智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合交叉注意力的突发事件多模态中文反讽识别模型

胡文彬, 陈龙, 黄贤波, 陈晨, 仲兆满

引用本文:
胡文彬,陈龙,黄贤波,陈晨,仲兆满. 融合交叉注意力的突发事件多模态中文反讽识别模型[J]. 智能系统学报, 2024, 19(2): 392–400.
HU Wenbin, CHEN Long, HUANG Xianbo, et al. A multimodal Chinese sarcasm detection model for emergencies based on cross attention[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 392–400.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202212011

DOI: 10.11992/tis.202212011
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20231220.1740.002

融合交叉注意力的突发事件多模态中文反讽识别模型

胡文彬[1,2], 陈龙[1], 黄贤波[1], 陈晨[1], 仲兆满[1,2]
（1. 江苏海洋大学 计算机工程学院, 江苏 连云港 222005; 2. 江苏省海洋资源开发研究院, 江苏 连云港 222005）

摘　要: 网民在社交媒体参与突发事件讨论时, 时常会采用反讽修辞方式表达对事件的看法, 此举导致情感分析的难度增加, 且已有中文反讽识别对社交媒体中网民发布的多模态评论研究较少, 有必要对图文多模态中文反讽识别进行深入研究。运用交叉注意力机制捕捉模态间的不一致性表达, 提出融合交叉注意力的多模态中文反讽识别模型 (fuse cross attention model, FCAM)。在模型中, 首先运用 TextCNN(text convolutional neural networks) 和 ResNet(deep residual network) 分别提取中文文本浅层特征和图像特征, 再运用交叉注意力机制分别得到文本层和图像层的注意力特征, 按照残差方式分别实现文本浅层特征和文本层注意力特征的连接、图像特征和图像层注意力特征的连接, 使用注意力机制融合 2 个特征表示, 经过分类层得到反讽分类结果。基于某一地区新冠疫情期间相关话题的微博评论数据, 构建出突发公共卫生事件多模态中文反讽数据集, 在该数据集上试验验证, 相较于基准模型, FCAM 具有一定的优越性。

关键词: 突发事件; 社交媒体; 多模态评论; 中文反讽识别; 中文反讽数据集; 交叉注意力机制; 注意力机制; 情感分析
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2024)02−0392−09

中文引用格式: 胡文彬, 陈龙, 黄贤波, 等. 融合交叉注意力的突发事件多模态中文反讽识别模型 [J]. 智能系统学报, 2024, 19(2): 392–400.
英文引用格式: HU Wenbin, CHEN Long, HUANG Xianbo, et al. A multimodal Chinese sarcasm detection model for emergencies based on cross attention[J]. CAAI transactions on intelligent systems, 2024, 19(2): 392–400.

A multimodal Chinese sarcasm detection model for emergencies based on cross attention

HU Wenbin[1,2], CHEN Long[1], HUANG Xianbo[1], CHEN Chen[1], ZHONG Zhaoman[1,2]
(1. School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222005, China; 2. Jiangsu Institute of Marine Resources Development, Lianyungang 222005, China)

Abstract: Internet users often use sarcasm when discussing emergencies on social media, which complicates emotional analysis. In addition, there is a lack of research on multimodal comments, particularly those in Chinese, and their use of sarcasm on social media platforms. Therefore, it is necessary to delve deeper into sarcasm detection in multimodal Chinese content, specifically within images and text. To address this need, we propose a multimodal Chinese sarcasm detection model called the fuse cross-attention model (FCAM). This model incorporates a cross-attention mechanism to identify inconsistencies between modes. The text convolutional neural network (TextCNN) is used to extract basic features of Chinese text, while the deep residential network (ResNet) is used to extract image features. The cross-attention mechanism is used to obtain attention features from the text and image layers. The residual method is employed to establish a connection between the basic text features and the text layer's attention features, as well as a link between the image features and the image layer's attention features. These two feature representations are fused using the attention mechanism, resulting in the sarcasm classification results through the classification layer. We have constructed a multimodal Chinese sarcasm data set based on Weibo comment data related to the COVID-19 pandemic in a specific region. Experimental testing on this data set confirms that FCAM holds certain advantages over the benchmark model.

Keywords: emergency; social media; multimodal comment; Chinese sarcasm detection; Chinese sarcasm dataset; cross-attention mechanism; attention mechanism; sentiment analysis

公众参与网络评论时发表的观点, 能够反映出网民对热点事件的态度和看法。情感分析通过挖掘网民评论的语义信息, 对舆情分析具有重大

收稿日期: 2022−12−08.　**网络出版日期**: 2023−12−21.
基金项目: 国家自然科学基金项目 (72174079); 江苏省“青蓝工程”优秀教学团队 (2022-29).
通信作者: 胡文彬. E-mail: hwb1008@163.com.

意义。而反讽作为一种修辞方式, 常常被网民当作隐性表达观点的工具, 挑战情感分析且易干扰管理者对突发事件中网民的情感认知, 引发舆情危机。Zhang 等[1] 发现反讽表达在政府、品牌和政治等突发事件话题中更常见, 在食物、健康等话题中较少。有效识别突发事件中网民的反讽表达, 可以精准识别突发事件中网民的情感, 从而把握网络评论的舆情导向。

