山西大学学报(自然科学版)47(1):48-58,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023134

面向目标多模态情感分析的双通道循环神经网络模型

王静红[1,2,3],高远[1],李昊康[4*]

(1.河北师范大学 计算机与网络空间安全学院,河北 石家庄 050024;
2.河北师范大学 河北省网络与信息安全重点实验室,河北 石家庄 050024;
3.河北师范大学 供应链大数据分析与数据安全河北省工程研究中心,河北 石家庄 050024;
4.河北工程技术学院 人工智能与大数据学院,河北 石家庄 050020)

摘 要:面向目标的多模态情感分析,其任务是对多模态帖子或评论中给定的目标词进行情感分类。针对目前该领域结合循环神经网络的模型只关注于一般的文本和图片表示,没有同时考虑模态内和模态间的信息交互,且忽略了图像信息中的噪声的问题,提出了一种双通道循环神经网络模型(DRNN)。该模型首先设计了一个基于注意力机制的循环神经网络模块,该模块利用门控循环单元(Gate Recurrent Unit, GRU)来过滤图像的噪声,之后通过注意力机制将文本和图像融合,最后将融合后的信息逐步加入目标信息中,得到模态间的动态表示。另外提出了一个目标文本交互循环神经网络模块,该模块通过计算目标信息与上下文中每个词的权重来学习模态内的上下文表示。最后将两部分模块得到的信息拼接后送入全连接层和softmax层预测情感极性。在两个基准数据集Twitter-15和Twitter-17上进行了大量实验,实验结果表明,与当前最先进的模型相比该模型能够有效增强面向目标的多模态情感分析的效果。

关键词:循环神经网络;多模态;面向目标的情感分析;注意力机制;噪声
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)01-0048-11

Dual-channel Recurrent Neural Network Model for Target-oriented Multimodal Sentiment Analysis

WANG Jinghong[1,2,3], GAO Yuan[1], LI Haokang[4*]

(1. *College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang* 050024*, China;*
2. *Hebei Key Laboratory of Network and Information Security, Hebei Normal University, Shijiazhuang* 050024*, China;*
3. *Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics & Security, Hebei Normal University, Shijiazhuang* 050024*, China;*
4. *Academy of Artificial Intelligence and Big Data, Hebei Polytechnic Institute, Shijiazhuang* 050020*, China*)

Abstract: The task of the target-oriented multimodal sentiment analysis is to classify sentiment for a given target word in a multimodal post or comment. Aiming at the problems that current models incorporating recurrent neural networks in this field only focus on general text and image representations, but never take intra-modal and inter-modal information interactions into account, and ignore noise in image information, in this paper, we propose a dual-channel recurrent neural network model (DRNN). The model de-

收稿日期:2023-06-08;**接受日期**:2023-08-04
基金项目:河北省自然科学基金(F2021205014;F2019205303);河北省高等学校科学技术研究项目(ZD2022139);中央引导地方科技发展资金项目(226Z1808G);河北省归国人才资助项目(C20200340)
作者简介:王静红(1967-),女,河北石家庄人,博士,教授,研究方向为人工智能、模式识别、自然语言处理、机器学习与数据挖掘。E-mail:wangjinghong@126.com
* **通信作者**:李昊康(LI Haokang),E-mail:lihaokang217@126.com
引文格式:王静红,高远,李昊康.面向目标多模态情感分析的双通道循环神经网络模型[J].山西大学学报(自然科学版),2024,47(1):48-58. DOI:10.13451/j.sxu.ns.2023134

signs a recurrent neural network module based on the attention mechanism, which first uses gate recurrent unit (GRU) to filter the noise of the image, then fuses the text and image through the attention mechanism, and finally adds the fused information to the target information step by step to obtain the dynamic representation between the modalities. In addition, we propose an recurrent neural network module for target-text interaction that learns the contextual representation within a modality by computing target information with the weight of each word in the context. Finally, we stitch together the information obtained from the two modules and send it to the fully connected and softmax layers to predict the sentiment polarity. Extensive experiments are conducted on two benchmark datasets, Twitter-15 and Twitter-17, which showed that the model is effective in enhancing target-oriented multimodal sentiment classification compared to current state-of-the-art models.

Key words: recurrent neural network; multimodal; target-oriented sentiment analysis; attention mechanism; noise

0 引言

随着互联网的普及,越来越多的用户习惯于在网上发表意见,由此会产生大量来自各个领域的包含情感的数据。在这个背景下,对这些数据进行情感分析就显得尤为重要。面向目标情感分析是情感分析中的一项关键任务,旨在确定特定目标的情感极性(积极、消极或中性)。例如,给定一条推文“我对这幅图片的背景很感兴趣,但是颜色我不是很喜欢”,用户分别对图片的背景和颜色表达了积极和消极的情绪。随着深度学习的发展,针对该任务提出了许多神经网络架构,例如卷积神经网络(Convolutional Neural Network,CNN)[1]和循环神经网络(Recurrent Neural Network,RNN)[2]。最近,为了更好地捕捉上下文词和目标实体之间的语义交互,许多研究试图将注意力机制应用于RNN,并且已被证明在大多数基准数据集中取得了较为先进的结果[3]。

