计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100041

基于注意力神经网络的多模态情感分析

林敏鸿　蒙祖强
广西大学计算机与电子信息学院　南宁 530004
(minhonglin1202@gmail.com)

摘　要　近年来,越来越多的人热衷于在社交媒体上同时用图片和文本等媒体形式表达自己的感受与看法,使得以图片和文本为主要内容的多模态数据不断增长。相比单模态数据,多模态数据包含的信息更丰富,更能揭示用户的真实情感。对这些海量多模态数据的情感进行分析有助于更好地理解人们的态度和观点,具有广泛的应用场景。为了解决多模态情感分类任务中的信息冗余的问题,在张量融合方案的基础上,提出了一种基于注意力神经网络的多模态情感分析方法。该方法构造了基于注意力神经网络的文本特征提取模型和图像特征提取模型,突出了图像情感信息关键区域和包含情感信息的单词,使得各单模态特征表达更简练精确。将各模态的张量积作为多模态数据的联合特征表达,采用主成分分析法剔除联合特征的冗余信息,进而使用支持向量机获取多模态数据的情感类别。在两个真实的 Twitter 图文数据集上对所提模型进行了评估,实验结果表明,与其他情感分类模型相比,该方法在分类准确率、召回率、F1 指标和准确率上都有较大的提升。

关键词:社交媒体;多模态数据;情感分析;注意力机制;张量融合
中图法分类号　TP391

Multimodal Sentiment Analysis Based on Attention Neural Network

LIN Min-hong and MENG Zu-qiang
School of Computer,Electronics and Information,Guangxi University,Nanning 530004,China

Abstract　In recent years,more and more people are keen to express their feelings and opinions in the form of both pictures and texts on social media,and the scale of multimodal data including images and texts keeps growing. Compared with single mode data,multimodal data contains more information. It can better reveal the real emotion of users. Sentiment analysis of these huge amounts of multimodal data helps to better understand people's attitudes and opinions. In addition,it has a wide range of applications. In order to solve the problem of information redundancy in multimodal sentiment analysis task,this paper proposes a multimodal sentiment analysis method based on tensor fusion scheme and attention neural network. This method constructs the text feature extraction model and image feature extraction model based on attention neural network to highlight the key areas of image emotion information and words containing emotion information,so as to make the expression of each feature more concise and accurate. It fuses each modal feature using tensor fusion method in order to obtain the joint feature vector. Finally,it uses support vector machine for sentiment classification. The experimental results of this model on two real Twitter data sets show that compared with other sentiment analysis models,this method has a great improvement in precision rate,recall rate,F1 score andaccuracy rate.

Keywords　Social media,Multimodal data,Sentiment analysis,Attention mechanism,Tensor fusion

1　引言

研究机构 We Are Social 在 2019 年 1 月 31 日发布了最新的 Global Digital 2019 Reports[1)]。该报告显示全球社交媒体(包括 Twitter,Facebook,Instagram 等)的用户数量已增长到 35 亿。平均每个用户每天都将自己 1/3 的互联网时间花在社交媒体上。越来越多的人们热衷于在社交媒体上表达自己的看法或观点。在社交媒体上每天都会有数以亿计的数据记录产生,这其中的大量数据是以文本和图像联合的形式出现,构成了海量的多模态数据。在海量的多模态数据中蕴含着丰富的情感信息,对多模态数据的情感分析有利于了解人们对某些事件的态度和看法,在票房预测[1]、政治选举[2-3]、股市预测[5-6]等方面有着很大的应用价值。因此,多模态情感分析在学术界和相关行业中受到了越来越多的关注。

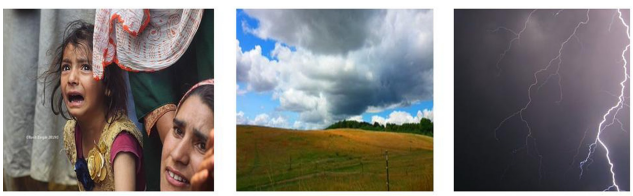
在社交媒体的图文数据中,文本和图像都分别包含了各自的情感信息,它们彼此不同又相辅相成。图 1 给出了 Twitter 的几个图文示例。其中,图 1(a)的图像和文本都表明了这条推文内容所传达的情绪是消极的;图 1(b)中的图片

1) https://wearesocial.com/global-digital-report-2019

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61762009)
This work was supported by the National Natural Science Foundation of China (61762009).
通信作者:蒙祖强(zqmeng@126.com)

对不同的人来说会有不同的感受，有的人会觉得空旷孤寂，有的人会觉得美丽，图片的情感极性并不强烈，而所对应的文本表达出了非常强烈的积极倾向，因此这条推文是积极的；图 1(c)中，文本是一个陈述句，没有明显体现情感的词语，但结合图片来分析，推文整体的情感倾向是消极的。在图文多模态数据中，文本和图片所包含的信息一般是相辅相成的。相比文本或图像的单模态数据，多模态数据包含了更为全面的信息，能更好地展现和揭示用户的真实情感。

然而，多模态情感分析在当下仍然是一项非常具有挑战性的任务。首先，不同模态数据所包含的情感信息是不同的，对多模态数据的情感分析需要有效地获取各模态数据的情感特征表示。我们注意到，对人而言，一张图片中并不是所有的图像区域都与情感表达相关。如图 1(a)所示，我们会更关注小女孩的面部，被她的悲伤表情所感染。而在阅读一段文字时，‘cry’‘damaged’‘killed’等词更能引起我们的情感共鸣。因此，在进行特征提取时应该突出图像情感关键区域和文本情感关键词的影响力。其次，不同模态的数据采用不同维度和不同属性的底层特征来表达[7]。与传统的单一模态情感分析相比，多模态情感分析需要正确结合各模态信息的有效方式，以最大化地保存各模态信息与各模态间的交互信息。



A girl is crying near a damaged house where two militants were killed in a gunfight with security forces in Awantipora village in Pulwama district.