深度学习在纯文本反讽识别方面解决了手工提取特征耗时的难题, 自动提取表征能力强的特征, 显著提高了反讽识别的准确率。Ghosh 等[2] 提出了一个由 CNN、LSTM 和 DNN 网络组成的神经网络模型, 试验证明是当时最好的反讽检测方法。Khotijah 等[3] 使用 LSTM 检测 2 种语言的反讽, 试验表明平衡数据集上的准确率高于不平衡数据集。Lou 等[4] 为检测上下文的不一致表达, 对每个句子构建情感图和依赖图, 构建出情感依赖图卷积网络框架, 得到 76% 的准确率。中文反讽识别方面, 孙晓等[5] 提出了一种基于神经网络输出层融合的混合模型, 融合 CNN 和 LSTM 输出的特征, 试验表明优于单一神经网络模型。卢欣等[6] 提出了一种融合反讽特征的卷积神经网络模型, 试验获得了 82% 的准确率。樊小超等[7] 使用 ELMo 从反讽文本中训练得到词嵌入表达, 并融合基于词性和风格信息的语义表示, 使用 Bi-LSTM 和 CNN 进行反讽分类, 准确率为 79%。

多模态反讽识别研究工作国外起步较早, 中文领域还处在探索阶段。Schifanella 等[8] 使用 2 种方法完成反讽识别任务, 一种是使用 SVM 方法将文本和图像特征相结合; 另一种将使用深度神经网络获取的图像特征与基于 unigram 的文本特征简单融合。试验证明第 2 种方法的性能不如第 1 种。Sharma 等[9] 使用 LSTM、BERT 和 USE 对输入分别编码并简单融合, 未考虑模态间的交互关系。在捕捉模态间不一致性的设计上, Sangwan 等[10] 使用 Bi-GRU 提取文本特征、VGG-16 提取图像特征, 通过使用 OCR 从图像中提取出文本, 作为一种模态输入至 Bi-GRU 中。Cai 等[11] 提出了一种结合文本特征、图像特征和图像属性的层次融合模型, 将 3 种模式的特征融合为一个特征向量进行反讽预测。Pan 等[12] 提出了一个基于 BERT 架构的模型, 通过设计注意力机制的方式来捕获多模态反讽识别模态内与模态间的不一致性。Yao 等[13] 结合门机制和引导注意力对多种模态的交互问题进行了研究。Gupta 等[14] 提出了基于 RoBERTa 模型的协同关注模型, 使用 RoBERTa 对文本编码, ResNet 提取图像特征, 通过 FiLMed ResNet 模块处理输入图像, 利用 GRU 获取多模态信息, 通过联合注意力解决输入文本和图像的不一致性问题。

中文反讽识别方面, 张继东等[15] 提出一种多模态深度学习模型, 在特征融合层仅采用了模态特征的简单加权, 未考虑模态间的信息交互。在这一问题上, 刘洋等[16] 使用图神经网络提取文本与图片中的交互信息, 通过注意力机制突出模态特征。对比中文与英文反讽识别研究发现, 目前中文多模态反讽识别研究主要存在以下问题: 1)缺乏规模较大且公开的权威中文反讽研究数据集, 数据集质量受研究者主观判断和手工标注的影响; 2)社交媒体中网民发布的评论多是短文本形式, 即使增加了图片信息, 也难以获取到上下文信息; 3)在图文多模态反讽识别研究中, “模态间矛盾”的捕捉是研究重点, 其中的模态间信息融合与交互模型设计是难点。

针对突发事件中如何捕捉“模态间矛盾”和多模态中文反讽识别无公开数据集问题, 运用交叉注意力机制捕捉模态间不一致性, 使用残差连接的方式和注意力机制进行模态间信息融合, 提出了融合交叉注意力机制的多模态中文反讽识别模型,该模型能有效捕捉模态间交互关系。结合新冠疫情期间网民对某一地区相关话题的多模态评论信息, 构造了试验验证缺乏的数据集。本文的主要贡献如下:

1)本研究运用交叉注意力机制捕捉“模态间矛盾”, 提出融合交叉注意力的多模态中文反讽识别模型 (fuse cross attention model, FCAM), 能精准捕捉中文评论中模态间的不一致性, 以较高的准确率识别出中文反讽评论。

2)针对突发事件中文反讽识别尚无公开数据集问题, 本研究构建了突发公共卫生事件多模态中文反讽识别数据集。

3)本研究提出的融合交叉注意力的多模态中文反讽识别模型, 采用设计良好的多模态融合方式可以更好地进行图文信息交互。

1　突发事件多模态反讽识别模型构建

1.1　突发事件的演化模型

《中华人民共和国突发事件应对法》第 3 条第 1 款规定突发事件是指突然发生, 造成或者可能造成严重社会危害, 需要采取应急处置措施予以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件[17]。信息技术的快速发展滋生网络

舆情，网民间信息传播影响着事件发展。一起突发事件发生之后很快受到民众和自媒体的关注，部分自媒体为获得网民关注，常常在社交媒体中发布未经证实的信息，并做出片面报道，当网民关注到这些信息后，会在表达自己态度看法的同时将信息传播出去，导致网络舆情大规模扩散。此时官方方面在网络舆情爆发时，需要汇集民意、分析并给出解决方案。

当网民评论中充斥着反讽表达时，管理者对网民情感认知产生偏差，对事件处置不当，产生再生舆情，陷入舆论漩涡，损害官方形象。突发事件的演化模型，如图 1 所示。

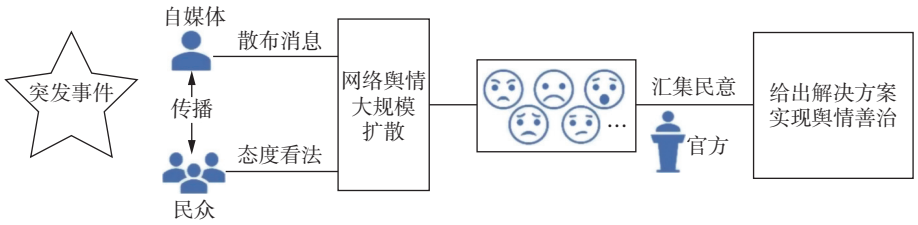

图 1　突发事件的演化模型

Fig. 1　Evolutionary model of emergencies

1.2　多模态中文反讽识别模型构建

为了解决模态间不一致性问题，提出了融合交叉注意力的多模态反讽识别模型。通过训练好的 Word2vec 模型[18]，生成文本词向量，利用 TextCNN 模型[19]提取文本特征，利用 ResNet 模型[20]提取图像特征。运用交叉注意力机制得到文本层面特征和图像层面特征。将得到的文本特征和图像特征分别与浅层特征按残差连接的方式融合，利用注意力机制得到图片与文本的最终特征表示。最后，融合深层次的文本特征与图像特征作为反讽识别特征的最终表示，进行反讽预测。融合交叉注意力的突发事件多模态中文反讽识别模型框架，如图 2 所示。

