,本文提出的层次化深度关联融合网络的方法 DDC+co-attention 和 DANDC+co-attention 的分类效果明显优于单模态图像 S-Visual 和单模态文本 S-Text 的分类效果,说明学习图文多媒体内容的特征能更好地理解用户的情感.此外,尽管 CNN-Multi 在多模态情感分析的任务上取得了一定的效果,然而其特征提取的网络模型比较简单,故修改 CNN-Multi 网络结构的 DNN-Multi 方法取得了更优异的效果,这一定程度上说明设计合适的网络结构有益于学习好的特征表示以更好地服务于情感分类.

Table 2 Accuracy of Different Methods on VCGI and VCGII Dataset

表 2 在 VCGI 和 VCGII 数据集上不同方法的准确率 %

Methods	VCGI	VCGII
S-Visual	65.82	75.71
S-Text	67.58	76.92
CNN-Multi	71.56	77.86
DNN-Multi	71.86	79.63
DDC+co-attention	**73.24**	**83.86**
DANDC+co-attention	**74.42**	**85.52**

Notes: The bold values are the accuracy obtained by our method.

然而 CNN-Multi 和 DNN-Multi 都是首先分别提取图像和文本的特征然后再进行融合,不是共同地学习成对的图像-文本数据,而社交媒体上共现的图像-文本数据往往是存在语义概念相关的,若分别提取图像特征和文本特征后再进行特征融合,这会割裂图像与文本之间对应的语义关联.本文提出的方法是同时共同地学习图像-文本的共现数据,且效果也优于 CNN-Multi 和 DNN-Multi,这表明在多模态情感分析任务上同时处理成对的图像-文本的共现数据是必要的.如表 2 所示,提出的方法在 VCGI 和 VCGII 数据集上相较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 3 分别展示了本文方法和对比方法在 MVSO-EN 数据集和 Flickr 数据集上的实验结果.尽管 MVSO-EN 数据集和 Flickr 数据集都是来自于 Flickr 社交网站,但是它们数据集的构造方式略有不同,其中 MVSO-EN 数据集和 VCG 数据集的构造方式相同,则针对 MVSO-EN 数据集的实验评估,采取了与表 2 中 VCG 数据集同样的对比方式,且本文的方法 DDC+co-attention 和 DANDC+co-attention 都展示了优异的性能.

Table 3 Accuracy of Different Methods on MVSO-EN and Flickr Dataset

表 3 在 MVSO-EN 和 Flickr 数据集上不同方法的准确率 %

Methods	MVSO-EN	Flickr-2	Flickr-3
S-Visual	66.06	79.36	59.17
S-Text	63.24	73.24	56.53
CNN-Multi	70.68	81.22	61.69
DNN-Multi	72.23	82.18	62.13
DDC+co-attention	**84.55**	**85.92**	**63.97**
DANDC+co-attention	**83.46**		

Notes: The bold values are the accuracy obtained by our method.

此外,由于 Flickr 数据集的标签来自于人工标注,故没有图像的 ANP 信息,则在 Flickr 数据集上不能评估 DANDC+co-attention 的性能,表 3 中空白表示无实验数据.但是由于本文使用的 Flickr 数据集来自于人工标注,其标签相比更准确,同时为了证明本文提出的 DDC+co-attention 同样适用于三分类的目标,故针对 Flickr 数据集,在二分类目标和三分类目标上都进行了分类性能评估,其中在 Flickr-2 数据集上是为了评估二分类目标,而在 Flickr-3 数据集上是为了评估三分类目标,且在 Flickr-2 和 Flickr-3 这 2 个数据集上 DDC+co-attention 均较对比方法展示了更好的性能.

3.4 实验 2: 局部模型效用评估

尽管表 2 和表 3 的实验已经展示了本文提出的方法可以达到更好的情感分类效果,但是在本文提出的层次化深度关联融合网络的模型中,不仅考虑了经过多模态深度多重判别性相关分析的优化而生成的最大相关的判别性视觉特征表示和判别性语义特征表示,还在多模态注意力的融合网络中序列化地研究了图像视觉特征和文本语义特征之间的协同关注(co-attention).为了探讨这 2 部分模型的设置对图像和文本融合的情感分类结果的贡献度以及合理性,则分别做实验来评估这 2 个部分的性能.

