文章编号：2095-2163(2023)11-0044-05 中图分类号：TP391.1 文献标志码：A

融合注意力机制的多模态影评情感分析

温作前，张云华
（浙江理工大学 信息学院，杭州 310018）

摘 要：对影评进行情感分析有助于为用户提供更好的服务。针对单模态模型只能选择单一的语义信息和多个模态间的信息无法进行共享等问题，本文提出一种融合注意力机制的 BiLSTM-VGG16 的中文影评情感分析模型。首先使用 BiLSTM、VGG16 分别提取文本信息和图像信息的特征值，在注意力机制的作用下，突出文本中情感信息量的部分。在决策层融合文本特征和图像特征，最后使用 *softmax* 函数实现影评情感级分类。通过爬虫获取腾讯视频的评论对模型进行训练和测试。模型准确率为 0.854，召回率为 0.875，*F* 值为 0.854，*AUC* 为 0.861。由实验结果得出，相比于其他单模态分析模型，多模态分析模型在影视评论情感分析方面取得更好的效果。

关键词：VGG16；BiLSTM；多模态；情感分析；注意力机制

Multimodal film review analysis of sentiment based on attention mechanism

WEN Zuoqian, ZHANG Yunhua
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Sentiment analysis of film reviews is helpful to provide better services for users. The single-modal model can only select a single semantic information, and the information between multiple modalities cannot be shared. In this paper, a BiLSTM-VGG16 Chinese movie review sentiment analysis model with attention mechanism is proposed. Firstly, BiLSTM and VGG16 are used to extract the eigenvalues of text information and image information respectively. Under the action of attention mechanism, emotional information in the text is highlighted. At the decision level, text features and image features are integrated. Finally, *softmax* function is used to achieve the emotion classification of film reviews. The model is trained and tested by comments obtained from Tencent Video via crawler. The model accuracy is 0.854, the recall is 0.875, the *F* - Measure is 0.854, and the *AUC* is 0.861. The experimental results show that compared with other single modal sentiment analysis models, the multimodal sentiment analysis model can achieve better results.

Key words: VGG16; BiLSTM ; multimodal; emotional analysis; attention mechanism

0 引 言

情感是个人面对客观事物的态度体验。也是个人对客观世界智能的、主观的一种表现。人们表达情感的方式是多样的。一段文字、一条语音、一张图片，都是人们在某种场景下对特定事件的情绪表现方式。而电影评论则是人们对电影本身的一种情感表达。通过收集网络上海量的影评文本和影评图片进行情感分析，能够有助于用户在网络上有更好的体验。随着信息技术的不断发展，B 站、优酷、腾讯视频等各类观影平台的普及使得文本数据和数据类型越来越丰富。自深度学习不断发展以来，越来越多的学者开始使用深层神经网络进行情感分析[1]。

近年来，国内外学者针对影评情感分析做了很多研究。张尚乾等学者[2]利用影评本体特征以及影评情感特征与长短期记忆网络(LSTM)融合进行文本级情感分类。张碧依等学者[3]提出基于 XLNet 预训练语言模型对影评信息进行分布式表示，再利用 BiLSTM 进行深层语义分析，最后使用 *softmax* 函数实现情感级分类。辛雨璇等学者[4]利用 TF-IDF 和贝叶斯分类对影评文本进行情感分析。

但是单模态文本数据中所包含的信息不够全面，在某些情况下只依靠目标文本难以准确判断目标的情感状态[5]。一个在影评中较为常见的例子

作者简介：温作前(1998-)，男，硕士研究生，主要研究方向：智能计算、数据挖掘。
通讯作者：张云华 (1965-)，男，博士，研究员，主要研究方向 ：软件工程、智慧医疗、智能信息处理。Email:605498519@qq.com

收稿日期：2022-11-10

是反讽。在反讽中，文本内容表达的情感往往是较为中性和积极的，但图片所表达的情感往往是消极的。如，“这电影可真好看啊！”，仅仅从文本上看情绪是积极的，但当配上一个”咒骂”的表情，整个句子的情感将发生本质变化。这种情况使用单模态模型很难彻底解决问题。

为此，本文以多模态影视评论为研究对象，在注意力机制的作用下突出文本中情感信息特征和图像特征，对高权重的数据向量进行特征融合再进行情感的分类，最后对普通的单模态模型效果进行分析。通过结论论证，本文构建的 VGG16-BiLSTM 多模态模型对于影视评论有更高的情感识别效率，深入挖掘文本信息，识别隐晦情感。