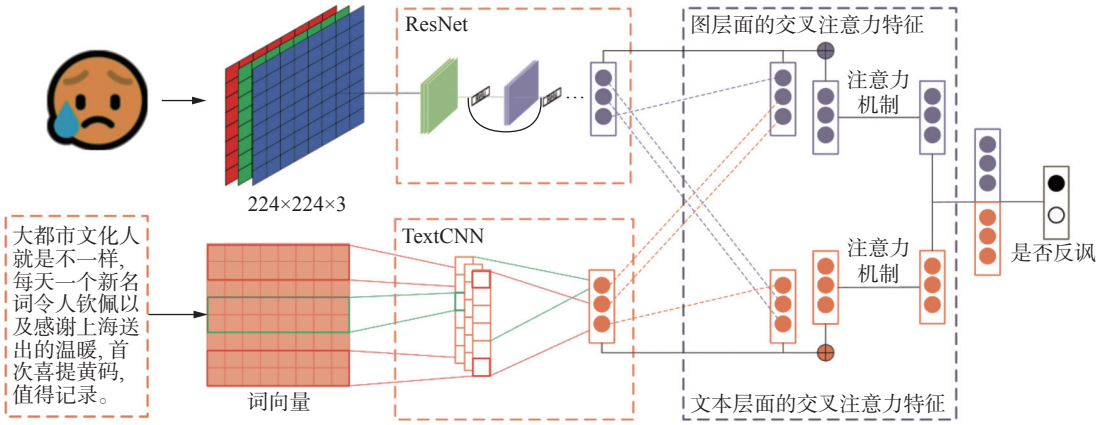

图 2　融合交叉注意力的突发事件多模态反讽识别模型框架

Fig. 2　Framework diagram of Multimodal Chinese sarcasm detection model for emergencies based on cross attention

1.2.1　文本特征提取

利用 Word2vec 模型在海量文本中学习突发事件网民评论语义信息得到文本向量化表示，作为文本模型的输入。突发事件多模态反讽识别数据集中的网民文本评论集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}$，$x_i$表示一条文本评论。通过分词得到的每条评论 $\boldsymbol{x}_i=(\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\cdots,\boldsymbol{x}_{ii},\cdots,\boldsymbol{x}_{ik})$，$x_{i1,ik}$表示第$i$条文本评论的第$k$个单词。每个通过 Word2vec 训练好的单词 $\boldsymbol{x}_{ik}=(x_{ik}^1,x_{ik}^2,\cdots,x_{ik}^m)$，$\boldsymbol{x}_{ik}^m$表示第$k$个单词的$m$维词向量。采用 TextCNN 模型捕捉文本局部特征，解决过拟合问题，训练速度大幅度提高。对特征进行组合和筛选后，获得不同抽象层次的语义信息文本特征。

将得到的词向量输入至卷积层，一次的卷积操作计算公式为

$$c_i=f(\boldsymbol{\omega}\cdot\boldsymbol{x}_{i:i+h-1}+b),\ \ \boldsymbol{\omega}\in\mathbf{R}^{h\times d} \tag{1}$$

式中：$\boldsymbol{\omega}$为卷积核；h为卷积核高度；d为卷积核宽度；$\boldsymbol{x}_{i:i+h-1}$为输入矩阵的第i行到第$i+h-1$行所组成的一个大小为$h\times d$的窗口；b 为偏置参数；f为非线性激活函数；$\boldsymbol{c}=(c_1,c_2,\cdots,c_i,\cdots,c_n)$为经多次卷积操作后得到的特征图。

对特征图 $\boldsymbol{c}$ 使用最大池化操作后得到$\tilde{c}$，其计

算公式为

$$\tilde{\boldsymbol{c}} = \max(\boldsymbol{c}) \tag{2}$$

对不同卷积核经池化之后形成的特征进行拼接得到文本特征向量$\boldsymbol{C}$, 作为后续交叉注意力融合机制模型的输入。

$$\boldsymbol{C} = \text{cat}(\tilde{\boldsymbol{c}}) \tag{3}$$

1.2.2　图像特征提取

ResNet 提出了残差结构, 解决了模型随着网络深度的增加所带来的退化问题。本研究使用 Resnet34 模型提取图片特征, 突发事件多模态反讽识别数据集中的网民文本评论对应的图片集 $P = \{p_1, p_2, \cdots, p_i, \cdots, p_n\}$, P中每张图片 p_i 的初始大小为224×224×3, 图片经若干卷积层至最后一层卷积层 (conv 5) 后, 图片大小变为7×7×512, 再经过平均池化层 (avg pool) 后, 输出大小变为1×1×512, 得到池化后的图像特征 $\boldsymbol{i}$。将图片特征 $\boldsymbol{i}$ 输入至全连接层中得到图像特征 $\boldsymbol{I}$。

$$\boldsymbol{I} = \text{Linear}(\boldsymbol{i}) \tag{4}$$

其中 Linear() 表示全连接层。

2　基于交叉注意力的特征表示

2.1　多模态交叉注意力融合机制

为了更好地学习不同模态间的交互信息, 捕捉到模态间差异点, 基于 self-attention 机制[21], 设计出多模态交叉注意力融合机制。

多模态交叉注意力融合机制由图像层面的交叉注意力机制、文本层面的交叉注意力机制、图像特征融合层注意力机制和文本特征融合层注意力机制组成。多模态交叉注意力融合机制结构, 如图 3 所示。