随着大量多模态数据的出现,许多学者尝试在文本之上融合图像信息,并取得了较好的成果。例如Xu等[4]设计了一种多交互记忆网络模型,该模型利用两个交互式记忆网络学习目标相关的文本信息与目标相关的视觉信息。Yu等[5]提出实体敏感注意力融合网络,利用注意力机制学习目标与文本和视觉间的信息,Zhang等[6]利用记忆网络挖掘两种模态内的信息,之后通过判别矩阵监督两种模态信息的融合。以上方法均通过捕获目标-文本和目标-图像的信息,或者捕获文本-图像间的信息来进行建模,且对于图像信息中的噪声过滤不充分。而这些信息会影响情感的表达,例如图1中的两个数据样本,在第一个数据样本中,若不考虑文本和图像之间的信息关联将无法确定文本中的“russwest44”指的是谁。对于不相关数据样本也可能会存在与目标无关的噪声,例如在第二个数据样本中图像存在多个目标,而那些无关的目标可能会将目标词“Lily”的情感预测为中性。

图1 数据集中的部分数据

Fig. 1 Selected data in the dataset

因此本文同时对上述两种方式进行建模,提出了一种双通道循环神经网络模型(Dual-channel Recurrent Neural Network,DRNN)。具体来说,首先通过BERT(Bidirectional Encoder Representation from Transformers)和BUA(Bottom-Up Attention)模型来提取文本和图像信息;之后构建了一个目标文本交互循环神经网络(Target-Text Interaction Recurrent Neural Network,TTIRNN)模型,该模型可以充分提取目标词与上下文之间的敏感表示;构建了一个基于注意力机制的循环神经网络(Attentional Recurrent Neural Network,ARNN)模型,该模型可以过滤图像噪声并逐步学习模态间的信息。最后将上述两个模块中提取到的信息拼接后送入全连接层进行情感分类。本文的主要贡献包括3个方面:

1)构建了DRNN模型,DRNN模型由TTIRNN和ARNN组成,TTIRNN可以捕获目标词与上下文词之间的语义相关性;ARNN可以过滤图像噪声并学习模态间的动态表示;

2)模型同时考虑了模态内和模态间的信息交互,且过滤了图像中的噪声;

3)大量的实验证明所提模型的有效性。在两个基准数据集上的实验表明,与一些强基线模型相比,DRNN在面向目标的多模态情感分析任务中具有一定的竞争性。

1 相关工作

1.1 面向目标的情感分析

面向目标的情感分析可以分为两大类:基于传统方法和基于深度学习方法。

基于传统方法主要使用机器学习方法进行情感分析。Sharma等[7]使用支持向量机(Support Vector Machines,SVM)作为基础分类器来利用Boosting的分类性能,实验结果表明所提出的方法显著提高了SVM的性能。Kamal等[8]提出了一种基于规则和机器学习的方法来识别目标词的情感极性。Yu等[9]提出了一种用于情感分类的半监督学习方法,取得了较好的性能。尽管基于传统学习的方法取得了较好的效果,但这种方法困难在于分类器的组合选择以及需要大量人工标注的数据集。

近年来,随着深度学习在自然语言处理中的应用越来越广泛。Dong等[10]将循环神经网络应用于面向目标的情感分析,并根据上下文和句法关系获得目标词的情感极性,模型的性能取得了较好的结果。Tang等[2]建立了两个目标相关的长短期记忆网络(Long Short-term Memory,LSTM)模型建立目标词和上下文之间的联系。He等[11]利用基于注意力的LSTM,在文档级数据上进行训练,以捕捉特定领域的情感词汇。由于注意力机制在建立目标词与上下文之间的关系具有很好的效果,Wang等[12]利用注意机制获得了方面词在词和句子层面上与情感高度相关的信息。随着记忆网络在面向目标的情感分析中具有有效性,Tang等[13]将注意力机制与记忆网络相结合,并取得了很好的效果。近年来,CNN在各个领域中获得了广泛的普及和应用。Fan等[14]将注意力机制与CNN结合,同时捕获单词之间的表达。Xue等[1]将门控机制应用于CNN,使模型的计算更加简洁高效。随着预训练语言模型的发展和广泛使用,Phan等[15]使用BERT进行语境嵌入,并结合语篇信息和句法信息进行方面级情感分析。最近研究人员发现,使用外部知识可以有效提升模型的性能,Cambria等[16]构建了情感分析知识库SenticNet,Gu等[17]将外部知识融入图卷积网络,使用外部情感词典对句子中每个词进行情感打分,取得了出色的实验结果。

1.2 面向目标的多模态情感分析

面向目标的多模态情感分析其目的是识别多模态数据对目标的情感极性。Xu等[4]提出了多模态方面级情感分析任务,构建了一个多交互记忆网络模型,用给定的方面监督文本和视觉信息。Yu等[18]采用额外的BERT层,从图像区域中提取目标相关的特征向量。之后,Zhang等[6]利用两个记忆网络来挖掘文本和图像的模态内信息,然后设计一个判别矩阵来融合这两种信息。随着注意力机制在多模态融合中的应用,Gu等[19]设计了一个注意力胶囊提取和多头融合网络,通过多头注意力机制(Multi-Head Attention,MHA)和胶囊网络的整合捕捉多模态输入之间的互动。Ju等[20]以端到端的方式处理多模态方面提取和多模态方面级情感分析这两个子任务。Yu等[21]提出了具有交叉模态对齐的双编码器转换器,并引入了两个辅助任务来增加注意力性能。端到端的面向目标的情感分析可以提高模型性能,Yang等[22]提出了一个多任务学习框架引入了两个辅助任务来学习模态内表示。受预训练语言模型的启发,Ling等[23]构建了一个统一多模态编码器-解码器架构,并设计了三类预训练任务。与之前的工作不同,本文提出了双通道循环神经网络模型DRNN,使用两个循环神经网络分别从多个层次充分融合目标文本信息、目标视觉信息和图文信息。