(a)

Best of my travel in Mecklenburg-Vorpommern, #Germany, Day 2 : Impressive #sunset, nice landscape and #clouds!

(b)

I just want to go out and get an ice cream…

(c)

图 1 Twitter 图文数据示例

Fig. 1 Examples of images and texts from twilter

2 相关工作

2.1 文本情感分析

早在 2000 年，情感分析就成为了自然语言处理领域中最活跃的研究方向之一。文本情感分析的方法可以分为基于情感字典的方法和基于机器学习的方法。

基于情感字典的方法[9-12]是使用情感词典，根据文本语句中的情感词来计算文本的情感分数以获得情感倾向。这些情感字典是由人工或者半人工构建的。文献[9]提出先抽取出句子中包含的形容词和副词的短语，将短语与“excellent”和“poor”之间的互信息之差作为短语语义情感计算，将短语的平均语义情感倾向作为句子的情感倾向。文献[10]提出了一种基于词汇的方法来提取文本中的情绪——语义取向计算器(SO-CAL)，它使用带有语义取向(极性和强度)注释的单词字典来分析文本的情感。这类方法的实现简单且快速，但是构建情感字典是相当困难的，而且现有的大多数情感字典都是基于某个领域总结而来的，不具备普适性。

基于机器学习的情感分析方法[13-18]是现阶段研究中更常见的方法。2002 年，文献[13]就在情感分析任务上应用机器学习方法对电影评论进行了情感分类。随着深度学习的快速发展，在计算机视觉、自然语言处理等领域的应用中都获得了优越的结果，广大学者也越来越重视深度学习在文本情感分析中的应用与研究。文献[14-15]通过卷积等操作显式获取文本的局部和全局的信息，能够快速地处理句子以获取文本特征表达，从而进行文本情感分类。文献[16]提出了基于深度学习的篇章情感分类方法，采用了循环卷积和循环相关操作(Circular Convolution and Circular Correlation)来计算评价文档中的单词与该评价文档的评价对象之间的相关性权重，并将文档中词向量的加权和作为文档向量的表达，从而进行情感分类。文献[17]提出了一种文档级的情感分类方法。该方法首先利用卷积神经网络或长短时记忆模型学习句子表示，然后利用门控递归神经网络对句子进行自适应编码以获取文档表示。该模型在情感分类任务中取得了较好的效果。

2.2 图像情感分析

由于图像的情感是更为抽象主观的，图像情感分析任务相比文本情感分析更为复杂。文献[20]提出了一种基于图像低级特征的方法，采用视觉词袋模型获取的图像特征和颜色分布来预测图像情感。文献[21]提出了一种基于图像中级特征的方法，构建了 1 200 个形容词-名词对(ANP)，并以此抽取视觉情感本体，从而对图像进行情感分类。随着深度学习的发展，神经网络模型获取图像高级特征的能力也越来越强。文献[22]提出了一种新的能够提高局部区域识别力的深度神经网络(NIN)来进行图像情感分析。文献[23]采用注意力机制自发检测到图像情感相关的视觉区域，证实了基于注意力机制在情感分析任务中的有效性。

2.3 多模态情感分析

现今，多模态情感分析研究[26-30]尚处于起步阶段，大多在文本情感分析和图像情感分析的现有技术的基础上进行研究。文献[26]为文本和图像构建了一个统一的词包模型，通过此模型获取文本和图像的表示，运用 Logistic 回归分类器进行情感分类。文献[27]运用情感词典提取文本特征，采用形容词-名词对(ANP)来抽取图片中的视觉情感本体，并将其作为其情感特征，文本特征及图片特征进行加权融合，进而进行情感分类。随着深度学习的快速发展，基于深度学习的多模态情感分析也取得了一系列的成果。文献[28]提出了一种跨模态一致性回归(CCR)模型。文献[28]认为图像特征、文本特征以及图像和文本的联合特征的情感倾向应该是一致的，因此在文本特征和图像特征的融合过程中增加一致性约束来获取联合特征向量，进而进行情感分类。文献[29]提出了一种与视觉注意力机制相结合的树形循环神经网络(TreeLSTM)，以此获取图像和文本之间的相关性特征来进行情感分类。特别地，视频本身可以看作是一种多模态数据，视频情感分析研究也取得了一些进展。文献[30]提出了张量融合的方法来提取视频数据的联合特征表示。实验证明，张量融合能够有效地学习到各模态的交互信息，能有效地提高情感分类的效果。

在社交媒体多模态数据的情感分析研究中主要有两个挑战。首先，不同模态数据所包含的情感信息是不同的，在进行多模态数据的情感分析时需要有效地获取各模态数据的情感特征。其次，不同模态的数据采用不同维度和不同属性的底层特征来表达[7]。与传统的单一模态情感分析相比，多模态情感分析需要正确结合各模态信息的有效方式，以最大化地保存各模态信息与各模态间的交互信息。文献[30]在视频情

