计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000074

基于视觉方面注意力的图像文本情感分析模型

袁景凌　丁远远　盛德明　李　琳
武汉理工大学计算机科学与技术学院　武汉 430070

摘　要　社交网络已经成为人们日常生活中不可分割的一部分,对社交媒体信息进行情感分析有助于了解人们在社交网站上的观点、态度和情绪。传统情感分析主要依赖文本内容,随着智能手机的兴起,网络上的信息逐渐多样化,除了文本内容,还包括图像。通过研究发现,在多数情况下,图像对文本有着支持增强作用,而不独立于文本来表达情感。文中提出了一种新颖的图像文本情感分析模型(LSTM-VistaNet),具体来说,LSTM-VistaNet 模型未将图像信息作为直接输入,而是利用 VGG16 网络提取图像特征,进一步生成视觉方面注意力,赋予文档中核心句子更高的权重,得到基于视觉方面注意力的文档表示;此外,还使用 LSTM 模型对文本情感进行提取,得到基于文本的文档表示。最后,将两组分类结果进行融合,以获得最终的分类标签。在 Yelp 餐馆评论的数据集上,所提模型在精确度上达到了 62.08%,比精度较高的模型 BiGRU-mVGG 提高了 18.92%,验证了将视觉信息作为方面注意力辅助文本进行情感分类的有效性;比 VistaNet 模型提高了 0.32%,验证了使用 LSTM 模型可以有效弥补 VistaNet 模型中图像无法完全覆盖文本的缺陷。

关键词:视觉方面注意力;LSTM;多模态;情感分析;社交图像
中图法分类号　TP391.1

Image-Text Sentiment Analysis Model Based on Visual Aspect Attention

YUAN Jing-ling,DING Yuan-yuan,SHENG De-ming and LI Lin
School of Computer Science and Technology,Wuhan University of Technology,Wuhan 430070,China

Abstract　Social network has become an integral part of our daily life. Sentiment analysis of social media information is helpful to understand people's views, attitudes and emotions on social networking sites. Traditional sentiment analysis mainly relies on text. With the rise of smart phones, information on the network is gradually diversified, including not only text, but also images. It is found that, in many cases, images can enhance the text rather than express emotions independently. We propose a novel image text sentiment analysis model (LSTM-VistaNet). Specifically, this model does not take the picture information as the direct input, but uses the VGG16 network to extract the image features, and then generates the visual aspect attention, and gives the core sentences in the document a higher weight, and get a document representation based on the visual aspect attention. In addition, we use the LSTM network to extract the text sentiment and get the document representation based on text only. Finally, we fuse the two groups of classification results to obtain the final classification label. On the Yelp restaurant reviews data set, our model achieves an accuracy of 62.08%, which is 18.92% higher than BiGRU-mVGG, which verifies the effectiveness of using visual information as aspect attention assisted text for emotion classification; It is 0.32% higher than VistaNet model, which proves that LSTM model can effectively make up for the defect that images in VistaNet model cannot completely cover text.

Keywords　Visual aspect attention, LSTM, Multimodel, Sentiment analysis, Social images

1　引言

随着网络信息的爆炸式增长,社交网络已经成为人们获取信息的重要来源之一。同时,人们习惯于通过这些平台分享自己的情感体验。基于社交媒体内容的情感分析有许多真实应用,如股市预测[1]、政治选举[2]甚至医疗保健[3]。同时,社交媒体的快速发展促使了多模式内容的出现。例如,Twitter 用户通常将图像与推文一起发布,以使推文更具表现力;Flickr 用户通常会在发布图像的同时给出描述,对图像进行解释;Yelp 用户会在评论一道菜品时配上菜品对应的图像,以供其他用户参考。因此,视觉文本内容的情感检测具有重要的实践意义。

了解用户评论包含的情感的技术称为情感分析,给定一份文档,对其进行情感分类,可分为 2 类(如积极和消极)、3 类(如积极、中性、消极)、5 类(如 1 类－5 类)等。目前,已提出许多情感分类技术,其中基于深度学习的方法更为有效[4-7]。

到稿日期:2020-10-14　返修日期:2021-04-19
基金项目:国家社会科学基金(15BGL048)
This work was supported by the National Social Science Fund of China(15BGL048).
通信作者:袁景凌(yuanjingling@126.com)

