第47卷 第12期 计 算 机 工 程 2021年12月
Vol.47 No.12 Computer Engineering December 2021

·人工智能与模式识别· 文章编号：1000-3428(2021)12-0071-07 文献标志码：A 中图分类号：TP181

基于注意力与多模态混合融合的谣言检测方法

陶 霄[1]，朱 焱[1]，李春平[2]

（1.西南交通大学 信息科学与技术学院，成都 611756；2.清华大学 软件学院，北京 100084）

摘 要：社交媒体内容结构具有复杂性，大量虚假信息掺杂在真实内容中，或者在真实图片上配以杜撰的文字内容，导致基于单个模态的方法难以有效检测谣言。提出基于注意力机制与Dempster’s组合规则的混合融合方法。通过新增用户模态，提取文本、视觉和用户3个模态的特征向量，利用注意力机制对词语和视觉进行双向匹配，给予对谣言检测具有更多贡献的词语和视觉神经元更大的权值。在前后期融合均加入注意力机制，实现特征和决策的自动加权，并使用Dempster’s组合规则实现混合融合。在真实的中文Weibo数据集和外文Twitter数据集上的实验结果表明，该方法准确率分别达到97.44%和92.35%，准确率和F1-score指标均高于基准方法和多模态方法。

关键词：谣言检测；多模态融合；注意力机制；混合融合；Dempster’s组合规则

开放科学（资源服务）标志码（OSID）：

中文引用格式：陶霄，朱焱，李春平.基于注意力与多模态混合融合的谣言检测方法[J].计算机工程，2021，47(12)：71-77.

英文引用格式：TAO X，ZHU Y，LI C P.Rumor detection method based on attention and multi-modal hybrid fusion[J].Computer Engineering，2021，47(12)：71-77.

Rumor Detection Method Based on Attention and Multi-Modal Hybrid Fusion

TAO Xiao[1]，ZHU Yan[1]，LI Chunping[2]

（1.School of Information Science and Technology，Southwest Jiaotong University，Chengdu 611756，China；
2.School of Software，Tsinghua University，Beijing 100084，China）

【Abstract】Content on social media is characterized by high structural complexity.Many rumors are mixed with real information，and real pictures are tagged with fabricated description. So it is difficult to detect rumors effectively based on single modal methods.In order to solve the problem，a method is proposed based on hybrid fusion of the attention mechanism and Dempster’s rule of combination.The method adds three kinds of modal feature vectors，including text，vision and user. Then the attention mechanism is utilized to give more weight to words and visual neurons that contribute more to rumor detection，making bidirectional matching between words and vision.The attention mechanism is added to early fusion and late fusion to achieve the automatic weighting of the features and decisions.The hybrid fusion of early fusion and late fusion is implemented by using Dempster’s rules of combination.The experimental results show that the proposed method displays an accuracy of 97.44% on the Chinese data set of Weibo and 92.35% on the data set of Twitter.The accuracy and F1-score of the method are both better than those of the base-line methods and advanced multimodal methods.

【Key words】rumor detection；multi-modal fusion；attention mechanism；hybrid fusion；Dempster’s combination rule

DOI：10. 19678/j. issn. 1000-3428. 0059683

0 概述

随着社交媒体的不断发展，在微博、Twitter等社交平台中转发即时新闻和热点事件已成为一种趋势。基于社交网络平台的即时性、互动性、便捷性等特点，信息以文本、图片等多种模态呈现，并迅速广泛地传播。另一方面，由于社交媒体的上述特性，每天有海量的文字或图片信息被发布，真实信息和谣言混杂在一起，网络平台很难对谣言或虚假新闻进行有效监管，若谣言被用户随意转发，很可能造成严重的社会问题。在2019年12月，社交媒体上有人发文称武汉8名新冠肺炎病人挖墙逃走，引发了一定

基金项目：四川省科技计划项目（2019YFSY0032）。

作者简介：陶 霄（1995—），男，硕士研究生，主研方向为虚假信息检测、多模态融合；朱 焱（通信作者），教授、博士；李春平，副教授、博士。

收稿日期：2020-10-10 **修回日期：**2020-12-01 **E-mail：**yzhu@swjtu. edu.cn

的社会恐慌，而实际上是一名疑似患者打穿夹板墙壁逃出病区，经检查后排除其患新冠肺炎。因此，谣言检测对改善社交媒体的生态和维护社会治安稳定具有重要意义。

早期的谣言检测大多是从相关的文本信息中提取手工特征进行分类，如CASTILLO等[1]初次使用机器学习算法进行谣言的检测，统计了标点符号、单词个数等文本特征，使用决策树进行分类。近年来，随着深度学习的发展，较新的谣言检测研究多数是基于深度学习网络进行分类[2-3]。由于目前社交媒体内容的多样性，用户分享信息更多是以图文共存的形式，因此可以结合多个模态进行学习，有助于提升谣言的分类准确性。由于每个模态的结构不同，选择高效的特征提取方法对各个模态特征进行有效的融合是多模态谣言检测中不可避免的问题。