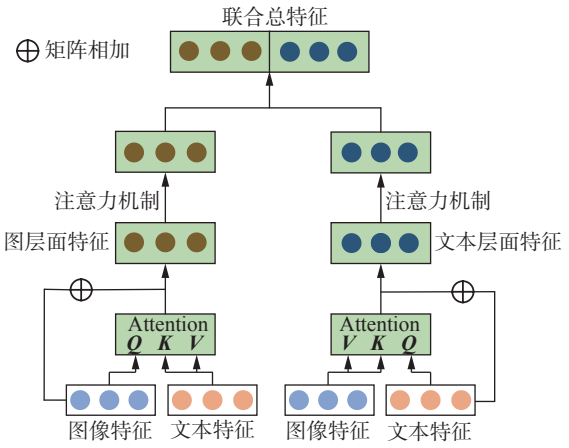

图 3　融合机制结构

Fig. 3　Fusion mechanism structure diagram

2.2　基于交叉注意力的图像层面特征表示

为了获取更丰富的图文不一致信息, 将图像特征作为查询 $\boldsymbol{Q}$(Query), 文本特征作为键 $\boldsymbol{K}$(Key) 和值 $\boldsymbol{V}$(Value), 通过交叉注意力机制得到多头注意力分数, 得到图像层面的模态间注意力特征。受交叉注意力机制必须维度相同的限制, 将文本特征维度和图像特征维度均设置为 1 000 维。

设图像特征$\boldsymbol{I}$作为查询 $\boldsymbol{Q}$, 文本特征$\boldsymbol{C}$作为键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$。通过下式计算得到一次注意力分数。

$$\text{Att}_i(\boldsymbol{I}, \boldsymbol{C}) = \text{softmax}\left(\frac{[\boldsymbol{W}^{\boldsymbol{Q}}\boldsymbol{I}][\boldsymbol{W}^{\boldsymbol{K}}\boldsymbol{C}]^{\text{T}}}{\sqrt{d_k}}\right) \tag{5}$$

式中: $\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$均为可学习参数; d_k为查询 $\boldsymbol{Q}$ 向量的维度大小。多次计算得到多头注意力分数, 将多头机制得到的不同注意力分数运用下式进行堆叠, 得到堆叠后的特征表示$\boldsymbol{A}_T$:

$$\boldsymbol{A}_T = \text{cat}(\text{Att}_i(\boldsymbol{I}, \boldsymbol{C})) \tag{6}$$

该特征表示$\boldsymbol{A}_T$进行线性变换, 得到图像层面交叉注意力机制最终特征表示$\boldsymbol{A}_{\text{TI}}$:

$$\boldsymbol{A}_{\text{TI}} = \text{Linear}(\boldsymbol{A}_T) \tag{7}$$

将图像层面的交叉注意力机制最终特征表示$\boldsymbol{A}_{\text{TI}}$和图像特征$\boldsymbol{I}$相加融合后, 得到特征$\boldsymbol{A}_{\text{TII}}$:

$$\boldsymbol{A}_{\text{TII}} = \boldsymbol{A}_{\text{TI}} + \boldsymbol{I} \tag{8}$$

运用注意力机制对特征$\boldsymbol{A}_{\text{TII}}$进行计算得到注意力分数, 计算得到最终图像总体特征表达$\boldsymbol{A}_{\text{TTI}}$:

$$\boldsymbol{M} = \tanh(\boldsymbol{W}_m \cdot \boldsymbol{A}_{\text{TII}} + b_m) \tag{9}$$

$$\boldsymbol{\alpha} = \text{softmax}\left(\boldsymbol{W}^{\text{T}} \cdot \boldsymbol{M}\right) \tag{10}$$

$$\boldsymbol{A}_{\text{TTI}} = \boldsymbol{A}_{\text{TII}} \cdot \boldsymbol{\alpha}^{\text{T}} \tag{11}$$

式中: $\boldsymbol{W}_m$为注意力机制层的权重矩阵; $\boldsymbol{b}_m$为偏置向量; $\boldsymbol{W}^{\text{T}}$为初始化的随机参数矩阵; $\boldsymbol{\alpha}$为输入分配的权重。

2.3　基于交叉注意力的文本层面特征表示

将文本特征作为查询 $\boldsymbol{Q}$(Query), 图像特征作为键 $\boldsymbol{K}$(Key) 和值 $\boldsymbol{V}$(Value), 使用上一节相同算法可使模型更关注突兀图像区域, 得到文本层面交叉注意力机制最终特征表示$\boldsymbol{A}_{\text{TTC}}$。将得到的文本层面总体特征$\boldsymbol{A}_{\text{TTC}}$和图像层面总体特征$\boldsymbol{A}_{\text{TTI}}$通过下式进行拼接得到最终反讽识别特征表达$\boldsymbol{A}_{\text{TTIC}}$, 将$\boldsymbol{A}_{\text{TTIC}}$输入至分类器中进行分类。

$$\boldsymbol{A}_{\text{TTIC}} = \text{cat}(\boldsymbol{A}_{\text{TTI}}, \boldsymbol{A}_{\text{TTC}}) \tag{12}$$