文档的不同部分如何提供不同的信息是情感分析中的一个新概念。表达多愁善感的句子(例如,“沙拉是新鲜美味的,汤汁是纯粹的,完美”)可能比一个中立的句子(例如,“我午餐吃了科布沙拉和巧克力蛋糕”)更重要。相应地,有些词(例如,“美味”)会更有影响力。这些信息的重要性差异可以通过注意力来捕捉,它赋予重点句子(或单词)更高的权重。

但是,如今的文档不仅仅包含文本。智能手机和平板电脑的存在,使得随时随地拍照非常方便,因此目前大多数的文档都是多模态的。“多模态”可能指图像、视频、音频等,本文只关注图像。博客帖子和评论通常都包括照片,以实现更加生动地描述发布者的经验。例如,Yelp 网站上洛杉矶最受欢迎的餐馆 Bottega Louie 有 1.5 万份评论(截至 2019 年),包含 2.6 万张图像。

评论中的图像和文本具有协同作用,图 1 给出了 Yelp 上关于商家 California Fish Market Restaurant 的评论示例,评论包括 3 张图像和 3 段文字。我们观察到,第一,评论中的图像往往只关注一种食物(例如,“Crudo Calabrese”“Oysters”“Lobster Ravioli”);第二,图像中展示的食物往往都是文本中提到的重点。

图 1 Yelp 上的评论示例

Fig. 1 Example of a comment on Yelp

评论包括令人特别难忘的事物或“方面”的图像,因为这些图像有助于更多地强调这些事物。传达信息的主要手段仍然是文本,特别是关于情感的信息,照片起着辅助作用,而不是一个独立的作用,它们自己不会讲述故事。有了这一观察,我们没有直接将照片作为特征输入情感分类模型,而是将它们作为视觉均值,突出评论中最重要的句子或“方面”[4]。

由于图像数量有限,无法完全覆盖到文本评论中所有重要的句子或单词,我们后续使用 LSTM(Long Short-Term Memory)模型专门对文本进行情感分析,用所得到的情感分析结果对使用视觉信息作为注意力网络得到的情感分析结果进行“补充纠正”,从而得到更加准确的情感分析结果。

本文的贡献主要如下:

(1)本文提出了一个名为 LSTM-VistaNet 的神经网络模型,该模型将视觉信息视为句子层次上的对齐来源。评论中的每一句都可以包含一些“方面”(我们没有预先指定一个“方面”清单)。图像有助于识别评论中的重要句子,即模型在对评论文本情绪进行分类时,应更加注意这些句子。在模型 LSTM-VistaNet 后期,我们使用 LSTM 模型专门针对文本进行情感分析,以此弥补图像无法完全覆盖文本、文本信息中的重要句子或单词无法完全识别,从而导致情感分类结果偏差较大的缺点。

(2)本文利用 Yelp 数据集对美国五大城市的餐厅评论进行了全面的实验。该数据集的训练集为细粒度的 5 级评级,测试集中含有真实标签,是一个绝佳的测试用例。实验结果非常具有竞争力,证明该模型可推广到其他类型的网络文档,如博客文章、推文或任何包含图像的文档。

2 相关工作

2.1 文本情感分析

传统情感分析模型主要针对文本[8],如文献[9]提出了一种深度学习体系结构,利用领域之间的语言重叠来推断情感极性。近年来,深度学习在文本分类方面取得了显著进展[4-7,10-11],基于注意力的 RNN 网络[12]的成功使文本分类取得了巨大进展[13]。但它们只依赖文本信息,而我们将图像信息作为视觉方面注意力。有的文献研究方面级情感分析[14],相比之下,我们关注的是整个文档的情感。

2.2 视觉情感分析

视觉情感分析被定义为图像分类[15],是一个日益热门的研究方向[16],目前视觉情感分析方法可分为以下 3 种类型:第一种是从图像中提取低级特征,如文献[17]提取图像的颜色直方图和 SIFT 特征来推断其情感;第二种是从图像中提取中级特征,如文献[18]检测到有 1 200 个形容词-名词对与情感表达密切相关,并用所提出的概念检测器库建立了一种新的中层表示;第三种是从图像中提取高级特征,如文献[19]设计了一个 CNN 架构,然后用渐进策略进行微调,以减少噪声对情绪分析的影响。这些单模态方法只依赖图像,即视觉特征。