2 方法

在本节中,我们主要介绍任务定义,并详细阐述DRNN的各个部分。

2.1 任务定义

给定一个长度为n的句子$T=\{w_1,w_2,\cdots,w_n\}$,与文本相关的图片I,其中文本中包括长度为m的目标短语$A=\{a_1,a_2,\cdots,a_m\}$。面向目标的多模态情感分析的任务是预测目标短语A的情感极性Y,其中Y可以是积极的、消极的、中性的。

2.2 模型概述

RNN有两种重要变体,长短期记忆网络和门控循环单元(Gated Recurrent Unit,GRU)。两种模型均是在RNN的基础上引入门控机制,有选择地遗忘和更新存储信息,从而有效处理序列数据。由于LSTM和GRU具有各自的优点,所以本文针对不同的任务在两种循环神经网络的基础上建模。在目标词与文本的交互中,由于文本的序列较长,而LSTM在复杂序列的任务中建模效果较好,所以本文将LSTM应用于TTIRNN。在目标词与文本-图像交互过程中会产生大量的参数,而GRU建模所需的参数量较少,所以本文将GRU应用于ARNN。针对不同的任务使用不同的模型进行建模可以有效提高模型性能,同时减少不必要的时间开销。

本文提出的双通道循环神经网络主要由四部分组成:特征编码层、目标文本交互循环神经网络、基于注意力机制的循环神经网络、输出层。模型总体结构如图2所示。

特征编码层:使用BERT作为文本编码器,预训练的BUA模型作为图像编码器。分别将文本和图像的嵌入表示作为输入,输出文本和图像的隐藏表示。目标文本交互循环神经网络:使用两个LSTM分别对目标词和上下文词进行建模。首先通过目标词为上下文词生成适当的注意力权重,然后将注意力权重与上下文词加权求和得到上下文表示,最后将上下文表示与目标表示经过文本融合层融合后获得最终的文本表示。基于注意力机制的循环神经网络:通过注意机制将文本与图像的特征在分词级别融合,之后将目标信息与融合后的信息通过GRU更新。本文提出的双通道循环神经网络模型图如图3所示。

2.3 文本编码层

将句子S_N拆分成三部分$S_N=(S_L,S_R,S_A)$。采用预训练的词嵌入模型BERT分别将S_L,S_R,S_A和S_N表示为对应的词嵌入$E^l=(e_1^l,e_2^l,\cdots,e_L^l)$,$E^r=(e_1^r,e_2^r,\cdots,e_R^r)$,$E^a=(e_1^a,e_2^a,\cdots,e_A^a)$和$E^w=(e_1^w,e_2^w,\cdots,e_W^w)$其中$E^l\in R^d$表示目标词的上文嵌入表示,$E^r\in R^d$表示目标词的下文嵌入表示,$E^a\in R^d$表示目标词嵌入表示,$E^w\in R^d$表示输入文本整体嵌入表示,$d$为隐藏特征的向量维度。$N=L+R+A$,其中$L$、$R$和$A$是每部分的输入长度。

2.4 图像编码层

为了充分利用图片信息,本文使用预训练的自下而上的注意力模型(bottom-up attention,BUA)作为图像编码器。对于给定的图像I,BUA模型可以检测图像I中的物体和其他区域,并输出相应的特征,然后通过线性层将特征向量投影到d维空间中。编码图像为$E^I=\{e_1^I,e_2^I,\cdots,e_k^I\}$,其中$k$表示图像中对象或区域的数量。