3.4.1 对比方法

首先,通过设定实验来评估多模态深度多重判别性相关分析的合理性.对比方法设置为:

1) DNN-S.利用 DNN-Multi 方法中的 DNN 网络结构分别提取图像和文本的特征,然后拼接特征向量输入 $softmax$ 层进行情感分类.

2) DC-S.利用文献[17]中提出的深度相关性分析的方法,不同于文献[17]中的网络结构,而是利用本文提出的视觉模态特征提取网络和文本模态特征提取网络来共同提取图像和文本的最大相关的视觉和语义的映射特征,将图文映射特征融合后通过 $softmax$ 层进行情感分类.

3) DDC-S.利用本文 DDC 的方法共同地提取图像和文本的最大相关的判别性视觉和语义的映射特征,将视觉和语义映射特征融合后通过 $softmax$ 层进行情感分类.

4) DANDC-S.利用本文 DANDC 的方法共同地提取图像和文本的最大相关的判别性视觉和语义的映射特征,将视觉和语义映射特征融合后通过 $softmax$ 层进行情感分类.

总之,前 3 组实验设置是为了评估简单的特征融合(DNN-S)、具有深度相关分析的特征映射(DCS)、具有深度多重判别性相关分析的特征映射(DDC-S)这三者在情感分类上的性能差异,而 DANDC-S 是为了评估在深度多重判别性相关分析阶段,融入图像中层语义特征对分类结果的影响.

其次,通过设定实验来评估多模态协同注意力(co-attention)设置的合理性,对比方法设置为:

1) co-attention-2.若称 2.4 节中提出的 co-attention 模型中基于语义的视觉注意力和基于视觉的语义注意力为 1 层 co-attention,则 co-attention-2 方法大致同 co-attention 模型,不同之处在于基于 2.4 节 1 层 co-attention 中形成的 $\tilde{\boldsymbol{v}}_{\mathrm{I}}$ 和 $\tilde{\boldsymbol{v}}_{\mathrm{S}}$ 再序列化地进行一次基于语义的视觉注意力和基于视觉的语义注意力来建模图像视觉和文本语义,称其为嵌套的 2 层 co-attention.

2) same-co-attention. 2.4 节中提出的 co-attention 模型是先后序列化地生成视觉的注意力和语义的注意力,不同于 co-attention 模型的序列化操作,而 same-co-attention 是同时平行地生成视觉的注意力和语义的注意力.具体而言,基于 2.4 节中视觉语义特征 $\boldsymbol{h}_{\mathrm{I}}$ 产生的注意力概率向量 $\boldsymbol{\alpha}$,共同构筑新的判别性视觉特征表示 $\tilde{\boldsymbol{v}}_{\mathrm{I}}$ 和判别性语义特征表示 $\tilde{\boldsymbol{v}}_{\mathrm{S}}$.

3.4.2 结果与讨论

图 4 的实验结果展示了在 5 个数据集上利用多模态深度多重判别性相关分析(DDC-S 和 DANDC-S)的分类性能均优于 DNN-S 和 DC-S,这说明利用多重深度判别性相关分析来学习最大相关的判别性特征表示是可行且必要的.此外,在视觉模态上共同学习图像视觉特征和图像中层语义特征的 DANDC-S 在除了 VCGI 数据集外的所有数据集上的分类结果上均优于仅利用视觉特征的 DDC-S.然而,在 VCGI 数据集上 DANDC+co-attention 的情感分类性能要优于 DDC+co-attention,如表 2 所示.此外,在表 3 中的 MVSO-EN 数据集上,DANDC+

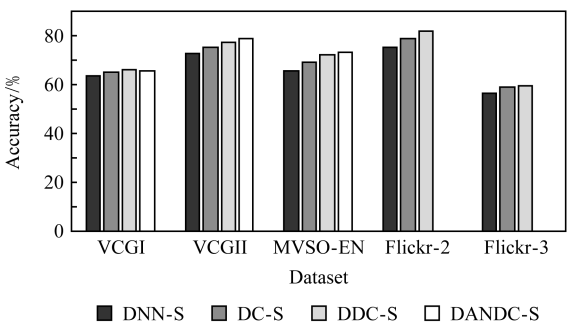

Fig. 4 Evaluate the performance of deep and discriminative correlation analysis on five datasets

图 4 在 5 个数据集上评估深度判别性相关分析的性能

co-attention 的性能次优于 DDC+co-attention,但是在多重深度判别性相关分析阶段 DANDC-S 的分类性能要优于 DDC-S,如图 4 所示.这表明融入图像的中层语义特征(ANP)在一定程度上对多模态情感分类的性能是起积极作用的.