虽然在过去几年中,对单一模态数据的情感分析取得了很大的成功,但它不能有效处理信息多样的社交媒体数据,因此多模态情感分析应运而生。

2.3 文本+视觉情感分析

由于社交媒体的日益普及,多模态情感分析作为一项具有挑战性的任务,近年来引起了学者们广泛的研究兴趣[7],目前的研究可以分为以下两种类型。

第一类研究分别处理不同来源的特征。一些研究将不同的特征连接到一个完整的特征向量中,然后使用连接向量输入情感分类器进行学习。文献[20]结合从深度 CNN 中提取的视觉和文本特征,将它们发送到多核学习分类器来分析情感。文献[6]将视觉和文本特征嵌入到统一的词向量表示中,然后利用逻辑回归来识别微博推文的情感。

第二类研究共同处理不同来源的特征。文献[6]提出以视觉特征为指导注意力的 LSTM,提取情感相关词汇进行情感分析。文献[7]提出了一种融合神经网络(MNN)的模型,用于提取图像和文本特征,利用早期/后期残差 RMNN 融合多模态特征进行情感分类。文献[16]通过共同考虑图顺序特征学习、视频-用户-标签交互以及标签的相关性,提出了图信息传播的多视图表示交互式嵌入模型,为微视频推荐标签。文献[21]首先将物体和场景识别为显著检测器,以提取图像的深层语义特征,然后提取对理解整个推文的情绪非常重要的单词,并汇总具有视觉语义特征、对象和场景的那些翔实单词的表示形式。这些研究有两个特点:1)没有考虑到图像只

是起到对文本内容的增强作用；2）数据集中图像与文本一一对应。在本文的问题中，一段文本可与多张图像关联，每张图像都与文本的特定部分相关。我们通过更多地关注与图像相关的句子（假设更重要）来学习有助于情感分类的对齐方式。

3 LSTM-VistaNet 模型

本节主要描述本文提出的 LSTM-VistaNet 模型。给定一份文档 C（如评论），对于文档中的每份文件 $c\in C$，它的文本成分为 L 个句子 $s_i(i\in(1,\cdots,L))$ 的序列，句子 s_i 由 T 个单词 $w_{i,t}(t\in(1,\cdots,T))$ 组成；它的视觉成分为 M 张图像 a_j 的集合，$a_j\in\{a_1,a_2,\cdots,a_M\}$。每份文件 c 都有一个情感标签。情感分析任务可描述为：给定文档 C，学习一个分类函数，对未知情感的文件进行分类。

VistaNet 有 3 层结构，如图 2 左半部分所示。底层是具有 soft attention 的词编码层，我们将词表示转化为句子表示。中间层是我们转换的句子编码层，在视觉方面注意力的帮助下，将句子表示形成文档级表示。顶层是为文档分配情感标签的分类层。接下来我们将分层进行讲解。