2.5 目标词表示

采用BERT得到目标词的嵌入向量表示后,通过标准的LSTM得到目标词的上下文表

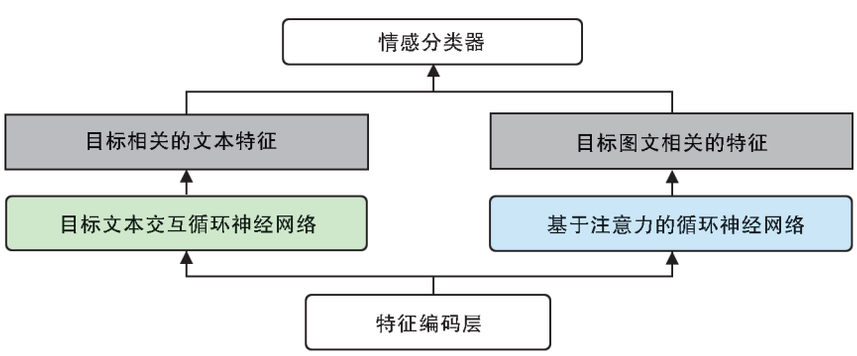

图2 DRNN模型总体结构

Fig. 2 General structure of the DRNN model

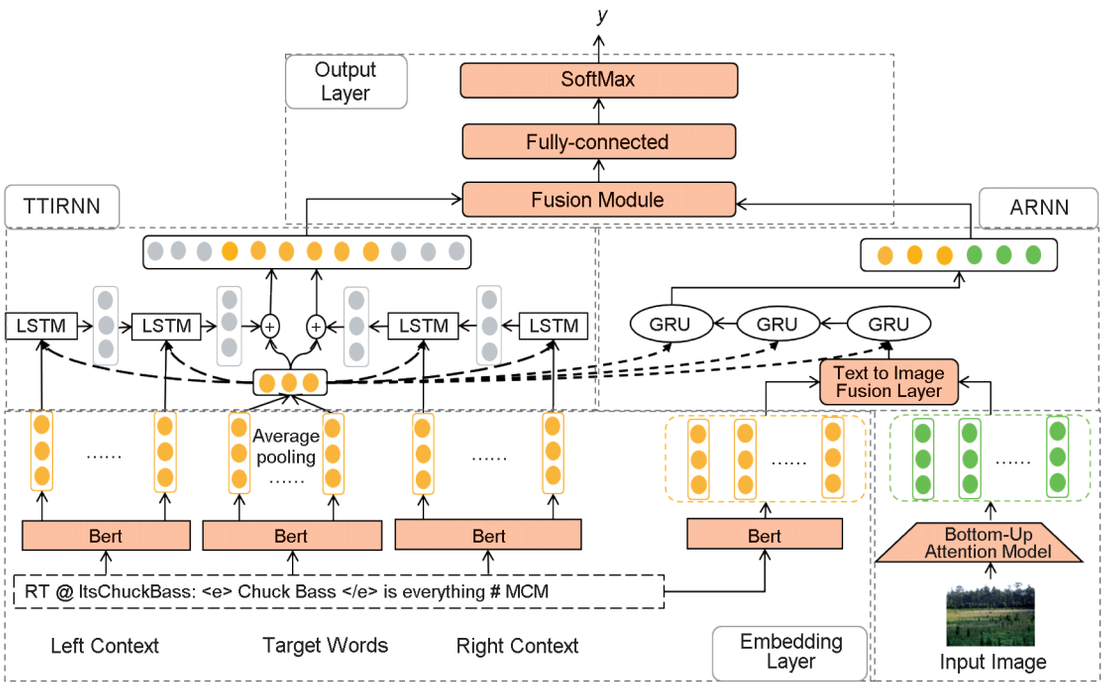

图3 DRNN示意图

Fig. 3 Illustration of DRNN

示,最后通过平均池化得到最终的目标词表示。

$$h_i^a = LSTM_\Theta(h_{i-1}^a, e_i^a), i \in [1, A], \quad (1)$$

其中,Θ 和 $h_i^a \in R^d$ 表示LSTM中的所有参数。在获得目标词的隐藏状态$[h_1^a, h_2^a, \cdots, h_A^a]$之后,采用平均池化操作得到目标词的最终表示$h_0$。

$$h_0 = \sum_{i=1}^{A} \frac{1}{A} h_i^a。 \quad (2)$$

2.6 目标文本交互循环神经网络

为了更好地捕捉上下文的语义含义,模型同样通过两个独立的LSTM网络将生成上下文的 d 维隐藏状态 $\widehat{x^l} = [x_1^l, x_2^l, \cdots, x_L^l]$, $\widehat{x^r} = [x_1^r, x_2^r, \cdots, x_R^r]$。基于这些隐藏状态,使用注意力机制学习上下文的语义表示。以目标表示作为输入h_0,计算上下文的隐藏状态的注意力权重a_i^l和a_i^r。

$$v_i^l = \tanh(e_i^l \cdot W_H^l \cdot h_0 + b^l), \quad (3)$$

$$v_i^r = \tanh(e_i^r \cdot W_H^r \cdot h_0 + b^r), \quad (4)$$

$$a_i^l = \frac{\exp(v_i^l)}{\sum_{k=1}^{L} \exp(v_k^l)}, \quad (5)$$

$$a_i^r = \frac{\exp(v_i^r)}{\sum_{k=1}^{R} \exp(v_k^r)}, \quad (6)$$

其中 W_H^l, W_H^r, b^l, b^r 是可学习参数。最终上下文的表示为$\widehat{x^{lr}}$:

$$\widehat{x^l} = \sum_{i=1}^{L} a_i^l x_i^l, \quad (7)$$

$$\widehat{x^r} = \sum_{i=1}^{R} a_i^r x_i^r。 \quad (8)$$

最后,使用特征拼接来融合目标和上下文的信息。

$$\widehat{x^{lr}} = [\widehat{x^l}; \widehat{x^r}], \quad (9)$$

其中,$\widehat{x^l}$和$\widehat{x^r}$表示目标相关的上下文,“;”表示拼接操作。算法过程如表1所示。

表1 TTIRNN算法过程

Table 1 TTIRNN algorithmic process

Algorithm1: 目标文本交互循环神经网络算法过程
Input: 上下文词嵌入E^l和E^r,目标词嵌入h_0,特征矩阵维度d
Output: 目标相关的上下文表示$\widehat{x^{lr}}$,
1 while $i = 1, 2, \cdots, d$ do
2 随机生成权重矩阵$\boldsymbol{W}_H^l, \boldsymbol{W}_H^r$, 偏置参数$b^l, b^r$
3 根据式(3),(5)计算上文注意力权重a_i^l
4 根据式(4),(6)计算下文注意力权重a_i^r
5 end while
6 根据式(7),(8)加权求和计算上下文表示$\widehat{x^l}, \widehat{x^r}$
7 $\widehat{x^{lr}} \leftarrow \widehat{x^l}$特征拼接$\widehat{x^r}$
8 return $\widehat{x^{lr}}$

2.7 基于注意力机制的循环神经网络

为了更好地结合文本表示与视觉表示,模型将不同模态的特征在分词级别融合。使用编码文本E^w中的每个标记向量e_i^w来查询图像特征E^I以获得相关的图像信息v_i^{img}。

$$v_i^{\text{img}}=f_{\text{attention}}(Q_i^w,K^I,V^I)=\text{soft max}\left(\frac{e_i^w\cdot W^Q(E^I\cdot W^K)^T}{\sqrt{d}}\right)E^I\cdot W^V, \tag{10}$$