然后,进一步评估 co-attention 方法设置的合理性,本实验仅利用提出的 DDC 模型生成的最大相关的判别性视觉特征和判别性语义特征做基准,比较其与 same-co-attention 和 co-attention-2 的性能差异.如图 5 所示,在 5 个数据集上的对比实验均显示序列化的 co-attention 相比于非序列化的 same-co-attention 都取得了略好的情感分类效果,这说明先后序列化生成视觉的注意力和语义的注意力的设置有益于探测图像视觉和文本语义之间的深层内部关联.另外,为了探讨嵌套 co-attention 网络的性能,在 5 个数据集上也相应做了实验评估.如图 5 所示,在 Flickr-2 和 Flickr-3 数据集上的分类结果 co-attention-2 略优于 co-attention,但在其他数据集上效果反而不如 co-attention 的性能.由于增加 co-attention 网络的迭代交互的次数,不仅会使模型变得更复杂,而且在实验中需要更多的训练时间.很显然,嵌套序列交互后的效果没有明显的提升甚至在几个数据集上反而下降,因此,实验设置中没有必要去设置更多嵌套 co-attention 层的模型.

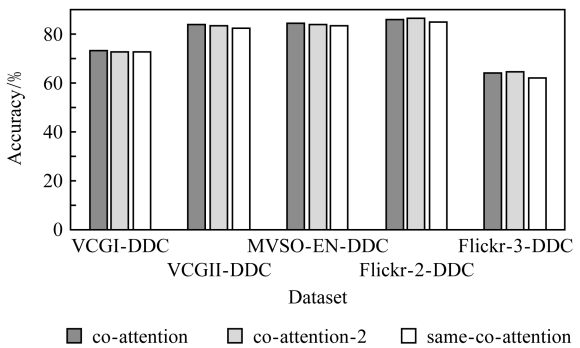

Fig. 5 Evaluate the performance of co-attention settings on five datasets

图 5 在 5 个数据集上评估 co-attention 设置的性能

4 总 结

近年来,多模态情感分析已经成为一个日益重要的研究热点,尤其在社交媒体大数据的环境下.本文提出一个新颖的层次化深度关联融合网络结构用于多模态情感分析.在提出的方法中,首先依赖提出的多模态深度多重判别性相关分析的模型共同学习最大相关的判别性视觉特征表示和判别性语义特征表示.基于这 2 种特征表示,进一步提出多模态注意力融合网络的情感分类模型,首先,序列化地生成语义的视觉注意力和视觉的语义注意力来交互视觉和语义,从而获得图像的和文本的更深层和更判别性的特征表示;然后,合并最新的图像视觉特征和文本语义特征后并通过全连接神经网络学习后再用于训练情感分类器.在 5 个真实数据集上已经评估了提出方法的有效性,且实验结果表明本文提出的层次化深度关联融合网络的图文媒体情感分析方法要优于其他相关的方法.

在未来的工作中将考虑不同的文本语言类型、图像的区域化语义,设计更好的多模态网络提取结构以及更合理的注意力网络模型用于情感分析,此外,还将研究更好的特征融合策略以进一步提高异构多模态特征融合的性能.

参 考 文 献

[1] Wang Min, Cao Donglin, Li Lingxiao, et al. Microblog sentiment analysis based on cross-media bag-of-words model [C] //Proc of the 2014 Int Conf on Multimedia Computing and Service. New York: ACM, 2014: 76-80

[2] Cao Donglin, Ji Rongrong, Lin Dazhen, et al. A cross-media public sentiment analysis system for microblog [J]. Multimedia Systems, 2016, 22(4): 479-486

[3] Poria S, Cambria E, Howard N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content [J]. Neurocomputing, 2016, 174: 50-59

[4] Katsurai M, Satoh S. Image sentiment analysis using latent correlations among visual, textual, and sentiment views [C] //Proc of the 2016 Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2016: 2837-2841

[5] Cai Guoyong, Xia Binbin. Convolutional neural networks for multimedia sentiment analysis [C] //Proc of the 2015 Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2015: 159-167

[6] Yu Yuhai, Lin Hongfei, Meng Jiana, et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks [J]. Algorithms, 2016, 9(2): 41-51