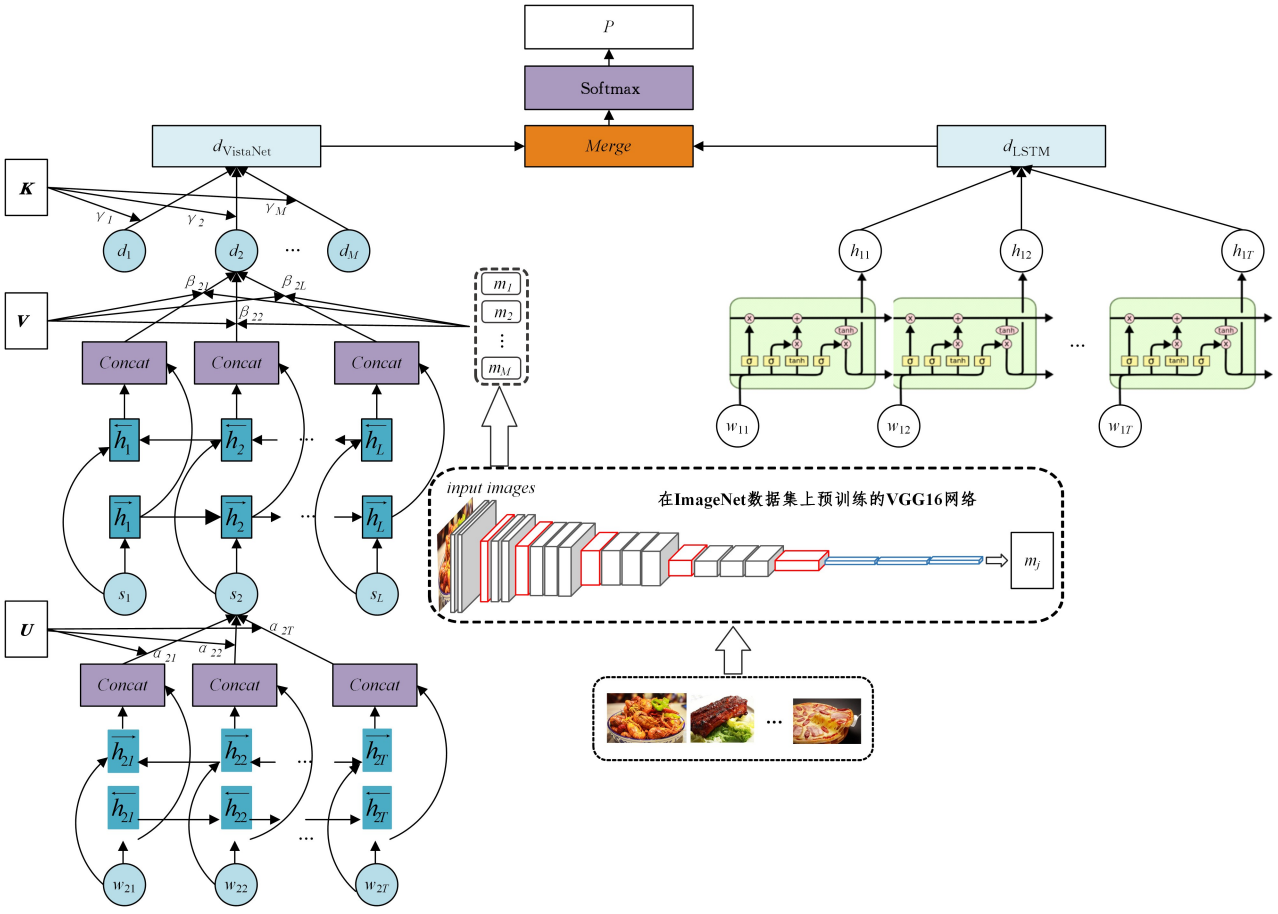

图 2 LSTM-VistaNet 网络结构图

Fig. 2 LSTM-VistaNet network structure diagram

3.1 VistaNet 对文本和图像进行情感分析

3.1.1 词编码层

对于每个单词 $w_{i,t}$，我们根据学习嵌入矩阵 W_e 得到它的嵌入表示 $x_{i,t}$，嵌入矩阵 W_e 由预先训练的单词嵌入模型[22]初始化，并在训练期间对其进行了调整。

$$x_{i,t}=W_e w_{i,t},t\in\{1,\cdots,T\} \tag{1}$$

对单词嵌入的整个序列进行编码，我们使用具有 GRU 单元的双向递归神经网络（Bi-RNN）[23-24]，输入为词嵌入序列 $x_{i,t}$，输出为一个新的隐藏状态向量 $\boldsymbol{h}_{i,t}=[\overrightarrow{\boldsymbol{h}_{i,t}},\overleftarrow{\boldsymbol{h}_{i,t}}]$，$\overrightarrow{\boldsymbol{h}_{i,t}}$ 由前向 RNN 产生，$\overleftarrow{\boldsymbol{h}_{i,t}}$ 由后向 RNN 产生。

$$\boldsymbol{h}_{i,t}=\text{Bi-RNN}(w_{i,t}) \tag{2}$$

句子中所有单词的权重是不相等的，有些单词对于情感分析更加重要。因此，当从句子的单词中获取句子的表示时，每个单词将被分配一个权重。我们采用了一种软注意力机制在单词之间学习和分配这些权重。

$$u_{i,t}=\boldsymbol{U}^{\mathrm{T}}\tanh(W_w\boldsymbol{h}_{i,t}+b_w) \tag{3}$$

$$\alpha_{i,t}=\frac{\exp(u_{i,t})}{\sum_t\exp(u_{i,t})} \tag{4}$$

$$s_i=\sum_t\alpha_{i,t}\boldsymbol{h}_{i,t} \tag{5}$$

我们将词嵌入 $w_{i,t}$ 的表示 $h_{i,t}$ 通过一层具有非线性激活函数 tanh 的神经元，得到在注意力空间的表示，其中 W_w 表示词嵌入 $w_{i,t}$ 的表示 $\boldsymbol{h}_{i,t}$ 所对应的权重矩阵。然后，我们将投影与上下文向量 $\boldsymbol{U}^{\mathrm{T}}$（在训练过程中随机初始化和学习）相乘得到 $u_{i,t}$，表示 $w_{i,t}$ 的相对重要性。使用 Softmax 进行归一化，产生注意力权重 $\alpha_{i,t}$。最后，句子 s_i 的向量表示由对当前句中所有单词表示 $\boldsymbol{h}_{i,t}$ 和它们的注意权重 $\alpha_{i,t}$ 加权求和产生。