其中,W^Q,W^K,W^V是可学习参数。对于分词w_i,将其词向量e_i^w和图像信息v_i^{img}作为GRU的输入。更新后的隐藏状态为e_i^{WI}:

$$e_i^{WI}=f_{\text{GRUCell}}(v_i^{\text{img}},e_i^w)。 \tag{11}$$

多模态信息的融合结果为$E^{WI}=\{e_1^{WI},e_2^{WI},\cdots,e_n^{WI}\}$。之后在第一个时间步将目标词表示$h_0$与$E^{WI}$送入GRU,更新目标表示得到$h_1$。在第$t$个时间步,多模态信息融合表示$E^{WI}$与前一时间步产生的目标表示$h_{t-1}$送入GRU。通过GRU根据获取的多模态信息不断更新目标表示,在最后一个时间步得到的融合图文信息的目标表示为:

$$\widehat{x^a}=h_t=f_{\text{GRUCell}}(E^{WI},h_{t-1}),t\geqslant 1。 \tag{12}$$

算法过程如表2所示。

表2 ARNN算法过程

Table 2 ARNN algorithmic process

Algorithm2: 基于注意力机制的循环神经网络算法过程
Input:目标词嵌入h_0,文本嵌入$E^W=(e_1^w,e_2^w,\cdots,e_W^w)$,图片嵌入$E^I$,时间步$N$。
Output:目标敏感的图文融合信息$\widehat{x^a}$。
while $i=1,2,\cdots,W$ do
随机生成权重矩阵W^Q,W^K,W^V
$Q^W\leftarrow e_i^wW^Q$
$K^I\leftarrow E^IW^K$
$V^I\leftarrow E^IW^V$
$v_i^{\text{img}}\leftarrow f_{\text{attention}}(Q^W,K^I,V^I)$
$i\leftarrow i+1$
$e_i^{\text{WI}}\leftarrow f_{\text{GRUCell}}(V_i^{\text{img}},E^W)$
end while
得到文本图像融合表示$E^{\text{WI}}=(e_1^{\text{WI}},e_2^{\text{WI}},\cdots,e_W^{\text{WI}})$
while $i=1,2,\cdots,N$ do
$h_t\leftarrow f_{\text{GRUCell}}(E^{\text{WI}},h_{t-1})$
$t\leftarrow t+1$
end while
$\widehat{x^a}\leftarrow h_t$
return $\widehat{x^a}$

2.8 输出层

为了丰富图文模态融合表示的情感语义信息,在对所有时间步进行更新后,将最后一个时间步产生的多模态目标表示$\widehat{x^a}$与目标相关的文本表示$\widehat{x^{lr}}$进行拼接:

$$\widehat{x^{alr}}=[\widehat{x^a};\widehat{x^{lr}}]。 \tag{13}$$

将融合的特征送入到全连接网络,然后利用softmax函数进行归一化处理,预测情感极性的概率分布。

$$p(y|\widehat{x^{alr}})=\text{soft max}(W_s\cdot\widehat{x^{alr}}+b), \tag{14}$$

其中,W_s和b是可学习参数。

3 实验

本节介绍了用于评估的数据集、参数设置和用于比较的基线方法,并且研究了不同参数和模块对模型的影响。

3.1 数据集及评估指标

实验采用Twitter-15和Twitter-17[5]两个基准数据集,数据集分别来自2014—2015年和2016—2017年用户在Twitter上发布的帖子,将两个数据集拆分为训练集、验证集、测试集。数据集中每个样本包含一段文本、一张图像、一个目标词以及相应的情感极性标签,情感极性标签为:积极的、消极的、中性的。数据集的基本统计数据如表3所示。

表3 数据集样本统计数据

Table 3 Sample statistics for the dataset

样本	Twitter-15			Twitter-17		
	训练集	验证集	测试集	训练集	验证集	测试集
句子总数	2 100	727	674	1 745	577	587
目标总数	3 179	1 122	1 037	3 562	1 176	1 234
积极的	928	303	317	1 508	515	493
中性的	1 883	670	607	1 638	517	573
消极的	368	149	113	416	144	168
平均长度	15	16	16	15	16	15
最大长度	35	40	37	39	31	38

采用Accuracy(Acc.)和Macro-$F1$($F1$)作为模型的最终评估指标[17]。具体公式如下:

$$A_{\text{acc}}=\frac{T}{N}\cdot 100\%, \tag{15}$$

$$P_{\text{precision}}=\frac{TP}{TP+FP}, \tag{16}$$

$$R_{\text{recall}}=\frac{TP}{TP+FN}, \tag{17}$$

$$F_{F1\text{-score}_i}=2\frac{R_{\text{recall}_i}\cdot P_{\text{precision}_i}}{R_{\text{recall}_i}+P_{\text{precision}_i}}\cdot 100\% \quad , \tag{18}$$

$$M_{\text{macro-}F1}=\frac{F_{F1\text{-score}_1}+F_{F1\text{-score}_2}+F_{F1\text{-score}_3}}{3}, \tag{19}$$

其中,T表示正确预测样本数量,N表示样本总数,TP为真正例,FP为假正例,FN为假负例。

3.2 实验环境及参数

实验代码基于Pytorch框架实现。将两个数据集中的文本输入最大序列长度分别设置为40和39,优化模型采用Adam优化器。BERT文本嵌入维度为768,注意头数设置为12;使用预训练的BUA模型,将最小特征数分别设置为3和36。时间步数的最大值被设置为4,批次大小设置为32,学习率为4×10^{-5}。使用不同的随机种子运行3次,取平均值作为最终结果。