1　相关知识

1.1　卷积神经网络 VGG16

卷积神经网络（CNN）是一种深层的监督学习神经网络，主要包含卷积层和池化层。其中，卷积层用于提取图像特征，池化层用于提取和优化特征。卷积神经网络在低隐藏层通常由卷积层和最大池化层组成，最大池化层可用来强化特征。高层是全连接层，起到分类器的作用。第一个全连接层的输入是由低隐藏层所提取且优化的图像特征。最后一层输出层使用逻辑回归、*softmax* 回归或者支持向量机对图像特征进行分类。

VGG16 网络模型共有 6 个块结构，每个块结构的通道数量相同，其中卷积层和全连接层均有权重系数，故也称权重层。权重层共 16 层，其中卷积层有 13 层，全连接层有 3 层。VGG 全部采用 3 * 3 的卷积核，步长和 *Padding* 均为 1，2 * 2 的最大池化核，步长为 2，*Padding* 为 0。VGG 通过叠加多个3 * 3卷积核使得最终拥有了 5 * 5 的卷积核以及 7 * 7 的卷积核的感受野。在感受野相同的情况下，多个 3 * 3 的卷积核可以大幅度增加非线性表达能力。

1.2　长短期记忆神经网络

RNN 常用于自然语言的处理，这依赖于 RNN 能够记忆已经学习到的信息，并结合当前的信息得到当前输出与之前信息的关系。RNN 的时序结构如图 1 所示。

由图 1 可以看出，t 时刻 RNN 的输入包含当前时刻的输入 X_t 和上一时刻隐藏层的状态 h_{t-1}。这样的设计在处理长序列时很容易将一些无效的信息也进行记忆传递，同时会出现梯度爆炸和梯度消失的问题，使得较长距离的文字相关性下降。

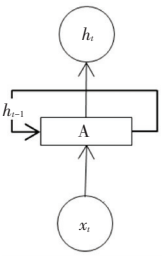

图 1　RNN 结构图

Fig. 1　Structure of RNN

长短期记忆神经网络（LSTM）在 RNN 的基础上利用门（gate）机制控制输入信息，输出信息，以此记忆或者遗忘长距离信息。LSTM 单元结构如图 2 所示。

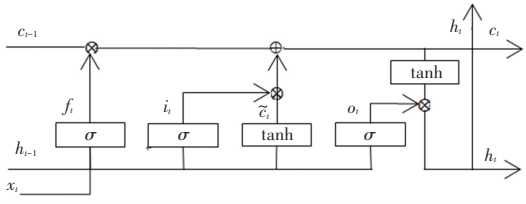

图 2　LSTM 单元结构

Fig. 2　The unit structure of LSTM

由图 2 可知，LSTM 的构成有记忆细胞 C、更新门 i、遗忘门 f 和输出门 o。其中，更新门用于决定当前时刻的信息对输出的影响程度；遗忘门用于保存或者遗忘之前记忆的信息；输出门用于描述当前时刻记忆细胞输出与下一时刻输入信息的相关性。记忆细胞 C 表示某一时刻所处理的特征信息。LSTM 工作过程中主要设计各符号的阐释解读见表 1。研究中，将用到以下数学公式：

$$\tilde{c}_t = \tanh(\boldsymbol{W}_C \cdot [h_{t-1}, x_t] + b_c) \tag{1}$$

$$i_t = \sigma(\boldsymbol{W}_u \cdot [h_t, x_t] + b_u) \tag{2}$$

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, x_t] + b_o) \tag{4}$$

$$c_t = i * {}_t\tilde{c}_t + f_t * c_t \tag{5}$$

$$a_t = o_t * \tanh(c_t) \tag{6}$$

表 1　符号解释说明表

Tab. 1　Symbolic interpretation

符号	说明
C_{t-1}	$t-1$ 时刻记忆细胞状态
h_{t-1}	$t-1$ 时刻网络输出值
c_t	t 时刻记忆细胞状态
h_t	t 时刻网络输出值
$\tilde{c}_t$	t 时刻记忆细胞的候选值
x_t	t 时刻网络输入值
$tanh$	反正切激活函数
f_t	t 时刻的遗忘门
i_t	t 时刻的更新门
o_t	t 时刻的输出门