3　试验与分析

3.1　试验数据集

通过爬取某一地区新冠疫情期间相关话题的微博评论数据, 在对其进行去噪音等一系列处理后, 得到突发公共卫生事件多模态中文反讽识别数据集。人工标注了 5 000 条数据, 经过筛选, 其中得到 1 009 条反讽评论、1 179 条非反讽评论和图片 3 989 张。本研究把反讽识别任务看作是二分类任务, 因此, 在数据标注的过程中将非反

讽标签标注为 0，反讽标签标注为 1。模型的评估指标采用准确率、精准率、召回率和 F_1。

3.2　模型参数

本研究以 9:1 的比例将数据集划分为训练集和验证集。文本词向量为 300 维，TextCNN 卷积核大小设置为 2、3、4 和 5，每种卷积核的数量为 400，学习率为 3×10^{-4}，Dropout 为 0.2，Batch size 为 32，epochs 为 100 轮，若 40 轮训练之后网络的准确率未提升，则停止该组训练，此时得到的准确率是全局最优。

3.3　对比分析

3.3.1　模型精度对比分析

本研究将提出的模型与文本单一模态、图片单一模态和图文双模态 3 类基线模型进行对比，各类基线模型如下。

1) 文本单一模态。

TextCNN[22]：处理短文本较常用的模型，利用 TextCNN 学习文本特征，进行反讽预测。

TextCNN+Att：利用 TextCNN 学习文本特征，对该特征使用注意力机制，进行反讽预测。

Bi-LSTM[23]：常用于分类预测任务中处理文本的模型，利用 Bi-LSTM 学习文本特征，进行反讽预测。

Bi-LSTM+Att：利用 Bi-LSTM 学习文本特征后，同样对特征加入注意力机制建模，预测是否反讽。

BERT[24]：仅使用基础的 BERT 进行预训练，不添加任何模块，预测反讽，表现非常好。

BERT+CNN[25]：使用 BERT 顶层的输出作为特征向量，与 CNN 联合预测反讽。

BERT+RNN[26]：使用 BERT 顶层的输出作为特征向量，与 RNN 联合预测反讽。

2) 图片单一模态。

ResNet：利用 ResNet 全连接层之后的图像向量作为图像特征，输入至 sofxmax 层预测是否反讽。

VGGNet[27]：使用 VGGNet 网络提取图像特征，并预测反讽。

3) 图文多模态。

TextCNN+ResNet：为本研究设计的部分模型，通过简单连接文本特征和图像特征作为总体特征，预测是否反讽。

TextCNN+ResNet+Att：得到文本特征和图像特征作为总体特征后，对该特征加入注意力机制建模并预测反讽。

Bi-LSTM+CNN[15]：利用 Bi-LSTM 学习文本特征，CNN 学习图片特征之后进行多模态特征融合，预测反讽。

BERT+ResNet[12]：利用 BERT 提取文本特征，ResNet 提取图片特征，经过融合后进行反讽预测。

BERT+ResNet+Att[12]：在分别得到文本特征和图片特征之后，使用注意力机制丰富特征后进行反讽预测。

FCAM：本研究提出的融合交叉注意力机制的多模态反讽识别模型。

本研究模型与基线模型对比的详细结果，如表 1 所示。各类基线模型的评价结果柱状图，如图 4 所示。

表 1　模型对比试验结果表

Table 1　Result of contrast test　　%

类别	模型	Acc	P	R	F_1
文本模态	TextCNN	86.76	87.06	86.04	86.39
	TextCNN+Att	84.47	84.75	83.66	84.02
	Bi-LSTM	85.84	87.00	84.65	85.25
	Bi-LSTM+Att	84.02	84.77	82.91	83.41
	BERT	87.61	87.68	86.65	87.06
	BERT+CNN	85.32	84.78	85.36	85.00
	BERT+RNN	83.49	83.24	82.47	82.78
图片模态	ResNet	59.82	58.84	57.14	56.19
	VGGNet	56.16	28.08	50.00	35.96
图文多模态	TextCNN+ResNet	86.30	87.09	85.29	85.80
	TextCNN+ResNet +Att	86.76	87.71	85.70	86.25
	Bi-LSTM+CNN	80.82	82.21	79.27	79.81
	BERT+ResNet	64.38	63.98	62.35	62.17
	BERT+ResNet+Att	60.27	69.41	55.03	47.51
	FCAM	**88.13**	**88.58**	**87.37**	**87.78**

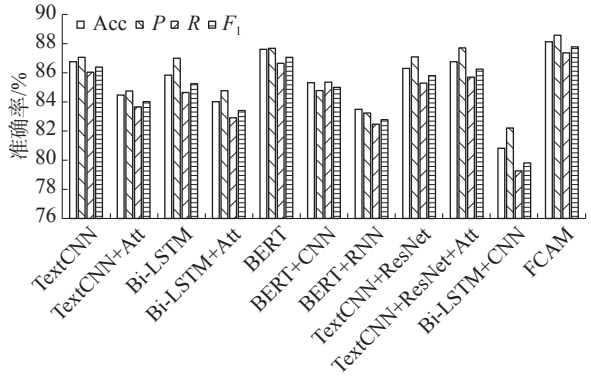

图 4　模型评价指标结果

Fig. 4　Results graph of model evaluation indicators

通过对比试验发现: 1)文本模态对比图片模态发现, 文本模态的表现好于图片模态, 说明通过文本更适合表达反讽意味。2)引入注意力机制的各类基线模型在单模态上表现不佳。3)BERT 模型对文本特征的表征能力较好, 但与其他模型再结合时表现变弱。4)图文多模态方面, 以 BERT 为基础设计出的模型表现不如预期, 分析认为所构造数据集较特殊且隐晦, 导致其预测准确率较低。5)3 类基线模型中, 双模态学习到更充分的特征, 表现更出色。