3.3 基线模型

为了评估所提出的DRNN模型的性能,将其与几个基线模型进行比较,包括基于文本的方法、基于视觉的方法和基于多模态的方法,实验结果如表4所示。

RES-Target:该方法将ResNet[24]用于图像编码,BERT[25]用于目标词编码,并将图像和目标词编码表示拼接后作为情感极性分类的特征。

MGAN[26]:通过使用粗粒度和细粒度的注意力机制来学习目标词和上下文之间的交互。

BERT:该方法采用多层Transformers[27]编码器来生成动态词向量表示。

BERT+BL:在BERT-Base模型之上再添加一个额外的BERT层。

Res-MGAN-TFN:通过使用张量融合网络(TFN)[28]融合ResNet和MGAN的编码结果。

MIMN[4]:使用两个交互式记忆网络学习方面词和文本以及图像之间的信息。

ESAFN[5]:利用注意力机制生成实体敏感的文本表示和视觉表示,然后使用门控机制消除噪声。

TomBERT[18]:该模型首先应用BERT获得目标敏感的文本表示,然后设计了目标注意机制进行目标图像匹配,从而获得目标敏感的视觉表示。

表4 Twitter-15和Twitter-17数据集上模型的对比实验结果

Table 4 Comparative experimental results of models on Twitter-15 and Twitter-17 datasets

方法	模型	Twitter-15		Twitter-17	
		Acc.	$F1$	Acc.	$F1$
视觉	Res-Target	59.88	46.48	58.59	53.98
文本	MGAN	71.17	64.21	64.75	64.16
	BERT	74.15	68.86	68.15	65.23
	BERT+BL	74.25	70.04	68.88	66.12
文本+视觉	Res-MGAN-TFN	70.30	64.14	64.10	59.13
	MIMN	73.53	66.49	67.22	63.85
	ESAFN	73.49	67.31	67.83	64.32
	TomBERT	77.15	71.75	70.34	68.03
	DRNN	**78.20**	**74.08**	**70.96**	**69.35**

3.4 对比实验

在两个数据集上进行了对比实验,实验结果如表4所示。通过将单模态和多模态的方法进行比较,可以看出,多模态方法普遍比单模态方法表现要好。说明在面向目标的情感分类任务中,图片信息可以补充文本信息,从而提升分类效果。通过比较纯文本与纯视觉模型,可以明显看出纯文本模型比纯视觉模型性能表现更好,这表明该任务对视觉信息的依赖性较小而更多依赖于文本。通过比较多模态方法如图4(a)、(b)所示,可以看出Res-MGAN-TFN模型在两个数据集上的性能最低,这主要是由于其对文本-图像模态间信息的交互较弱。TomBERT模型的效果比较好,这主要是由于TomBERT模型在建模目标词与文本和图片之间的交互能够显著提升任务性能。与多模态方法TomBERT模型相比,本文模型在Twitter-15和Twitter-17数据集上实现了1.05%和0.62%的精度提升和2.33%和1.32%的$F1$提升。与纯文本方法BERT+BL模型相比,本文模型在Twitter-15和Twitter-17数据集上实现了3.96%和1.18%的精度提升和4.04%和3.23%的$F1$提升。与纯视觉方法Res-Target模型相比,本文模型在Twitter-15和Twitter-17数据集上实现了18.32%和12.31%的精度提升和27.60%和15.37%的$F1$提升,实验验证了本文模型的有效性。

3.5 模型参数分析

为了验证参数对DRNN的影响,本节在两个数据集上对时间步ts,学习率lr,批量大小bs三个参数进行了定量分析。

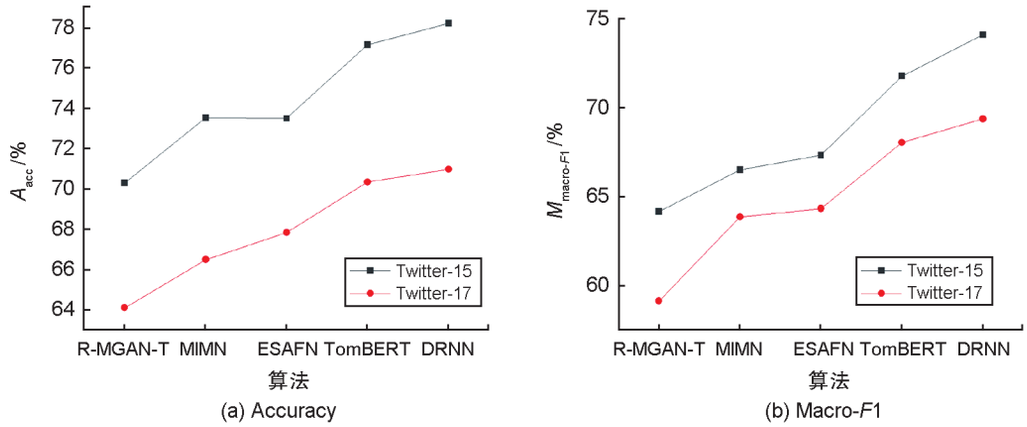

图4 多模态算法在数据集Twitter-15和Twitter-17上两种评价指标的实验结果

Fig. 4 Experimental results of the multimodal algorithms on two evaluation metrics on the datasets Twitter-15 and Twitter-17

表5 消融实验结果

Table 5 Results of Ablation studies

方法	Twitter-15		Twitter-17	
	Acc.	*F*1	Acc.	*F*1
移除图像部分	72.90	65.75	66.07	63.39
移除 TTIRNN	75.22	68.60	69.57	68.03
移除GRU	75.23	68.20	68.52	67.13
移除图文交互层	76.86	71.62	69.72	67.41
DRNN	78.20	74.08	70.96	69.35

时间步:为了探索时间步 ts 对性能的影响,设置参数 $lr=3\times10^{-5}$, $bs=32$,观察 $ts\in\{1,2,3,4,5\}$时对实验的影响。实验结果如图5(a)所示,可以发现模型的性能随着 ts 的增加而逐渐提升,这表明增加时间步可以有效融合注意力特征。当参数较小时,模型没有学习到有效融合特征,当参数继续增加时,模型性能下降,这有可能是模型出现了过拟合。