感分析任务中提出了一种张量融合的方案，并证明了张量融合方案在多模态情感分析任务中能够较好地保留多模态数据中各模态之间的交互信息，从而提升多模态情感分析模型的性能。构造的多模态、联合特征向量的维度非常大，等于各模态特征向量维度之积。而且联合特征除了有效的信息之外，还包含了大量的冗余信息。因此，若要模型取得更好的效果，需要有效提取各模态数据的情感信息的方法和剔除冗余信息的机制。近年来，注意力机制已被广泛应用在自然语言处理[31-33]和图像处理领域[18-19,34]的多个方面的研究。研究表明，注意力机制让神经网络在执行预测任务时可以更多关注输入中的相关部分，更少关注不相关的部分，以提高模型的性能。

在张量融合方案中存在的最大问题就是信息冗余。那么，如果对各模态数据的特征表达越精炼，多模态联合特征的冗余信息就越少，模型的计算效率就越高。因此，本文将注意力机制引入多模态情感分析任务，以在对各单模态数据提取特征表达时更多地保留包含情感信息的部分，忽略掉与情感表达无关的部分，使得各单模态特征表达更简练精确，进而使多模态联合特征表达减少了大量的冗余信息。本文在文献[30]提出的张量融合的方案的基础上，提出了一种基于注意力神经网络的多模态情感分析模型(Attention-based Neural Network model for Multimodal sentiment analysis，ANNM)。本模型采用基于注意力机制的方法来构造文本和图像的特征的提取网络，以突出图像情感信息关键区域和包含情感信息的单词。并将各模态的张量积作为多模态数据的联合特征表达，采用主成分分析法对联合特征向量进行降维，进而使用支持向量机获取多模态数据的情感类别。

3 多模态情感分析模型

为了能够更有效地获取各模态的情感特征表示和模学习到模态间的交互信息，本文采用基于注意力机制的方法来构建多模态情感分析模型。本文所提出的 ANNM 模型的整体结构如图 2 所示。首先提出了两个基于注意力机制的单模态特征提取模型，分别用于获取图像和文本的情感特征，然后采用张量融合的策略来获得多模态联合特征表示，从而进行情感分类。

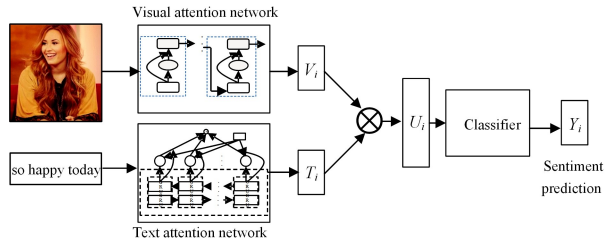


图 2 基于注意力神经网络的多模态情感分析模型的结构

Fig. 2 Framework of attention-based neural network model for multimodal sentiment analysis

3.1 图像特征提取网络

图像的情感信息通常与视觉区域的某一部分联系得更紧密，如图 1(a)所示，人物悲伤的表情比图像中的其他部分更能引起人们的情感共鸣，是图像情感信息更相关区域。因此，提取图像特征时，应突出图1(a)中人物表情这一局部特征而减弱其他部分的影响。对图像进行有侧重的信息提取，使得特征表达更精炼，模型的计算效率更高。

文献[4]提出了一种综合考虑通道域注意力和空间域注意力的权重计算方法。对于每一次卷积计算后产生的多个特征图，模型需要知道哪个特征图应该更重视，特征图的哪一个部分包含的信息更多。因此，注意力权重计算主要分为两部分：1)对各个特征图的权重进行计算；2)对特征图局部权重进行计算。在本模型的图像特征提取网络——基于注意力机制的卷积神经网络(Convolutional Neural Networks with attention for image sentiment analysis，CNNa)中，借鉴文献[4]中的注意力权重的计算方法对卷积层输出的特征图进行注意力加权计算，通过神经网络算出梯度，并且通过前向传播和后向反馈来自主学习从而得到注意力的权重。

本文的图像情感特征提取网络结构如图 3 所示。在这个基于注意力机制的多层卷积神经网络中，共有 13 个卷积层，每个卷积核大小为 3×3，每一个卷积步骤都经历了卷积、注意力权重计算以及特征图加权计算 3 个步骤计算得到最终的注意力特征图。然后，将得到的注意力特征图输入到下一个卷积步骤继续计算。最后，将最终卷积步骤的输出通过全连接层来获取图像情感特征向量。

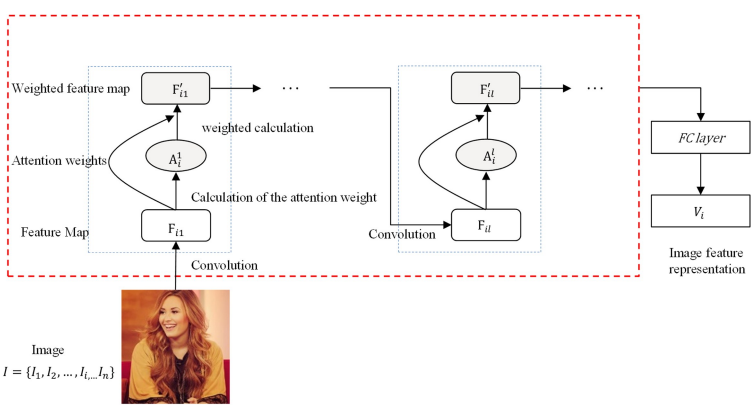


图 3 图像特征提取网络结构

Fig. 3 Architecture of visual attention network

$I=\{I_1, I_2, \cdots, I_i, \cdots, I_n\}$ 表示数量为 n 的图像数据集。基于注意力机制的卷积神经网络(Convolutional Neural Networks with attention，CNNa)完成了 CNNa(I)→V_i 的特征映射。将图片输入 CNNa 模型，获取图像特征向量 V_i。图 3 中，F_{il} 表示第 i 张图片经过第 l 层卷积层后所得到的特征图(Feature Map)。F'_{il} 为经注意力加权后得到的注意力特征图。