通常情况下,文档中每个词汇不但依赖之前的元素,而且还与之后的元素关系密切。为此可知,BiLSTM 是一种双向的 LSTM,如 t 时刻的 BiLSTM 包含的信息为 t 时刻之前 LSTM 的信息加上 t 时刻之后 LSTM 的信息。如句子向前的 LSTML 依次输入“电影”、“好”、“看”得到 3 个向量 $\{\boldsymbol{a}_{l0},\boldsymbol{a}_{l1},\boldsymbol{a}_{l2}\}$。后向的 LSTMR 依次输入“看”、“好”、“电影”得到 3 个向量 $\{\boldsymbol{a}_{r0},\boldsymbol{a}_{r1},\boldsymbol{a}_{r2}\}$. 再进行拼接得到 $\{[\boldsymbol{a}_{l0},\boldsymbol{a}_{r0}],[\boldsymbol{a}_{l1},\boldsymbol{a}_{r1}],[\boldsymbol{a}_{l2},\boldsymbol{a}_{r2}]\}$,即 $\{\boldsymbol{a}_0,\boldsymbol{a}_1,\boldsymbol{a}_2\}$。

1.3 注意力机制

注意力机制是一种筛选信息的方法,能够进一步缓解 LSTM 中长期依赖的问题。注意力机制实现分 3 步进行,如:

(1)通过人工设置的超参数或者通过动态生成的向量确定查询向量。

(2)使用打分函数中的加性模型计算出输入特征与查询向量的相关性,得到概率分布。

(3)利用注意力机制对输入的特征进行加权平均,得到最终的特征信息。

2 模型构建

2.1 融合注意力机制的多模态影评情感分析研究框架

本文选取腾讯视频的影评中的文本和图像作为研究对象,提出融合注意力机制的影评情感分析模型,主要思路是在融合注意力机制的情况下,针对文本和图像进行训练,强化用户情感词,更全面地捕获全文信息。模型组成部分有:使用 Word2Vec 并结合负采样对影评文本进行词向量化;使用 BiLSTM 模型对影视评论的文本信息进行特征信息的提取;在表情图像特征识别上,使用 VGG16 对表情图像的特征进行提取;利用注意力机制对文本和表情图像中的情感信息特征进行强化;在中间层进行多模态信息特征的融合。最后,由决策层根据融合的特征信息进行情感分类。融合注意力机制后如图 3 所示。

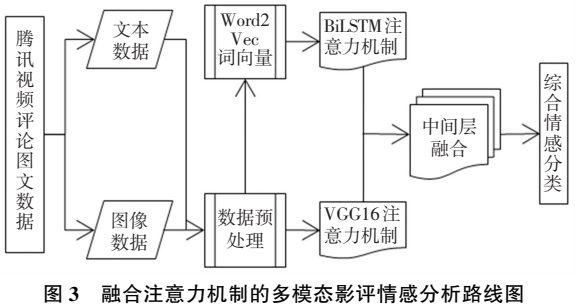

图 3 融合注意力机制的多模态影评情感分析路线图

Fig. 3 Roadmap of multimodal emotion analysis of movie reviews integrating attention mechanism

2.2 融合注意力机制的 BiLSTM 的文本特征提取

关于融合注意力机制的 BiLSTM 的模型有 3 层,涉及词向量化、特征提取和注意力层。Word2Vec 将传入的文本编码转化为特征向量,使用卷积过滤器进行特征提取,再进行注意力分析,最后实现情感分析。形成 ATT-BiLSTM 模型。融合注意力机制的文本情感分析流程如图 4 所示。