学习率:为了探索 lr 对模型性能的影响,设置参数 $ts=3$, $bs=32$,观察 $lr\in\{1\times10^{-5}, 2\times10^{-5},\cdots,8\times10^{-5}\}$时对实验的影响。实验结果如图5(b)所示。可以发现随着学习率的变化,模型性能在 3×10^{-5} 和 4×10^{-5} 表现较好。

批量大小:为了探索 bs 对性能的影响,设置参数 $ts=3, lr=3\times10^{-5}$,观察 $bs\in\{4,8,16,32,64\}$时对实验的影响。实验结果如图5(c)所示。可以发现,当批大小适中(即32、16和8),整体性能较好。说明更大或更小的批量大小可能会影响模型的通用性和参数的收敛性。

3.6 消融实验

为了验证DRNN不同模块的有效性,进行了消融实验,实验结果如表5所示。

移除图像部分:模型只将方面文本信息融合而不考虑图像信息。实验结果发现,模型在两个数据集上出现了大幅度地下降。说明图像信息对模型有着明显的影响。

移除 TTIRNN:模型移除了方面词对上下文之间的交互,只单独考虑模态间的交互。实验结果发现,模型在两个数据集上的表现有所下降。该结果表明,在移除该模块的情况下,会减弱文本中方面词与上下文信息的相关性,从而导致了实验结果下降。

移除图文交互层:模型直接将图像信息与方面-文本表示进行拼接送入情感分类层进行分类。实验结果表明,移除模态间的表示会导致模型在两个数据集上的效果出现下降。

移除 GRU:模型将方面文本表示与图像文本表示直接进行融合。实验结果表明,忽略了图像信息的噪声过滤会导致实验结果有所下降。综上所述,DRNN模型中的各部分模块对模型的效果都产生了一定的贡献。

4 结论

本文提出了一种双通道循环神经网络模型,有效解决了在面向目标的多模态情感分析任务中模态内和模态间信息融合不充分以及图像信息中存在噪声的问题。通过与其他跨模态模型、纯文本和纯视觉模型在两个基准数据集上进行对比实验,实验结果表明DRNN相较于其他基线模型在性能方面有着较大的优势,并且通过消融实验说明本文模型中的每一个模块都对模型具

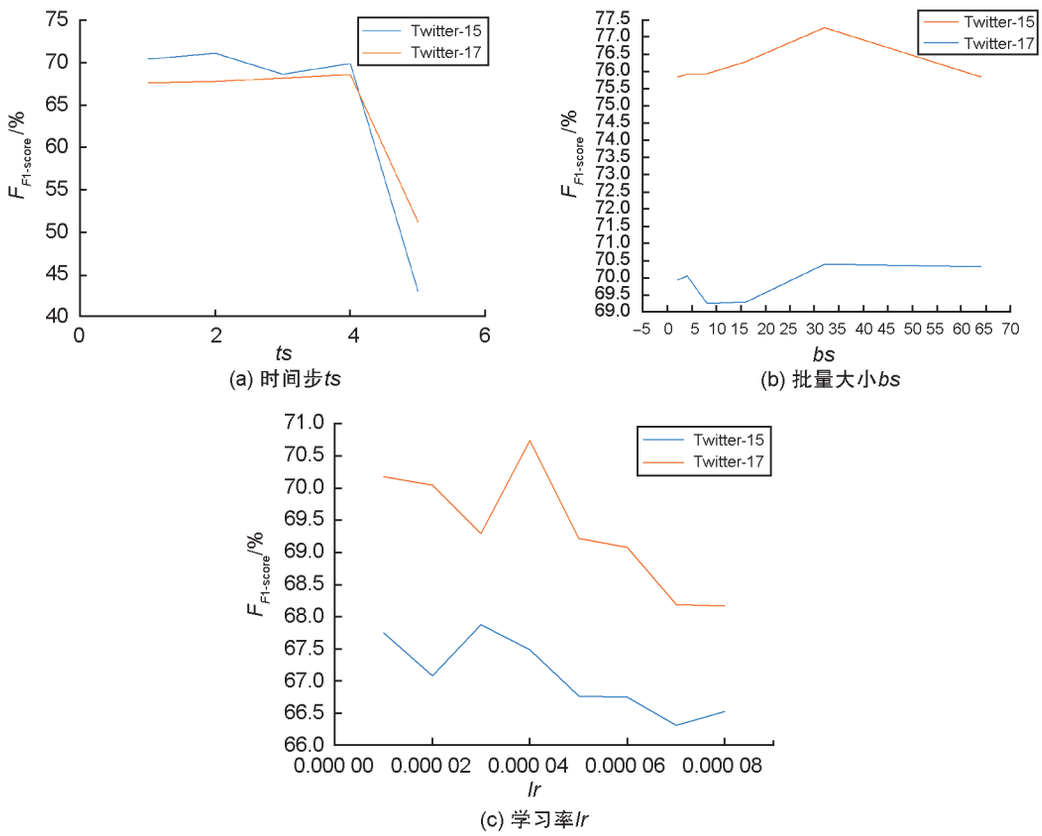

图5 DRNN模型在数据集Twitter-15和Twitter-17上的不同参数对*F*1-Score的影响

Fig. 5 Effects of different parameters of DRNN model on *F*1-Score on two datasets Twitter-15 and Twitter-17

有贡献。接下来的工作中我们会对模型进行进一步改进。例如,可以考虑引入情感领域知识,从而提升面向目标的多模态情感分类的效果。

参考文献:

[1] XUE W, LI T. Aspect Based Sentiment Analysis with Gated Convolutional Networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 2514-2523. DOI: 10.18653/v1/p18-1234.

[2] TANG D, QIN B, FENG X, *et al*. Effective LSTMs for Target-Dependent Sentiment Classification[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, 2016: 3298-3307.