其中 F_{il}，$F'_{il}\in\mathbb{R}^{C\times H\times W}$，$C$ 为通道数，H 为特征图的长，W 为特征图的宽。A_i^l 为第 i 个图像的第 l 个特征图的注意力权重，其表达式如下：

$$A_i^l=\{\alpha_{il}^c,\alpha_{il}^s\} \tag{1}$$

其中，α_{il}^c 为第 i 个图像的第 l 个特征图的通道注意力权重。而 α_{il}^s 为第 i 个图像的第 l 个特征图的空间注意力权重。

通道注意力体现了经过卷积后的特征图的每个特征图对于关键信息的贡献大小。通道注意力权重 α_{il}^c 的计算公式如下：

$$\alpha_{il}^c=\sigma(W_1(W_0(global_{\text{avg}}(F_{il})))+W_1(W_0(global_{\max}(F_{il})))) \tag{2}$$

其中，$global_{\text{avg}}(\cdot)$表示全局平均池函数，计算每个特征图的所有特征点的平均值，所得结果特征空间为$\mathbb{R}^{C\times1\times1}$，其中 C 为特征图数；$global_{\max}(\cdot)$表示全局最大池化函数，计算每个特征图的最大特征值，所得结果特征空间为$\mathbb{R}^{C\times1\times1}$，其中 C 为特征图的通道数；$\sigma(\cdot)$为 sigmoid 函数，将结果映射到(0，1)以获得标准的通道注意力权重；通道注意力权重 $\alpha_{il}^c\in\mathbb{R}^{C\times1\times1}$，$C$ 为特征图数。在式(2)中，W_1，W_0 是该神经网络中的参数，可以通过前向传播和后向反馈来自主学习。

空间注意力权重体现了图片局部区域对关键信息的贡献大小，能够找出图片信息中需要被关注的区域。空间注意力权重 α_{il}^s 的计算公式如下：

$$\alpha_{il}^s=\sigma(f^{7\times7}([avg(\alpha_{il}^c\odot F_{il}),\max(\alpha_{il}^c\odot F_{il})])) \tag{3}$$

其中，$\odot$表示逐元素相乘；$avg(\cdot)$为平均池化函数，沿着通道轴对特征点求平均值，输出结果的特征空间为$\mathbb{R}^{1\times H\times W}$。$max(\cdot)$为最大池化函数，沿着通道轴对求最大值，输出结果的特征空间为$\mathbb{R}^{1\times H\times W}$。$avg(\cdot)$和 $max(\cdot)$实现了对特征图的信息的聚合，同时减少了计算量。$[\cdot]$表示拼接操作，输出结果向量空间为$\mathbb{R}^{2\times H\times W}$。$f^{7\times7}(\cdot)$为卷积运算，通过卷积计算来获取特征图不同局部区域对关键信息的影响力。卷积核大小为 7×7，作为该神经网络中参数的一部分，通过前向传播和后向反馈来自主学习。$f^{7\times7}(\cdot)$的输出结果的特征空间为$\mathbb{R}^{1\times H\times W}$。$\sigma(\cdot)$为 $sigmoid$ 函数，将结果映射到(0，1)以获得标准的空间注意力权重。

注意力特征图的计算公式如下：

$$F'_{il}=F_{il}\odot\alpha_{il}^c\odot\alpha_{il}^s \tag{4}$$

最后将注意力特征图作为下一个卷积层的输入继续计算。将最终卷积结构的输出经过一个全连接层转换为一维向量，即最终的图像特征表示 V_i。

3.2 文本特征提取模型

在文本的情感分类任务中，文本的情感信息往往与某些单词更相关。如‘cry’‘damaged’‘killed’等词比‘house’‘village’等词更能体现文本所传达出来的情感。因此，在对文本进行特征提取的过程中应该要增大关键词的影响力。在本模型的文本特征提取网络中使用双向门控循环单元(Bi-directional Gated Recurrent Unit，Bi-GRU)来构建基于注意力的文本特征提取网络。并对 Bi-GRU 层的输出进行加权以突出关键部分的影响力，从而获得更精确的文本特征表达。文本特征提取网络结构如图 4 所示。