3.1.2 句子编码层

从词编码层得到句子级别的表示后，对重要的句子赋予更大的权重，即使用视觉方面注意力，将句子表示聚合到文档

级表示中。对于每一个输入句子s_i，双向 RNN 输出的隐藏状态向量为：$\boldsymbol{h}_i=[\overrightarrow{\boldsymbol{h}_i},\overleftarrow{\boldsymbol{h}_i}]$。

$$\boldsymbol{h}_i=\text{Bi-RNN}(s_i) \tag{6}$$

得到文档 d 的最终表示的一个方法是使用基于文本的 soft attention 方案[9]。然而，我们使用视觉注意力增强机制。一个文档可能与几个图像相关联，这些图像可能与不同的“方面”有关。给定一个图像，句子可以是不同的信息，即图像可以突出显示文档的不同但重要的部分。

我们首先对输入图像进行编码。VGG 卷积神经网络已被证明在学习许多图像相关任务的图像表示方面是有效的[4,24-25]。我们利用 VGG16 模型获得输入图像 α_j 的表示m_j，并在分类层之前的最后一个完全连接层(FC7)得到输出。图像表示m_j是图像 α_j 编码的 4096 维向量。

$$m_j=VGG(a_j) \tag{7}$$

关于每个图像表示m_j，我们学习句子表示h_i的注意力权重$\beta_{j,i}$。

$$p_j=\tanh(W_p m_j+b_p) \tag{8}$$

$$q_i=\tanh(W_q h_i+b_q) \tag{9}$$

$$v_{j,i}=\boldsymbol{V}^{\mathrm{T}}(p_j\cdot q_i+q_i) \tag{10}$$

$$\beta_{j,i}=\frac{\exp(v_{j,i})}{\sum_i \exp(v_{j,i})} \tag{11}$$

为了学习这些注意力权重，我们首先将图像表示m_j和句子表示h_i分别与其对应权重W_p和W_q相乘，并将结果投影到一个注意空间，其中b_p和b_q表示偏差；然后通过一个非线性激活函数 tanh 将m_j和h_i缩放到相同的值范围内；最后分别输出p_j和q_i。为了学习句子的图像特定注意权重，我们让图像投影p_j与句子投影q_i以元素乘法与求和两种方式交互。学习向量$\boldsymbol{V}^{\mathrm{T}}$在词级上起着类似于$\boldsymbol{U}^{\mathrm{T}}$的全局上下文注意力的作用，生成了一个注意力值$v_{j,i}$，用 Softmax 进行归一化获得$\beta_{j,i}$。

计算$v_{j,i}$需要元素乘法与求和来确保图像和句子之间的交互作用有意义。元素乘法可保障在计算注意力权重$\beta_{j,i}$时，Softmax 函数不会清除视觉部分的影响；求和可保障文本部分的效果不会因为视觉部分的稀疏性而明显减弱。因此，两者同样重要。

使用图像特定的注意力权重$\beta_{j,i}$，将句子表示h_i聚合成图像特定的文档表示d_j，具体表达式如下：

$$d_j=\sum_i \beta_{j,i} h_i \tag{12}$$

我们对文档中的每张图像都应用这种视觉方面注意机制，产生一组特定于方面的文档表示d_j，$j\in\{1,\cdots,M\}$。在分类之前，所有d_j都需要聚合到最终的文档表示 d 中。每个文档中的图像数量可能不同，因此我们学习重要性权重γ_j，这表明每个特定于图像的文档表示d_j将对最终结果做出多大贡献。

$$k_j=\boldsymbol{K}^{\mathrm{T}}\tanh(W_d d_j+b_d) \tag{13}$$

$$\gamma_j=\frac{\exp(k_j)}{\sum_j \exp(k_j)} \tag{14}$$

特定于图像方面的文档表示d_j通过具有非线性激活函数 tanh 的神经元投射到注意空间，其中W_d表示每个文档所对应的权重，b_d为偏差。表示d_j重要性的标量k_j是通过与全局上下文注意力向量$\boldsymbol{K}^{\mathrm{T}}$($\boldsymbol{K}^{\mathrm{T}}$在训练过程中随机初始化和学习)相乘所得。如图 2 所示，文档表示d_j与文本对图像的注意力权重共同被聚合到最终的文档表示d_{VistaNet}中，计算过程如式(15)所示：