[7] Baecchi C, Uricchio T, Bertini M, et al. A multimodal feature learning approach for sentiment analysis of social network multimedia [J]. Multimedia Tools and Applications, 2016, 75(5): 2507-2525

[8] You Quanzeng, Luo Jiebo, Jin Hailin, et al. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia [C] //Proc of the 9th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2016: 13-22

[9] Xu Nan, Mao Wenji. A residual merged neutral network for multimodal sentiment analysis [C] //Proc of the 2nd Int Conf on Big Data Analysis (ICBDA). New York: IEEE, 2017: 6-10

[10] Dos Santos C, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts [C] //Proc of the 25th Int Conf on Computational Linguistics (COLING 2014). Stroudsburg, PA: ACL, 2014: 69-78

[11] Liang Bin, Liu Quan, Xu Jin, et al. Aspect-based sentiment analysis based on multi-attention CNN [J]. Journal of Computer Research and Development, 2017, 54(8): 1724-1735 (in Chinese)
(梁斌, 刘全, 徐进, 等. 基于多注意力卷积神经网络的特定目标情感分析[J]. 计算机研究与发展, 2017, 54(8): 1724-1735)

[12] Chen Ke, Liang Bin, Ke Wende, et al. Chinese micro-blog sentiment analysis based on multi-channels convolutional neural networks [J]. Journal of Computer Research and Development, 2018, 55(5): 945-957 (in Chinese)
(陈珂, 梁斌, 柯文德, 等. 基于多通道卷积神经网络的中文微博情感分析[J]. 计算机研究与发展, 2018, 55(5): 945-957)

[13] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing(EMNLP). Stroudsburg, PA: ACL, 2015: 2539-2544

[14] You Quanzeng, Luo Jiebo, Jin Hailin, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2015: 381-388

[15] Campos V, Salvador A, Giro-I-Nieto X, et al. Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction [C] //Proc of the 1st Int Workshop on Affect & Sentiment in Multimedia. New York: ACM, 2015: 57-62

[16] 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

[17] Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis [C] //Proc of the 2013 Int Conf on Machine Learning. Atlant, GA: Machine Learning Society, 2013: 1247-1255

[18] Dorfer M, Kelz R, Widmer G. Deep linear discriminant analysis [J]. arXiv preprint, arXiv:1511.04707, 2015

[19] Borth D, Ji Rongrong, Chen Tao, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs [C] //Proc of the 21st ACM Int Conf on Multimedia. New York: ACM, 2013: 223-232

[20] Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: A survey [J]. Multimedia Systems, 2010, 16(6): 345-379

[21] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint, arXiv: 1409.0473, 2014

[22] Yang Zicao, He Xiaodong, Gao Jianfeng, et al. Stacked attention networks for image question answering [C] //Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016:21-29

[23] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ : IEEE, 2015: 3156-3164

[24] Siersdorfer S, Minack E, Deng F, et al. Analyzing and predicting sentiment of images on the social Web [C] //Proc of the 18th ACM Int Conf on Multimedia. New York: ACM, 2010: 715-718

[25] Jou B, Chen Tao, Pappas N, et al. Visual affect around the world: A large-scale multilingual visual sentiment ontology [C] //Proc of the 23rd ACM Int Conf on Multimedia. New York: ACM, 2015: 159-168

[26] Li Zuhe, Fan Yangyu, Liu Weihua, et al. Image sentiment prediction based on textual descriptions with adjective noun pairs [J]. Multimedia Tools and Applications, 2018, 77(1): 1115-1132

[27] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint, arXiv:1409.1556, 2014

[28] Kim Y. Convolutional neural networks for sentence classification [J].arXiv preprint, arXiv:1408.5882, 2014

[29] Ruder S. An overview of gradient descent optimization algorithms [J].arXiv preprint, arXiv:1609.04747, 2016

[30] Islam J, Zhang Yangqing. Visual sentiment analysis for social images using transfer learning approach [C] //Proc of the 2016 IEEE Int Conf on Big Data and Cloud Computing. Piscataway, NJ: IEEE, 2016: 124-130

Cai Guoyong, born in 1971. PhD, professor. Senior member of CCF. His main research interests include social media mining, sentiment analysis, machine learning.

Lü Guangrui, born in 1989. PhD candidate. His main research interests include natural language processing, multimodal sentiment analysis, and deep learning.

Xu Zhi, born in 1977. PhD, associate professor. Member of CCF. His main research interests include pattern recognition, image classification, and machine learning.