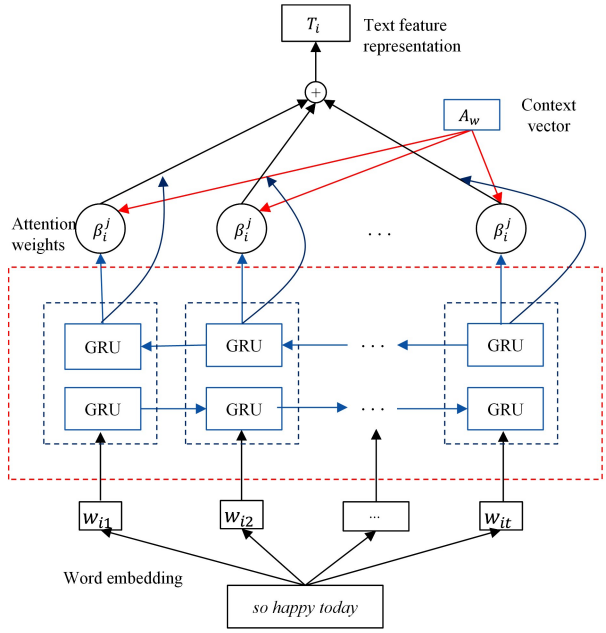


图 4 文本特征提取网络结构

Fig. 4 Architecture of text attention network

记 $T=\{t_1,t_2,\cdots,t_n\}$为大小为 n 的文本记录。本文通过词嵌入工具将单词嵌入向量空间第 i 个文本的第 t 个单词的词向量，用 w_it 表示，第 i 个文本可以表示为$\{w_{i1},w_{i2},\cdots,w_{it},\cdots,w_{iL}\}$，其中 L 为文本长度。单门控循环单元 GRU 的计算过程如下：

$$\begin{aligned}r_t&=\sigma(W_r[h_{t-1},w_{it}]+b_r)\\z_t&=\sigma(W_z[h_{t-1},w_{it}]+b_z)\\\tilde{h_t}&=\tanh(W_{\tilde{h_t}}[r_t*h_{t-1},w_{it}]+b_{\tilde{h_t}})\\h_t&=(1-z_t)*h_{t-1}+z_t*\tilde{h_t}\end{aligned} \tag{5}$$

其中，$[\cdot]$表示两个向量相连接，$*$ 表示对应元素相乘。z_t 为更新门，r_t 为重置门。$\tanh(\cdot)$是激活函数。W_r，W_z，$W_{\tilde{h_t}}$皆为参数，需要训练得到。在 Bi-GRU 神经网络中，将词向量$\{w_{i1},w_{i2},\cdots,w_{it}\}$按正向输入得到对应的前向隐藏层输出$\{\overrightarrow{h_{i1}},\overrightarrow{h_{i2}},\cdots,\overrightarrow{h_{it}}\}$。前向隐藏层输出$\overrightarrow{h_{it}}$的计算如下：

$$\overrightarrow{h_{it}}=GRU(\overrightarrow{h_{i(t-1)}},w_{it}) \tag{6}$$

而将词向量$\{w_{i1},w_{i2},\cdots,w_{it}\}$按反向输入得到对应的后向隐藏层输出$\{\overleftarrow{h_{i1}},\overleftarrow{h_{i2}},\cdots,\overleftarrow{h_{it}}\}$。反向传播状态信息输出$\overleftarrow{h_{it}}$的计算如下：

$$\overleftarrow{h_{it}}=GRU(\overleftarrow{h_{i(t-1)}},w_{it}) \tag{7}$$

由前向隐藏层输出$\overrightarrow{h_{it}}$与反向隐藏层输出$\overleftarrow{h_{it}}$，通过拼接操作得到 Bi-GRU 网络的输出 h_{it}，其计算式如下：

$$h_{it}=[\overrightarrow{h_{it}},\overleftarrow{h_{it}}] \tag{8}$$

其中，$[\cdot]$表示向量的拼接。

所得到的 h_{it} 可以看作是第 t 个单词，包含了上下文信息的表示。与计算视觉注意力相似，文本注意力权重是单词对于文本情绪分类的相关程度的衡量。

$$y_{it}=\tanh(W_o\cdot h_{it}) \tag{9}$$

$$\beta_{it}=\frac{\exp(y_{it}^{\mathrm{T}}A_w)}{\sum_t\exp(y_{it}^{\mathrm{T}}A_w)} \tag{10}$$

首先将 h_{it} 输入一层隐藏层，并用 $tanh$ 函数激活得到 y_{it}。W_o 为隐藏层参数。我们将 A_w 称为上下文向量，其可以看作

是一个关键信息词的查询的向量，它被添加到文本特征提取网络中联合训练得到，能在训练中自主学习信息。将词表示 y_{it} 与 A_w 的点积通过 $softmax$ 函数归一化得到标准化的注意力权重。对隐藏层输出加权求和得到文本特征表示 T_i，其计算过程如下：

$$T_i=\sum_t \beta_{it} h_{it} \tag{11}$$

3.3 多模态融合

张量是多向阵列，可看作向量、矩阵的高阶扩展[35]，其维度被称为张量的阶。向量是一阶张量，矩阵是二阶的张量。

对于 x 阶张量 $\boldsymbol{A}\in\mathbb{R}^{N_1\times\cdots\times N_x}$ 与 y 阶张量 $\boldsymbol{B}\in\mathbb{R}^{M_1\times\cdots\times M_y}$ 之间的张量积为 $A\otimes B\in\mathbb{R}^{N_1\times\cdots\times N_x\times M_1\times\cdots\times M_y}$，定义如下：

$$(\boldsymbol{A}\otimes\boldsymbol{B})_{n_1,n_2,\cdots,n_x,m_1,m_2,\cdots,m_y}=a_{n_1,n_2,\cdots,n_x}b_{m_1,m_2,\cdots,m_y} \tag{12}$$