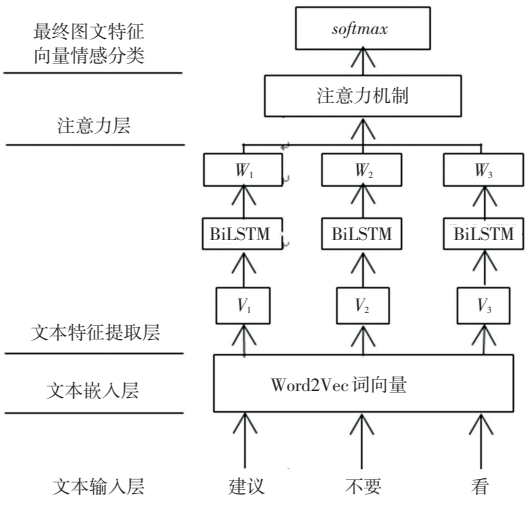

图 4 融合注意力机制的文本情感分析流程图

Fig. 4 Flow chart of text emotion analysis integrating attention mechanism

2.3 VGG16 图像特征提取与注意力加权

在图像情感分析中,VGG16 提取影视评论中表情图像特征,利用注意力机制,提取图像局部关键位置的信息,形成 Att-VGG16 模型,进行情感分析。注意力加权和图像特征情感分析流程如图 5 所示。

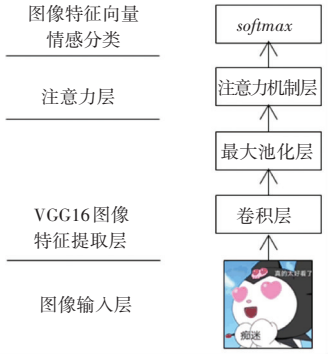

图 5 注意力加权和图像特征情感分析流程图

Fig. 5 Flow chart of attention weighting and image feature emotion

2.4 图像文本特征加权融合

对于影视评论情感进行分析时,虽然图片能够直观提供视觉信息,但是图片描述情感过于单一。尽管文本特征描述情感更丰富,但是文本描述情感不直观,所以独立的文本输入或者单独的图片输入无法满足高精度的情感分类需求。因此,需要融合

图片特征和文本特征。融合方式采用决策级融合，也称后期融合。在决策层将文本分类结果与图像情感分类结果相融合，附上对应权重，能够较大限度地保留不同模态对情感倾向的影响，以此获得最终的结果分类。

在权重分配过程中，P_t 表示文本分类的概率，P_i 表示图像的分类的概率，P_c 是分别给 P_t 和 P_i 分配 W_t（文本权限）、W_i（图像权限）并且相加得到，根据 P_c 得出后期融合后的输出分类。融合函数 P_c 如式(7)所示：

$$P_c = P_t * W_t + P_i * W_i \tag{7}$$

情感分类在图文特征融合之后，$\boldsymbol{o}^{\mathrm{T}}$ 作为最终表示，采用 *softmax* 函数作为输出层。函数表达为：

$$y = softmax(\boldsymbol{W}_s\boldsymbol{o}^{\mathrm{T}} + b_s) \tag{8}$$

其中，b_s 是可学习的偏置向量；$\boldsymbol{W}_s$ 是可学习的输出层的权重矩阵；y 是预测的情感极性分布。

通过使用交叉熵损失函数 $L(\theta)$ 对所提出的模型进行测试。计算公式如下：

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_i^j \log \hat{y}_l^l + \lambda \sum_{k=0}^{N} \theta_k^2 \tag{9}$$

其中，N 是训练集中影评片段；M 是情感类别的数量；y_s 是第 s 个影评的真实情感类别。

3　实验与分析

3.1　数据集

本实验的数据集选用腾讯视频，通过爬虫软件，在视频评论中，爬取评论的文本信息和图片信息。对于文本数据需要进行适当处理，如删除不合规的字符，删除标点符号。在词嵌入方面使用 Word2Vec，将执行词进行向量化。对于图像，先删除广告图像等无关图像，再将图像调整成大小为 227×227×3，进行图像裁剪。

3.2　评价指标

为了更加准确计算出模型所预测的情感分类与实际情感分类的区别，采用多种评价标准。如准确率、召回率、F 值、AUC 等评价指标进行模型性能的综合判断，具体见式(10)～(13)：

$$ACC(\text{准确率}) = \frac{TP + TN}{TP + TN + FP + TN} \tag{10}$$

$$REC(\text{召回率}) = \frac{TP}{TP + FP} \tag{11}$$

$$F(F\text{ 值}) = \frac{ACC * (\frac{TP}{TP + FN}) * 2}{ACC + (\frac{TP}{TP + FN})} \tag{12}$$

$$AUC = \frac{1}{2}(\frac{TP}{TP + FP} + \frac{FP}{FP + TN}) = \frac{1}{2}(R_C + S_p) \tag{13}$$