[3] LI X, BING L D, LAM W, *et al*. Transformation Networks for Target-oriented Sentiment Classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 946-956. DOI: 10.18653/v1/p18-1087.

[4] XU N, MAO W J, CHEN G D. Multi-interactive Memory Network for Aspect Based Multimodal Sentiment Analysis[J]. *Proc AAAI Conf Artif Intell*, 2019, **33**(1): 371-378. DOI: 10.1609/aaai.v33i01.3301371.

[5] YU J F, JIANG J, XIA R. Entity-sensitive Attention and Fusion Network for Entity-level Multimodal Sentiment Classification[J]. *IEEE/ACM Trans Audio Speech Lang Process*, 2020, **28**: 429-439. DOI: 10.1109/TASLP.2019.2957872.

[6] ZHANG Z, WANG Z, LI X N, *et al*. ModalNet: an Aspect-level Sentiment Classification Model by Exploring Multimodal Data with Fusion Discriminant Attentional Network[J]. *World Wide Web*, 2021, **24**(6): 1957-1974. DOI: 10.1007/s11280-021-00955-7.

[7] SHARMA A, DEY S. A Boosted SVM Based Ensemble Classifier for Sentiment Analysis of Online Reviews[J]. *SIGAPP Appl Comput Rev*, 2013, **13**(4): 43-52. DOI: 10.1145/2577554.2577560.

[8] KAMAL A, ABULAISH M. Statistical Features Identification for Sentiment Analysis Using Machine Learning Techniques[C]//2013 International Symposium on Computational and Business Intelligence. 2014: 178-181.

DOI: 10.1109/ISCBI.2013.43.

[9] YU Z W, WONG R K, CHI C H, *et al*. A Semi-supervised Learning Approach for Microblog Sentiment Classification[C]//2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity). 2016: 339−344. DOI: 10.1109/SmartCity.2015.94.

[10] DONG L, WEI F R, TAN C Q, *et al*. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 49−54. DOI: 10.3115/v1/p14-2009.

[11] HE R D, LEE W S, NG H T, *et al*. Exploiting Document Knowledge for Aspect-level Sentiment Classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 579−585. DOI: 10.18653/v1/p18-2092.

[12] WANG J J, LI J, LI S S, *et al*. Aspect Sentiment Classification with both Word-level and Clause-level Attention Networks[C]//Proceedings of the Twenty−Seventh International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 4439−4445. DOI: 10.24963/ijcai.2018/617.

[13] TANG D Y, QIN B, LIU T. Aspect Level Sentiment Classification with Deep Memory Network[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 214−224. DOI: 10.18653/v1/d16-1021.

[14] FAN C, GAO Q H, DU J C, *et al*. Convolution-based Memory Network for Aspect-based Sentiment Analysis [C]//SIGIR'18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2018: 1161−1164. DOI: 10.1145/3209978.3210115.

[15] PHAN M H, OGUNBONA P O. Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 3211−3220. DOI: 10.18653/v1/2020.acl-main.293.

[16] CAMBRIA E, LI Y, XING F Z, *et al*. SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 105−114. DOI: 10.1145/3340531.3412003.

[17] GU T, ZHAO H, HE Z, *et al*. Integrating External Knowledge into Aspect-based Sentiment Analysis Using Graph Neural Network[J]. *Knowl Based Syst*, 2023, **259**: 110025. DOI: 10.1016/j.knosys.2022.110025.

[18] YU J F, JIANG J. Adapting BERT for Target-oriented Multimodal Sentiment Classification[C]//Proceedings of the Twenty−Eighth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019: 5408−5414. DOI: 10.24963/ijcai.2019/751.

[19] WANG J, GU D, YANG C, *et al*. Targeted Aspect Based Multimodal Sentiment Analysis: an Attention Capsule Extraction and Multi-head Fusion Network" [EB/OL]. arXiv Preprint: 2103.07659, 2021.

[20] JU X C, ZHANG D, XIAO R, *et al*. Joint Multi-modal Aspect-sentiment Analysis with Auxiliary Cross-modal Relation Detection[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 4395−4405. DOI: 10.18653/v1/2021.emnlp-main.360.

[21] YU Z W, WANG J, YU L C, *et al*. Dual-Encoder Transformers with Cross-modal Alignment for Multimodal Aspect-based Sentiment Analysis[C]//Proceedings of the 2nd Conference of the Asia−Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Stroudsburg, PA: ACL, 2022: 414−423.

[22] YANG L, NA J C, YU J F. Cross-modal Multitask Transformer for End-to-end Multimodal Aspect-based Sentiment Analysis[J]. *Inf Process Manag*, 2022, **59**(5): 103038. DOI: 10.1016/j.ipm.2022.103038.

[23] LING Y, YU J F, XIA R. Vision-language Pre-training for Multimodal Aspect-based Sentiment Analysis[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 2149−2159. DOI: 10.18653/v1/2022.acl-long.152.

[24] HE K M, ZHANG X Y, REN S Q, *et al*. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770−778. DOI: 10.1109/cvpr.2016.90.

[25] DEVLIN J, CHANG M, LEE K, *et al*. BERT: Pre-

training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. arXiv Preprint: 1810.04805, 2018.

[26] FAN F F, FENG Y S, ZHAO D Y. Multi-grained Attention Network for Aspect-level Sentiment Classification [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 3433-3442. DOI: 10.18653/v1/d18-1380.

[27] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all You Need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010. DOI: 10.5555/3295222.3295349.

[28] ZADEH A, CHEN M H, PORIA S, *et al*. Tensor Fusion Network for Multimodal Sentiment Analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 1103-1114. DOI: 10.18653/v1/d17-1115.