本文面向文本、图片和用户属性3个模态，提出一种融合注意力机制和混合融合的多模态谣言检测方法。对词和视觉特征进行双向匹配，生成融合注意力机制的文本词特征和视觉特征，基于自注意力机制进行特征的前期融合和后期融合，实现特征和决策的自动加权，在此基础上，设计基于Dempster's组合规则的多模态混合融合方法统一融合方式，以提升谣言检测分类的准确性。

1 相关技术

1.1 谣言检测

早期多数研究致力于基于文本内容和传播结构的谣言检测。ZHOU等[4]提出C-GRU模型，该模型使用卷积神经网络(Convolutional Neural Networks，CNN)自动构建微博的深层文本特征，利用门控循环单元(Gated Recurrent Unit，GRU)挖掘某一事件下相关微博的时间序列信息，该模型结合两者的优点，大幅提高了检测效率。最近也有研究从视觉模态来进行谣言检测，QI等[5]提出一种基于多域视觉的谣言检测模型，通过提取频域特征和像素域特征，并根据BiGRU提取CNN不同阶段的深层视觉特征，将两者融合输入进分类器，该模型优于使用VGG等单一网络进行谣言图像分类。

1.2 注意力机制

2014年，文献[6]在自然语言处理中引入注意力机制，它利用了人类视觉焦点的思想，目前已成为人工智能领域中的重要组成部分。例如，在机器翻译中，若输入序列中的某些单词对所预测的下一个单词更为重要，就需要赋予这些单词更大的权重。注意力机制本质上就是Query和Key-Value之间的映射函数，在多数情况下Key和Value相同，计算公式如式(1)所示：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K})=\text{similarity}(\boldsymbol{Q},\boldsymbol{K})\boldsymbol{V} \tag{1}$$

函数similarity($\boldsymbol{Q}$,$\boldsymbol{K}$)得到的是Query和Key之间进行相关性计算之后的权值。相关性计算有多种方法，例如点积、拼接、加法等。VASWANI等[7]提出完全依赖点积注意力机制的Transformer模型，该模型完全摒弃了以往在NLP中经常使用的递归结构，可以获得全局信息。

1.3 多模态融合

多模态特征融合分为前期融合和后期融合，已经有研究人员对深度学习中的多模态融合技术进行了研究[8-9]，其在语音识别、目标检测、情感识别等多任务中具有非常广泛的应用。

前期融合又称特征级的融合，是将各个模态的特征直接进行拼接，最后将融合后的特征输入进分类器。文献[10-11]均采用前期融合，在提取多模态特征后直接进行拼接。文献[12]提出一种稀疏自编码算法完成特征的有效融合。前期融合将各个模态的信息相融合，可以为分类器提供全面的特征信息。前期融合有2个缺点：1)忽略了模态间的对齐关系，而且容易造成信息冗余；2)因为不同模态的信息可能来自不同的表征空间，进行融合需要将各模态特征映射到统一的空间，这一过程可能会造成信息损失。

后期融合又称决策级的融合，将每个模态的特征输入进独立的分类器，并将每个分类器的决策结果进行集成得到最终的分类结果。GENG等[13]在3个模态上采取了voting策略的后期融合，取得了优于单模型的检测效果。后期融合的2个优点是：1)不需要将每个模态的特征转换成相同的格式，可以根据同模态的特点设计适合模态本身模式的分类器；2)有利于缓解单个分类器的学习过拟合等问题。后期融合也存在一定的缺点，比如决策融合会累加独立分类器内部的误差，进行融合会造成误差的放大。

2 多模态特征提取和混合融合方法

2.1 多模态特征提取

本文设计一种多模态特征提取方法，对于用户发布的微博或推特(本文统称为微博)分别提取文本、视觉和用户3个模态特征。提取文本模态和视觉模态特征后，借鉴文献[10]提出的视觉神经元注意力，采用双向注意力机制从文本和视觉中生成新的词特征和视觉特征。多模态特征提取方法如图1所示。