其中，TP 表示观众对影视作品持积极情感、并且预测为积极情感；FN 表示观众对影视作品持积极情感、预测为消极情感；FP 表示观众对影视作品持消极情感、预测为积极情感；TN 表示观众对影视作品持消极情感、预测也是消极情感。

3.3　实验结果

利用训练集数据训练后统计的训练集损失结果如图 6 所示。从图 6 结果可知，BiLSTM-VGG16 模型的 AUC 值为 0.86.相比于 BiLSTM 和 VGG16，分别增加了 0.127 和 0.11。AUC 的值越趋近于 1，模型的处理能力越好。这体现了图像和文本在影视评论的情感分析中起到了相互引导、相互弥补的作用。模型的训练集损失曲线如图 6 所示。

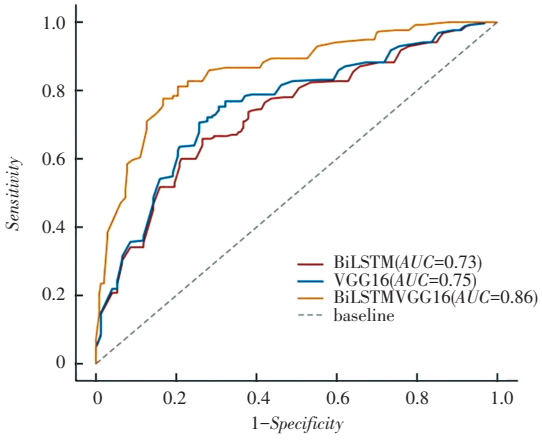

图 6　模型的训练集损失

Fig. 6　Training loss of the model

为了进一步证实本实验模型的有效性，基于同一数据集对 VGG16、BiLSTM 、BiLSTM-VGG16 等模型使用准确率、召回率、F 值等指标进行评估具体评价，结果见表 2。

表 2　预测模型的评价结果

Tab. 2　Evaluation results of prediction model　%

模型	准确率	召回率	F 值	AUC
BiLSTM	78.69	76.73	75.86	73.37
VGG16	79.91	81.77	75.93	75.18
BiLSTM-VGG16	85.37	87.53	85.41	86.07

实验数据表明，相比于单独使用文本或者图像，多模态下对影视评论进行情感分析的效果更好。仿真后可知，准确率为 85.37%、召回率为 87.53%、F 值为 85.41%、AUC 为 86.07%。相比于 VGG16、BiLSTM 都有所提升。

4 结束语

本文针对现有的单模态影评情感分析模型研究存在的分类不精准、各模态间信息无法共享、难以分辨反讽文本等问题,提出了基于注意力机制的多模态 BiLSTM-VGG16 模型。利用 BiLSTM 和 VGG16 分别对影视评论的文本和影视评论的表情图像进行特征的提取和分类,再将提取的特征信息进行融合。在理论上,不同模态形式是相互独立,但是出现在同一语境中时,不同模态会相互影响。例如,图像和文本的情感表达倾向一致时,会增强情感的表达,当二者相反时则会出现反讽的现象。在注意力机制的作用下,提高对正确情感的捕获能力。通过对采集到的影视评论数据进行实验,验证本模型较好的情感分析能力,分析效果好于 VGG16、BiLSTM 等模型。该模型可为影评情感分析提供参考。

参考文献

[1] NAMGIL L, ANDRZEJ C. Fundamental tensor operations for large-scale data analysis using tensor network formats[J]. Multidimensional Systems and Signal Processing, 2018, 29(3): 921-960.

[2] 张尚乾,刘知一. 基于关键特征的影评细粒度情感分析[J]. 现代电影技术,2022(06):16-21.

[3] 张碧依,陶宏才. 基于 XLNet-BiLSTM 模型的中文影评情感分析[J]. 成都信息工程大学学报,2021,36(03):264-269.

[4] 辛雨璇,王晓东. 基于文本挖掘的电影评论情感分析研究[J]. 牡丹江师范学院学报(自然科学版),2021(01):25-28.

[5] 陈艳,陈宏松,安俊秀,等. 基于 BERT-VGG16 的多模态情感分析模型[J]. 成都信息工程大学学报,2022,37(04):379-385.

[6] 周宁,钟娜,靳高雅,等. 基于混合词嵌入的双通道注意力网络中文文本情感分析[J]. 数据分析与知识发现, 2023,7(03):58-68.