一阶张量 $\boldsymbol{C}\in\mathbb{R}^n$ 与一阶张量 $\boldsymbol{m}\in\mathbb{R}^m$ 的张量积计算如式(13)所示：

$$(\boldsymbol{C}\otimes\boldsymbol{D})_{i,j}=c_i d_j \tag{13}$$

本文采用张量融合方法[30]对图像特征 $V_i=\{v_1,v_2,\cdots,v_n\}$ 和文本特征 $T_i=\{t_1,t_2,\cdots,t_n\}$ 进行融合。第 i 个图文数据对的联合特征记作 U_i，其计算式如下：

$$U_i=[V_i,1]\otimes[T_i,1] \tag{14}$$

$$U_i=\begin{bmatrix} v_1\cdot t_1 & v_1\cdot t_2 & \cdots & v_1\cdot t_n & v_1 \\ v_2\cdot t_1 & v_2\cdot t_2 & \cdots & v_2\cdot t_n & v_2 \\ \vdots & \vdots & & \vdots & \vdots \\ v_n\cdot t_1 & v_n\cdot t_2 & \cdots & v_n\cdot t_n & v_n \\ t_1 & t_2 & \cdots & t_n & 1 \end{bmatrix} \tag{15}$$

在式(12)、式(13)及式(14)中，$\otimes$ 为求张量积运算。[·]表示拼接操作。首先在每个单模态特征的末尾增加一个值为1的特征点再进行张量积计算。使得在联合特征 U_i 中，不仅包含了图像与文本的模态交互信息，还包含了各单模态特征信息。最后为了便于计算，将 U_i 转换成为向量表示来进行情感分类。张量融合的优点不仅在于能够充分获取模态间的交互信息，还在于能很容易地拓展到更多模态的融合，使得算法的应用性更广，然而也更容易造成冗余，所需计算量也更大。本文算法引入了注意力机制来获取图像和文本的更精确的情感特征，减小了不相关信息的影响，但联合特征向量 U_i 的向量维度仍然较大，而且其包含的交互信息仍有冗余。因此，首先采用主成分分析方法（Principal Component Analysis，PCA）对联合特征进行降维，以减小冗余信息造成的误差，减少计算量，然后再运用支持向量机（Support Vector Machine，SVM）进行情感分类。

支持向量机是机器学习中最为常用的分类模型之一，特别是在图像分类的经典方法中获得了广泛的使用。相比其他分类器，SVM 对于高维度的输入数据和大量的样本均具有良好的适应性，是有着良好泛化能力的预测工具，无论是对于文本情感分类还是对图像情感分类都取得了很好的效果。因此，本文方法使用 SVM 对融合后的特征向量进行分类。

在融合特征向量中，除了包含图文多模态数据关键的交互信息，还包含许多冗余了信息，这些冗余信息对情感分类任务的作用小，还增加了分类器的计算量，因此我们需要对计算结果进行降维处理。本文运用主成分分析方法 PCA（principal Component Analysis）来对数据进行降维，则计算未知样本的情感类别的计算式如下：

$$label=SVM(PCA(\boldsymbol{U})) \tag{16}$$

其中，$\boldsymbol{U}$ 为图文数据的联合特征矩阵。

$$\boldsymbol{U}=\begin{bmatrix} U_1 \\ \vdots \\ U_n \end{bmatrix} \tag{17}$$

4 实验与结果

本节通过实验来评估所提出多模态情感分析模型的性能。

4.1 数据集

由于带标注的 Twitter 图文数据集的稀缺不利于模型的训练，本文采用现有较成熟的图像情感数据集和文本情感数据集来对模型进行预训练。实验数据集的详细信息如表1所列。

表1 实验数据集的详细信息

Table 1 Details of experimental data set

Type	Name	Positive	Negative
Image dataset	Twitter1269	769	500
	Twitter10699	8443	2256
Text dataset	TwitterText	58659	73221
Multimodaldataset	Twitter603	470	133
	Twtter2014	1008	1006

本文使用到的图像情感训练数据集有两个：Twitter1269 数据集以及 Twitter10699 情感图像数据集。前者是文献[35]中构建的图像情感数据集，一共包含 1269 张 Twitter 的图片，其中有 769 张图片为情绪积极的图片，另外 500 张为情绪消极的图片。后者是数据标注公司 Figure Eight[2] 提供的图片情感分析数据集，一共包含了 10699 张 Twitter 图片，其中有 8443 张情绪积极的图片和 2256 张情绪消极的图片。

对于文本情感分析模型的训练，本文采用的是 Analytics Vidhya 机构提供的 Twitter 文本情感分析数据集 TwitterText[2]，其中包含 131880 条文本数据，包括 73221 条情绪消极的文本与 58659 条情绪积极的文本。

本文所采用的多模态数据集是 Twitter603 与 Twtter2014。前者是文献[21]所构造的实验数据集，一共包含 603 条图文数据对，其中有 470 条情绪积极的图文数据对与 133 条情绪消极的图文数据对。而后者则是，本文为了实验而在 Twitter 平台上收集的多模态数据集，总共收集了 2250 条图文数据对，并由 3 个研究员对数据进行数据情感标注。将 3 个研究员判定类别一致的数据保留，将有争议的数据剔除，最终形成了包含 1008 条情绪积极的图文数据对与 1006 条情绪消极的图文数据对的多模态数据集。

4.2 模型的预训练

由于多模态数据集的稀缺，模型的训练变得困难。为解决这个问题，采用图像情感分类数据集与文本情感分类数据集分别对模型的图像特征提取网络和文本特征提取网络进行预训练。本文在图像特征提取网络与文本特征提取网络最后