3.3.2　模型参数对比分析

本研究通过评估模型参数量的大小来反映模型运行效率, 参数量越小其运行效率越高。对比结果如表 2 所示。

表 2　模型参数量对比结果表

Table 2　Comparison results of model parameters

模型	参数量
ResNet34	21.80M
VGGNet11	132.86M
TextCNN+ResNet	24.92M
TextCNN+ResNet+Att	25.05M
BiLSTM+CNN	14.50M
BERT+ResNet	124.85M
BERT+ResNet+Att	126.50M
FCAM	36.14M

通过对比发现: 1)以 BERT 为基础设计的模型的参数量远大于其他模型。2)ResNet 的参数量远小于 VGGNet。3)加入注意力机制后参数量略微增加, 说明在提高模型性能的同时, 也会带来一些额外的计算成本。

3.4　消融试验

本研究针对注意力机制模块和交叉注意力机制模块对最终模型性能的影响进行了进一步评估, 这 2 个模块对模型产生的性能影响, 如表 3 所示。1)-att: 移除模态融合之前所做的注意力机制。2)-ca: 移除交叉注意力机制模块。3)FCAM: 本研究提出的模型。

表 3　消融试验结果

Table 3　Ablation experimental results　　%

方法	Acc	P	R	F_1
-att	86.30	87.09	85.29	85.80
-ca	84.47	85.65	83.21	83.79
FCAM	**88.13**	**88.80**	**87.26**	**87.74**

通过消融试验发现, 将图文特征融合前的注意力机制 att 去除后, 总体性能降低, 说明这时注意力机制模块将最有用的信息作为特征表示, 发挥了较好的作用。当去除交叉注意力模块 ca 后, 同样发现性能降低, 这说明交叉注意力的引入对于反讽识别任务来说是必要的。

3.5　不同维度对性能的影响分析

本研究对经过注意力机制后得到的文本特征 $\boldsymbol{A}_{\mathrm{TTC}}$和图像特征$\boldsymbol{A}_{\mathrm{TTI}}$不同维度的变化对模型性能带来的影响进行了评估。

1)图文特征分别为 1 000 维时的性能分析。

设置文本特征$\boldsymbol{C}$的维度和图像特征$\boldsymbol{I}$的维度均为 1 000 维, 在最终融合之前, 设置文本特征 $\boldsymbol{A}_{\mathrm{TTC}}$和图像特征$\boldsymbol{A}_{\mathrm{TTI}}$的维度是可变的, 对文本特征 $\boldsymbol{A}_{\mathrm{TTC}}$和图像特征$\boldsymbol{A}_{\mathrm{TTI}}$分别从 100~1 000 维上的准确率和 F_1 进行对比试验, 固定模型中所有参数, 共 100 组试验, 对准确率和 F_1 影响的试验结果, 如图 5、图 6 所示。图 5、图 6 中文本维度和图片维度, 其数值皆为 100维升至 1 000 维。跨度为 100, 每张图各 100 个点, 每个点表示设置不同的文本维度与图片维度对应的评估指标数值大小, 通过颜色的深浅表示其数值的高低。准确率和 F_1 越高, 对应点的颜色越亮。

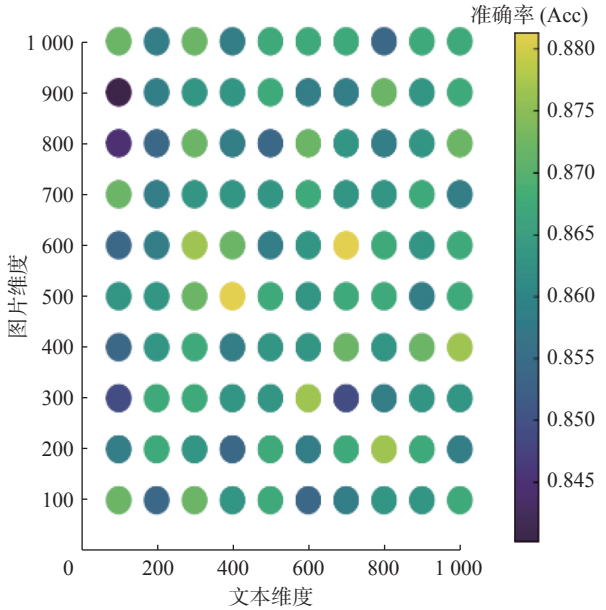

图 5　图文特征 1 000 维度对准确率的影响

Fig. 5　Influence of 1 000 dimensions on accuracy

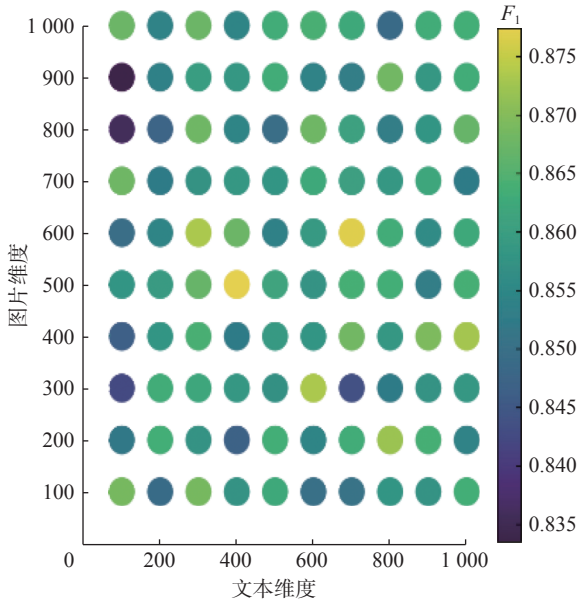

图 6　图文特征 1 000 维度对 F_1 的影响

Fig. 6　Influence of 1 000 dimensions on F_1

通过图 5 和图 6 可以发现，当文本维度为 400 维、图片维度为 500 维时组合形成的最终特征 $\boldsymbol{A}_{\mathrm{TTIC}}$ 与文本维度为 700 维、图像维度为 600 维时组合形成的最终特征 $\boldsymbol{A}_{\mathrm{TTIC}}$ 模型性能最好，此时的准确率达到最大值 0.881 3。此时，前者这对组合对应的 F_1 也达到最大值 0.877 4，后者这对组合的 F_1 为 0.877 0。