$$d_{\text{VistaNet}}=\sum_j \gamma_j d_j \tag{15}$$

3.2 LSTM 对文本进行情感分析

长短期记忆(LSTM)网络由循环神经网络(Recurrent Neural Network，RNN)扩展而来，可用于解决 RNN 长期依赖的问题。LSTM 是一种特殊的循环神经网络，也具有链状结构，但是与 RNN 的重复模块的结构不同。网络模块示意图如图 2 右上部分所示。本文使用 LSTM 模型对文本进行情感分析，并对 LSTM 模型的输出h_t进行拼接，将其命名为文档级向量表示 d_{LSTM}，其计算过程如下：

$$d_{\text{LSTM}}=H(h_t) \tag{16}$$

3.3 LSTM-VistaNet 模型结果获取

对由 VistaNet 模型和 LSTM 模型得到的文档向量表示进行拼接得到最终的文档表示，将其作为全连接层 Dense 层的输入得到文档向量表示。

$$d=Dense(d_{\text{VistaNet}},d_{\text{LSTM}}) \tag{17}$$

获得文档的向量表示 d 后，将其作为 Softmax 情感分类器的输入，产生 ρ(ρ 为评论等级 1－5)类上的概率分布。

$$\rho=\text{Softmax}(W_c d+b_c) \tag{18}$$

其中，W_c表示 d 所对应的权重，b_c为偏差。通过最小化情感分类结果的交叉熵误差，以监督的方式训练模型：

$$loss=-\sum_d \log \rho_{(d,l)} \tag{19}$$

其中，l 是评论的真实标签。

4 实验

4.1 数据集

数据集为从 Yelp.com 网站上食品和餐馆类别爬取的在线评论数据集，涵盖了美国 5 个不同的主要城市：波士顿(BO)、芝加哥(CH)、洛杉矶(LA)、纽约(NY)和旧金山(SF)。数据集总共有超过 4.4 万条评论，包括 24.4 万张图像。每个评论至少有 3 张图像，目的是验证视觉方面的注意效果。

由于 Yelp 的评论包括 1 到 5 的评分，将其分为 5 种情绪水平，把每个评级作为一个类别。为了保持不同类别的示例数量平衡，将数据的 80％用于训练，5％用于验证，15％用于测试。由于有些城市的数据集较小，我们将 5 个城市的训练集和验证集合并，而单独存储测试集，以便在评估模型时保持统计属性。

4.2 训练细节

对于预处理，我们使用 python 中的 NLTK 工具包对句子和单词进行标记，并从训练集和验证集中出现 3 次以上的单词中构建词汇表，再将其他不常见的单词替换为特殊的 UNK 标记，然后使用 GloVe 模型对 200 维词嵌入矩阵W_e进行初始化。词嵌入在训练过程中进行微调，以适应当前域。

为了获得在验证集上的最佳性能，所有模型都使用超参数进行调整。对于词编码和句子编码，GRU 单元为 50 维，由于双向 RNN 的缘故，可得到 100 维词编码与句子编码。上下文向量$\boldsymbol{U}$、$\boldsymbol{V}$和$\boldsymbol{K}$对于单词、句子和文档的注意空间同样为

100 维。对于图像，我们使用 VGG16 模型进行特征提取，图像特征向量为在分类层之前 FC7 层的输出。我们使用在 Image Net 数据集上预训练的 VGG16 模型来初始化图像编码器的权重，在训练过程中图像编码器的所有权重均为固定数值。在训练中，我们使用 RMSprop 进行基于梯度的优化，批量大小为 32。

4.3 对比模型

(1) BiGRU-aVGG 和 BiGRU-mVGG。BiGRU-aVGG 和 BiGRU-mVGG 是将 BiGRU[26] 从文本中学习的特征向量和 VGG 从图像中学习的特征向量连接起来，并输入到分类层。对于图像，我们使用从 ImageNet 数据集预训练的 VGG16 模型作为编码器。图像特征取自分类层前的全连接层 FC7。因为每个评论都有多个图像，所以有两个变体，对于图像特征向量的处理方式，BiGRU-aVGG 使用平均池化，BiGRU-mVGG 使用最大池化。在训练期间，图像编码器的权重是固定的(VistaNet 同理)。