1) https://www.figure-eight.com/data-for-everyone/

2) https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/

3) https://download.pytorch.org/models/vgg16-397923af.pth

添加一个三层全连接神经网络分类器，将其构造成为图像情感分类模型和文本情感分类模型。

在图像情感分类模型的训练中，使用在 ImageNet 数据集上训练好的图像分类预训练模型[3]对图像情感分类模型的卷积模块进行参数初始化，对注意力计算模块部分的参数采用正态分布随机小数进行初始化。在图像情感数据集 Twitter10691 及 Twitter1269 数据集上对模型进行微调(fine-tuning)，得到本文的图像情感特征提取网络的预训练模型。同样地，在对文本情感分类模型的训练中，采用正态分布随机小数对模型的参数进行初始化。将模型在 TwitterText 数据集上进行训练，从而得到文本情感特征提取网络的预训练模型。

4.3 实验设置

对于图像情感特征提取网络，我们将对图像进行缩放裁剪后大小为 224×224 的三通道 RGB 彩色图像作为输入。对于文本内容，我们首先对文本数据进行了预处理。删除了所使用到的数据集的文本中没有实际语义的用户名、数字及特殊字符，然后进行分词处理。通过在大规模的 Twitter 文本数据集上训练得到的预训练模型 GloVe，获取维度为 100 的词向量作为文本特征提取网络的输入。通过 GloVe 在大规模的 Twitter 文本数据集上训练得到的预训练模型，获得文本单词的维度为 100 词向量表示作为文本特征提取网络的输入。文本特征提取网络的最大输入长度限制为 50 个词，不足 50 词的数据将会用 0 补足，超过的则截断。

在训练过程中，我们使用 Dropout 正则化来提高模型在数据集上的泛化能力。采用交叉熵函数作为损失函数，采用小批量梯度下降法来优化参数，其动量设置为 0.9。文本情感分类模型的初始学习率为 0.01。图像情感分类模型的注意力权重计算模块参数的学习率设置为 0.01。学习率按指数衰减调整学习率，学习率调整倍数的底设置为 0.95。本文所有的模型训练是在 GPU(NVIDIAGeforceGTX 1080)上进行的。

4.4 实验结果

4.4.1 比较模型与评价指标

为验证本文所提模型的有效性，将本文模型与单模态情感分类模型和基于神经网络的多模态情感分类模型进行对比。

(1)单模态情感分类模型

VGG16 是经典的图像分类模型，本文对其进行微调用于图像情感分类。

CNNa 是基于本文的图像特征提取网络的图像情感分类模型。

BiGRU-Text 模型是采用双向 GRU 网络结构的文本情感分类模型。

ABiGRU-Text 模型是一个基于注意力机制的 BiGRU-Text 模型。其注意力权重的计算与本文所提文本情感特征提取网络一致。

(2)多模态情感分类模型

NNM 模型，是采用张量融合方案的基于神经网络的多模态情感分析模型，是 ANNM 未采用注意力机制的简化模型。

本文实验是一个二分类问题，为评估模型的性能，本文选取了精确度(Precision)、召回率(Recall)、F1score 及准确率(Accuracy)作为实验的评价指标。其计算式如下：

$$\begin{cases} Precision=\dfrac{TP}{TP+FP} \\ Recall=\dfrac{TP}{TP+FN} \\ F1\ score=\dfrac{2*Precision*Recall}{Precision+Recall} \\ Accuracy=\dfrac{TP+TN}{TP+FN+TN+FP} \end{cases} \tag{18}$$

其中，TP 是正确地被标记为正例的样本数，FP 是被错误地标记为正例但实际上是反例的样本数，TN 是被正确地标记为反例的样本数，FN 是被错误地标记为反例但实际上是正例的样本数。

4.4.2 结果分析

表 2 和表 3 列出了不同模型在两个 Twitter 图文数据集上的实验结果。总体来说，本文提出的模型在两个实验数据集上的表现都要优于其他模型。由表 2 中的结果可以看出，多模态情感分类模型的效果与单模态情感分类模型相比，多模态情感分类模型在各个评价指标上都有更优越的表现。NNM 多模态情感分类模型与文本情感分类模型 BiGRU-Text 模型、图像情感分类模型 VGG16 模型相比，在准确率上有了 4.5%和 16.9%的提升。ANNM 多模态情感分类模型与文本情感分类模型 ABiGRU -Text 模型、图像情感分类模型 CNNa 模型相比，在准确率上有了 4.3%和 15.8%的提升。这充分说明了多模态数据更能揭示用户的真实情感。也证实了用张量融合的方法结合各模态特征向量，能够有效地利用各模态间的互补信息，从而提升情感分类的效果。

基于注意力机制的图像情感分类模型 CNNa 模型相比 VGG16 模型在准确率上提高了 8.69%，基于注意力机制的文本情感分类模型 ABiGRU-Text 模型相比 BiGRU-Text 模型在准确率上提高 7.75%，这证明了注意力机制能更有效地提取图像和文本的情感信息，从而提高了分类效果。特别地，ANNM 模型相比 NNM 模型在各个指标上都有大于 7%的提升。采用基于注意力神经网络的模型对分类结果有较大的提升，这说明了本文的 ANNM 模型的可行的。

表 2 在 Twitter2014 数据集上的实验结果

Table 2 Experimental results on Twitter2014 dataset

Model	*Precision*	*Recall*	*F1 score*	*Accuracy*
VGG16	0.6216	0.5833	0.6018	0.6137
BiGRU -Text	0.7505	0.7133	0.7314	0.7378
NNM	**0.8107**	**0.7391**	**0.7732**	**0.7830**
CNNa	0.7325	0.6329	0.6791	0.7006
ABiGRU -Text	0.8313	0.7917	0.8110	0.8153
ANNM	**0.8817**	**0.8284**	**0.8542**	**0.8585**