2）图文特征分别为 500 维时的性能分析。

同样地，文本特征 $\boldsymbol{C}$ 的维度和图像特征 $\boldsymbol{I}$ 的维度设置为 500 维，从 100 维至 500 维，跨度为 100，做了相同的试验对比，共 25 组试验，对准确率和 F_1 影响的试验结果，如图 7、图 8 所示。

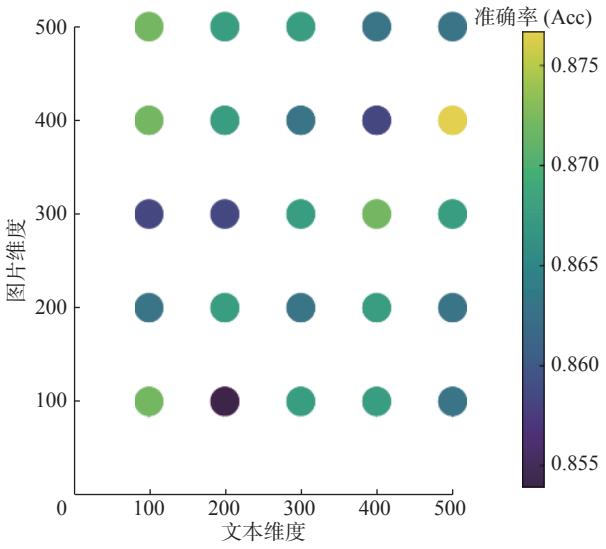

图 7　图文特征 500 维度对准确率的影响

Fig. 7　Influence of 500 dimensions on accuracy

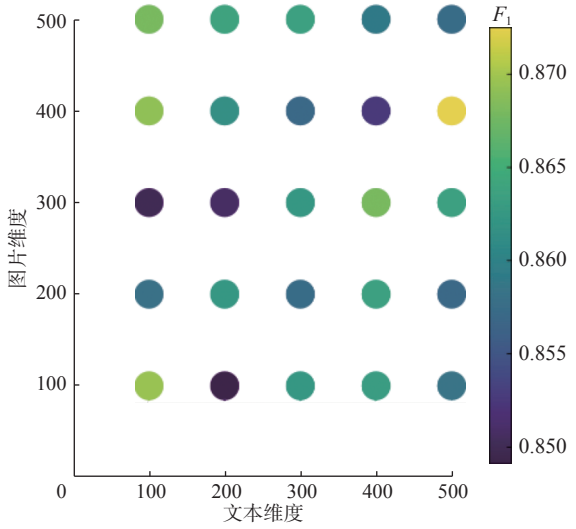

图 8　图文特征 500 维度对 F_1 的影响

Fig. 8　The influence of 500 dimensions on F_1

从图 7 和图 8 可以得出，当文本维度为 500 维、图像维度为 400 维组合形成最终特征 $\boldsymbol{A}_{\mathrm{TTIC}}$ 时，准确率达到最大值 0.876 7，其对应的 F_1 也达到最大值 0.872 5。2 组试验对比发现，与初始维度设置为 500 维度时的准确率相比，1 000 维度时可以得到最好的准确率。

4　结束语

本研究提出了一种融合交叉注意力机制的多模态中文反讽识别模型，通过对爬虫获取某一地区新冠疫情期间相关话题的微博评论数据做处理后，构建出突发事件多模态中文反讽识别数据集。运用 TextCNN 提取文本特征，运用 ResNet 提取图片特征，利用交叉注意力机制表示文本与图片各自引导的特征，最后经过注意力机制融合构成最终的特征表示。该模型在构建的数据集上的表现优于单模态分类模型和基于单纯注意力机制的模型，验证了图文双模态对反讽识别的必要性。未来将从以下几点内容作进一步的研究：1）在双模态基础上加入其他模态对反讽识别进行研究，例如用户评论音频和视频信息；2）加入广义上的上下文信息，改进模型，以提高模型性能，例如用户的个人信息、用户发帖量大小及发帖信息等；3）扩充构建的多模态中文反讽识别数据集；4）尝试将突发事件中的反讽语言特征与情感分析任务相结合，将反讽识别任务划分为多分类问题。

参考文献：

[1]　ZHANG M, ZHANG Y, FU G. Tweet sarcasm detection using deep neural network[C]// International Conference on Computational Linguistics. Osaka: The COLING 2016

Organizing Committee, 2016: 2449−2460.

[2] GHOSH A, VEALE D T. Fracking sarcasm using neural network[C]//Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Stroudsburg: Association for Computational Linguistics, 2016: 161−169.

[3] KHOTIJAH S, TIRTAWANGSA J, SURYANI A A. Using LSTM for context based approach of sarcasm detection in twitter[C]//Proceedings of the 11th International Conference on Advances in Information Technology. New York: ACM, 2020: 1−7.

[4] LOU Chenwei, LIANG Bin, GUI Lin, et al. Affective dependency graph for sarcasm detection[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 1844−1849.

[5] 孙晓, 何家劲, 任福继. 基于多特征融合的混合神经网络模型讽刺语用判别 [J]. 中文信息学报, 2016, 30(6): 215–223.
SUN Xiao, HE Jiajin, REN Fuji. Pragmatic analysis of irony based on hybrid neural network model with multi-feature[J]. Journal of Chinese information processing, 2016, 30(6): 215–223.