(2) TFN-aVGG 和 TFN-mVGG。TFN-aVGG 和 TFN-mVGG 将 Tensor Fusion Network 作为主要组件[5]。利用张量融合层将由 HAN-ATT 模型获得的文本特征向量与 VGG 网络获得的视觉特征向量相结合，通过情感分类层得到最终的情感标签。我们同样应用平均池化与最大池化产生两个对比模型 TFN-aVGG 和 TFN-mVGG。

(3) VistaNet。Truong 等[4] 假设在情感分析任务中视觉信息对文本起支撑作用，利用 VGG16 网络提取视觉特征，生成基于视觉的方面级注意力，指导文本进行情感分析，提出了 VistaNet 网络。

4.4 实验结果分析

4.4.1 LSTM-VistaNet 模型结果分析

表 1 列出了对比实验的结果，以及各自方法的关键属性，除了显示 5 个城市的结果外，还显示了所有城市的平均值，以及相对于性能最低的基础模型的改进程度(见表 1 的第 1 行－2 行)。

表 1 实验结果对比

Table 1 Comparison of experimental results

(单位：%)

模型	文本特征	视觉特征	层次结构	视觉方面注意力	BO	CH	LA	NY	SF	平均值	提升
TFN-aVGG	√	√	√		46.35	44.69	43.91	43.79	42.81	43.89	—
TFN-mVGG	√	√	√		48.25	47.08	46.70	46.71	47.54	46.87	6.8
BiGRU-aVGG	√	√			51.23	51.33	48.99	49.55	48.60	49.32	12.4
BiGRU-mVGG	√	√			53.92	53.51	52.09	52.14	51.36	52.20	18.9
VistaNet	√	√	√	√	63.81	65.74	62.01	61.08	60.14	61.88	41.0
LSTM-VistaNet	√	√	√	√	64.93	66.86	61.92	61.29	61.22	62.08	41.4

有趣的是，TFN 模型在文本特征和视觉特征之间提供了丰富的交互作用，在比较方法中(除 LSTM 模型外)表现最差，其准确率分别是 43.89%(TFN-aVGG)和 46.87%(TFN-mVGG)。该实验结果验证了我们的假设，即将图像作为模型的输入不如强调图像对文本的增强作用有效。

BiGRU-aVGG(平均池)精度为 49.32%。通过最大池化，BiGRU-mVGG 获得了 52.20%的较高精度，比 BiGRU-aVGG 提高了 5.8%，比 TFN-aVGG 提高了 18.9%。这些模型包含了评论文本和图像的连接特性。

本文模型 LSTM-VistaNet 在精确度上达到了 62.08%，比精度较高的模型 BiGRU-mVGG 提高了 18.9%，这证明了将视觉信息作为方面注意力指导文本进行情感分类的有效性；比 VistaNet 模型提高了 0.32%，表明使用 LSTM 模型对文本提取情感，可有效弥补 VistaNet 模型中图像无法完全覆盖文本的缺陷。

4.4.2 LSTM-VistaNet 模型消融实验结果分析

为了研究 LSTM-VistaNet 体系结构的各个组件各自的贡献，我们进行了消融分析实验，从基础模型 BiRNN 开始，逐步添加一个组件来构建完整的体系结构，汇总结果如表 2 所列。

表 2 LSTM-VistaNet 模型消融实验结果

Table 2 Ablation experimental results of LSTM-VistaNet model

(单位：%)

BiRNN	Hierarchical Structure	Soft Text Attention	Visual Aspect Attention	LSTM	BO	CH	LA	NY	SF	平均值	提升
√					57.70	60.01	56.74	56.59	55.84	56.83	—
√	√				60.39	64.39	59.08	59.58	59.18	59.54	4.8
√	√	√			63.38	64.47	60.65	59.85	58.34	60.56	6.6
√	√	√	√		63.81	65.74	62.01	61.08	60.14	61.88	8.9
√	√	√	√	√	64.93	66.86	61.92	61.29	61.22	62.08	9.2