表 3 在 Twitter603 数据集上的实验结果

Table 3 Experimental results on Twitter603 dataset

Model	*Precision*	*Recall*	*F1 score*	*Accuracy*
VGG16	0.7762	0.7085	0.7408	0.6136
BiGRU -Text	0.8043	0.7872	0.7957	0.6849
NNM	**0.8228**	**0.8298**	**0.8263**	**0.7280**
CNNa	0.8084	0.7809	0.7944	0.6849
ABiGRU -Text	0.8463	0.8319	0.8391	0.7512
ANNM	**0.8565**	**0.8766**	**0.8665**	**0.7894**

由表 3 的结果可以看出，本文所提出的 ANNM 模型仍然是在众多模型中效果最好的。另外，在表 3 中所有模型的精确率、召回率、F1 score 的结果数据相比准确率都更好。而准确率与其他评估指数相比更低。这种情况说明，模型对正

例即积极的数据的识别能力很强而对反例即消极的数据识别能力较弱。这可能是由于 Twitter603 数据集中数据正负样本比例较大造成的。但即使存在数据不平衡的问题，本文所提出的模型仍然有很好的表现，这也说明了 ANNM 模型是有效的。

结束语 本文提出了一种基于注意力机制的多模态情感分析方法。结果表明，本文提出的模型产生了更好的分类效果。由于现有的数据资源与水平的限制，本文工作还有进一步改善的空间。我们提出的模型是仅考虑完善的社交媒体的图文多模态数据的情况，然而在现实生活中，社交媒体数据是文本、图像、图文数据对共存的多模态数据集。在未来，我们会补充我们的数据资源，并且提出多模态数据存在模态缺失的情况下更有效的模型。

参 考 文 献

[1] ASUR S, HUBERMAN B A. Predicting the future with social media [C] // Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Toronto, Canada, 2010: 492-499.

[2] O'CONNOR B, BALASUBRAMANYAN R, ROUTLEDGE B R, et al. From tweets to polls: Linking text sentiment to public opinion time series [C] // Proceedings of the International AAAI Conference on Weblogs And Social Media. Washington, United States, 2010: 11: 122-129.

[3] TUMASJAN A, SPRENGER T O, SANDNER P G, et al. Predicting elections with twitter: What 140 characters reveal about political sentiment [C] // Proceedings of the International AAAI Conference on Weblogs And Social Media. Washington, USA, 2010: 10: 122-129.

[4] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module [C] // Proceedings of the European Conference on Computer Vision 2018. ECCV, Munich, Germany, 2018: 3-19.

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

[6] NGUYEN T H, SHIRAI K. Topic modeling based sentiment analysis on social media for stock market prediction [C] // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China, 2015: 1354-1364.

[7] ZHANG L, ZHAO Y, ZHU Z F. Advances in Semantically Shared Subspace Learning for Cross－Media Data [J]. Chinese Journal of Computers, 2017.

[8] ZHANG L, WANG S, LIU B. Deep learning for sentiment analysis: A survey [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018: e1253.

[9] TURNEY P D. Thumbs up or thumbs down: semantic orientation applied to unsupervised classification of reviews [J]. Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002: 417-424.

[10] TABOADAM, BROOKE J, TOFILOSKI M, et al. Lexicon-Based Methods for Sentiment Analysis [J]. Computational Linguistics, 2011, 37(2): 267-307.

[11] BACCIANELLA S, ESULI A, SEBASTIANI F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining [C] // Proceedings of the International Conference on Language Resources and Evaluation. Valletta, Malta, European. 2010, 2010(10): 2200-2204.

[12] HU M, LIU B. Mining and summarizing customer reviews [C] // Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004: 168-177.

[13] PANG B, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques [C] // Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. 2002.

[14] KIM Y. Convolutional neural networks for sentence classification [J]. arXiv, 2014: 1408.5882.

[15] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional Neural Network for Modelling Sentences [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2014: 655-665.

[16] TAY Y, TUAN L A, HUI S C. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis [C] // Thirty-Second AAAI Conference on Artificial Intelligence, Louisiana, USA, 2018.

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

[18] SONG J, YU Q, SONG Y Z, et al. Deep spatial-semantic attention for fine-grained sketch-based image retrieval [C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017). Venice, Italy, 2017: 5551-5560.

[19] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv: 1409.1556, 2014.

[20] 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. Irenze, Italy, 2010: 715-718.

[21] 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. Barcelona, Spain, 2013: 223-232.

[22] XU C, CETINTAS S, LEE K C, et al. Visual sentiment prediction with deep convolutional neural networks [J]. arXiv: 1411.5731.

[23] YOU Q, JIN H, LUO J. Visual sentiment analysis by attending on local image regions [C] // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA, AAAI Press, 2017: 231-237.

[24] YANG Y, JIA J, ZHANG S, et al. How do your friends on social media disclose your emotions? [C] // Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Québec City, Québec, Canada, AAAI Press, 2014: 306-312.

[25] YUAN J, MCDONOUGH S, YOU Q, et al. Sentribute: image sentiment analysis from a mid-level perspective [C] // Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2013). Chicago, IL, USA, ACM, 2013: 10: 1-10: 8.

(下转第 548 页)