[6] 卢欣, 李旸, 王素格. 融合语言特征的卷积神经网络的反讽识别方法 [J]. 中文信息学报, 2019, 33(5): 31–38.
LU Xin, LI Yang, WANG Suge. Linguistic features enhanced convolutional neural networks for irony recognition[J]. Journal of Chinese information processing, 2019, 33(5): 31–38.

[7] 樊小超, 杨亮, 林鸿飞, 等. 基于多语义融合的反讽识别 [J]. 中文信息学报, 2021, 35(6): 103–111.
FAN Xiaochao, YANG Liang, LIN Hongfei, et al. Irony recognition based on multiple semantic fusion[J]. Journal of Chinese information processing, 2021, 35(6): 103–111.

[8] SCHIFANELLA R, DE JUAN P, TETREAULT J, et al. Detecting sarcasm in multimodal social platforms[EB/OL]. (2016−08−08)[2021−01−01]. http://arxiv.org/abs/1608.02289.pdf

[9] SHARMA D K, SINGH B, AGARWAL S, et al. Sarcasm detection over social media platforms using hybrid auto-encoder-based model[J]. Electronics, 2022, 11(18): 2844.

[10] SANGWAN S, AKHTAR M S, BEHERA P, et al. I didn't mean what I wrote! exploring multimodality for sarcasm detection[C]//2020 International Joint Conference on Neural Networks. Glasgow: IEEE, 2020: 1−8.

[11] CAI Yitao, CAI Huiyu, WAN Xiaojun. Multi-modal sarcasm detection in twitter with hierarchical fusion model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 2506−2515.

[12] PAN Hongliang, LIN Zheng, FU Peng, et al. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Online. Stroudsburg: Association for Computational Linguistics, 2020: 1383−1392.

[13] YAO Fanglong, SUN Xian, YU Hongfeng, et al. Mimicking the brain's cognition of sarcasm from multidisciplines for twitter sarcasm detection[J]. IEEE transactions on neural networks and learning systems, 2023, 34(1): 228–242.

[14] GUPTA S, SHAH A, SHAH M, et al. FiLMing multimodal sarcasm detection with attention[C]//International Conference on Neural Information Processing. Cham: Springer, 2021: 178−186.

[15] 张继东, 蒋丽萍. 基于多模态深度学习的旅游评论反讽识别研究 [J]. 情报理论与实践, 2022, 45(7): 158–164.
ZHANG Jidong, JIANG Liping. Research on irony recognition of travel reviews based on multi-modal deep learning[J]. Information studies:theory & application, 2022, 45(7): 158–164.

[16] 刘洋, 马莉莉, 张雯, 等. 基于跨模态深度学习的旅游评论反讽识别 [J]. 数据分析与知识发现, 2022, 6(12): 23–31.
LIU Yang, MA Lili, ZHANG Wen, et al. Detecting sarcasm from travel reviews based on cross-modal deep learning[J]. Data analysis and knowledge discovery, 2022, 6(12): 23–31.

[17] 刘美萍. 重大突发事件网络舆情协同治理机制构建研究 [J]. 求实, 2022(5): 64–76,111.
LIU Meiping. Research on mechanism construction of collaborative governance of online public opinion on major emergencies[J]. Truth seeking, 2022(5): 64–76,111.

[18] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013−01−16)[2021−01−01]. http://arxiv.org/abs/1301.3781.pdf.

[19] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1746–1751.

[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016

IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems 30. Long Beach, USA, 2017: 5998−6008.

[22] WU Suyan, SU Entong, LEI Binyang, et al. TextCNN-based text classification for E-government[C]//2019 6th International Conference on Information Science and Control Engineering. Shanghai: IEEE, 2020: 929−934.

[23] SHARFUDDIN A A, TIHAMI N M, ISLAM S M. A deep recurrent neural network with BiLSTM model for sentiment classification[C]//2018 International Conference on Bangla Speech and Language Processing . Sylhet: IEEE, 2018: 1−4.

[24] DEVLIN J, CHANG MINGWEI , LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: [s. n.], 2019: 4171–4186.

[25] 陆晓蕾, 倪斌. 基于预训练语言模型的 BERT-CNN 多层级专利分类研究 [J]. 中文信息学报, 2021, 35(11): 70–79.
LU Xiaolei, NI Bin. BERT-CNN: a hierarchical patent classifier based on pre-trained language model[J]. Journal of Chinese information processing, 2021, 35(11): 70–79.

[26] LI Zhengguang, LIN Hongfei, SHEN Chen, et al. Cross 2 Self-attentive bidirectional recurrent neural network with BERT for biomedical semantic text similarity[C]//2020 IEEE International Conference on Bioinformatics and Biomedicine. Seoul: IEEE, 2021: 1051−1054.

[27] KAUR T, GANDHI T K. Automated brain image classification based on VGG-16 and transfer learning[C]//2019 International Conference on Information Technology. Bhubaneswar: IEEE, 2020: 94−98.

作者简介：

胡文彬，副教授，博士，中国计算机学会会员，江苏省人工智能学会会员，主要研究方向为社会网络隐私保护、智能信息处理。作为主要成员主持、参与完成省、市级项目 5 项，获省级教学成果奖 1 项，省级教育教学与研究成果奖 1 项。参与撰写专著 1 项，发表学术论文近 20 篇。E-mail：hwb1008@163.com。

陈龙，硕士研究生，主要研究方向为自然语言处理、舆情分析。E-mail：956779521@qq.com。

黄贤波，硕士研究生，主要研究方向为情感分析、舆情管控。E-mail：764157719@qq.com。