我们从基本模型 BiRNN 开始，只依赖于文本。如表 2 第 1 行所列，5 个城市的情感分类平均准确值为 56.83%。探索文档的层次结构，通过对句子表示应用最大池化，情感分类结果与基本模型 BiRNN 相比提高了 4.8%，如表 2 第 2 行所列，这展示了文本层次结构建模的价值。如果我们在聚合句级表示时单独应用基于文本的软注意层，则比 BiRNN 提高了 6.6%，如表 2 第 3 行所列。通过进一步结合视觉方面的注意力，实现了比基础模型效果提升 8.9%的结果，如表 2 第 4 行所列，平均准确率为 61.88%。利用 LSTM 模型弥补图像无法完全覆盖文本的漏洞，我们将模型效果提升到了62.08%，比基础模型提升了 9.2%，如表 2 第 5 行所列。

结束语 本文提出了一种利用视觉信息对文本信息进行增强，并用 LSTM 弥补视觉信息无法完全覆盖文本重要信息缺陷的新方法——LSTM-VistaNet。VistaNet 模型具有 3 层结构，从单词到句子，然后到特定于图像的文档表示，最后到

最终的文档表示，使用图像作为对齐方式来指出文档中的重要句子，用 LSTM 模型对图像无法完全覆盖文本的缺陷进行弥补，在 Yelp 数据集上获得了具有竞争力的实验结果。本实验中未考虑文本没有完全覆盖图像的情况，在后续实验中将对这一问题进行研究。

参 考 文 献

[1] LI X, XIE H, CHEN L, et al. News impact on stock price return via sentiment analysis[J]. Proceedings of the Knowledeg Based System, 2014, 69: 14-23.

[2] KAGAN V, STEVENS A, SUBRAHMANIAN V S. Using twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election[J]. Proceedings of the IEEE Intelligent Systems, 2015, 30(1): 2-5.

[3] YADAV S, EKBAL A, SAHA S, et al. Medical sentiment analysis using social media: towards building a patient assisted system [C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). 2018: 2790-2797.

[4] TRUONG T Q, LAUW H W. VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis[C]// Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-2019). 2019: 305-312.

[5] YU D, FU J, MEI T, et al. Multi-level attention networks for visual question answering[C]// Proceedings of the Computer Vision and Pattern Recognition (CVPR). 2017: 4187-4195.

[6] PORIA S, CHATURVEDI I, CAMBRIA E, et al. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis[C]// Proceedings of 2016 IEEE 16th International Conference on Data Mining (ICDM). 2017: 439-448.

[7] XU N, MAO W, CHEN G. Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1), 371-378.

[8] CHANSW K, CHONGMW C. Sentiment analysis in financial texts [J]. Proceedings of the Decision Support Systems, 2017, 94(2017): 53-64.

[9] SEVERYN A, MOSCHITTI A. Twitter sentiment analysis with deep convolutional neural networks[C]// Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 959-962.

[10] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015: 2267-2273.

[11] NGUYEN T, KAVURI S, LEE M. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips [J]. Proceedings of the Neural Networks. 2019, 118: 208-219.

[12] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2014.

[13] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]// Proceedings of North American Chapter of the Association for Computational Linguistics (HLT-NAACL). 2016: 1480-1489.

[14] TANG D, QIN B, LIU T. Aspect level sentiment classification with deep memory network[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2016: 214-224.

[15] TRUONG Q, LAUW H W. Visual sentiment analysis for review images with item-oriented and user-oriented CNN[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017: 1274-1282.

[16] LI M, GAN T, LIU M, et al. Long-tail Hashtag Recommendation for Micro-videos with Graph Convolutional Network[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 509-518.

[17] SIERSDORFER S, MINACK E, DENG F, et al. Analyzing and predicting sentiment of images on the social web[C]// Proceedings of the 18th ACM International Conference on Multimedia. 2010: 715-718.

[18] YOU Q, LUO J, JIN H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]// Proceedings In AAAI. 2015: 381-388.

[19] BORTH D, JI R, CHEN T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C]// Proceedings of the 21st ACM International Conference on Multimedia. 2013: 223-232.

[20] YOU Q, JIN H, LUO J. Visual sentiment analysis by attending on local image regions[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017: 231-237.

[21] XU N, MAO W. 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.

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

[23] CHO K, VAN M, GULCEHRE C, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation[C]// Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014). 2014: 1724-1734.

[24] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// Proceedings of the International Conference on Learning Representations. 2015.

[25] YUE W, WAEL A, PREMKUMAR N. Multi-Modality Image Manipu-lation Detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 9543-9552.

[26] TANG D, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2015: 1422-1432.

YUAN Jing-ling, born in 1975, doctor, is a member of China Computer Federation. Her main research interests include mechine learning, intelligent analysis and green computing.

（责任编辑：柯颖）