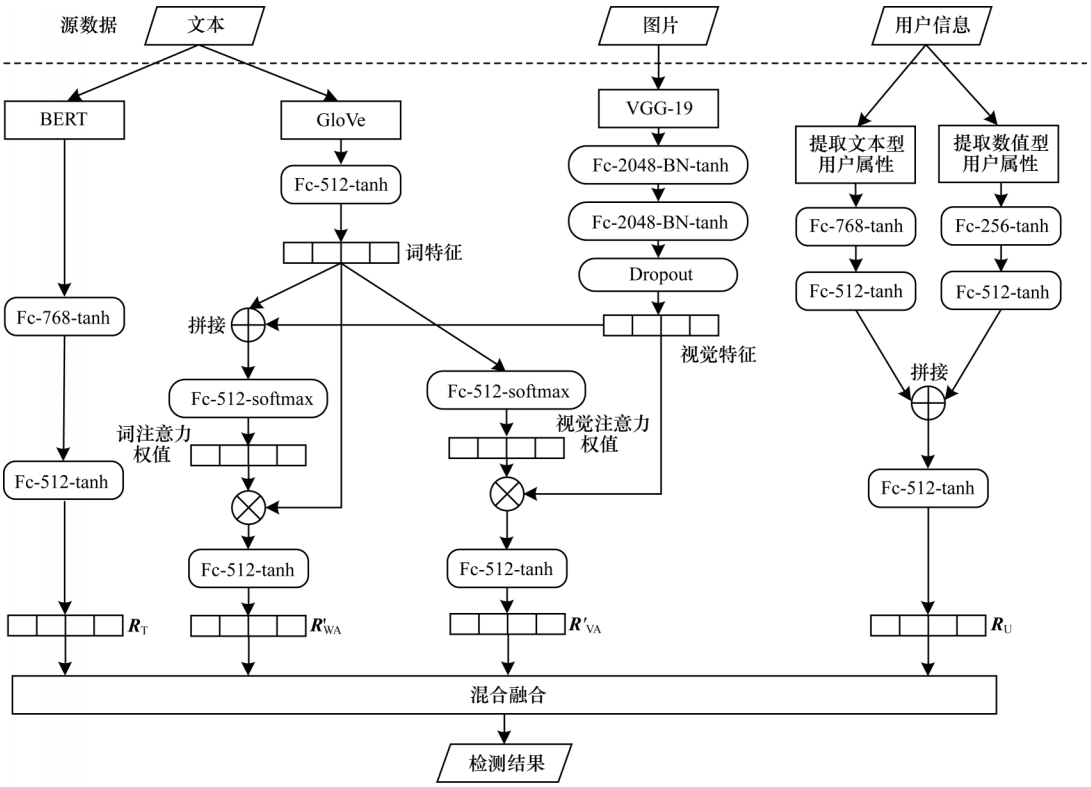

图1　多模态特征提取方法

Fig.1　Method of multi-modal feature extraction

2.1.1　文本特征

针对微博文本，本文分别使用GloVe模型[14]和BERT[15]模型提取了其词特征向量$\boldsymbol{R}_{\mathrm{G}}$和句特征向量$\boldsymbol{R}_{\mathrm{T}}$。GloVe模型与常用的word2vec模型[16]类似，都能得到每个单词的统一编码$\boldsymbol{R}_{\mathrm{W}}$，每条微博就可以用集合$\boldsymbol{R}_{\mathrm{G}}=\{\boldsymbol{R}^1_{\mathrm{W}},\boldsymbol{R}^2_{\mathrm{W}},\cdots,\boldsymbol{R}^n_{\mathrm{W}}\}$（$n$为微博单词个数，经预处理后所有微博文本长度统一）来表示。2种模型没有绝对的优劣，而GloVe更容易并行化，速度更快。BERT是以Transformer为核心的自然语言处理模型，可以真正意义上地双向提取文本上下文信息。词特征用于与视觉特征进行匹配，不再用于提取其词语间上下文信息。

文本中任何特殊字符或词语都有可能作为谣言的特征（比如微博谣言中经常使用大量感叹号）。本文将完整的微博文本输入进BERT模型，使用预训练的BERT模型提取了768维的句向量特征$\boldsymbol{T}$，将$\boldsymbol{T}$输入进2个相连的全连接层。

$$\boldsymbol{R}_{\mathrm{T}}=\tanh(\boldsymbol{W}_{t2}\tanh(\boldsymbol{W}_{t1}\boldsymbol{T}+\boldsymbol{b}_{t1})+\boldsymbol{b}_{t2}) \tag{2}$$

其中：$\boldsymbol{W}_{t1}$、$\boldsymbol{W}_{t2}$是权重矩阵；$\boldsymbol{b}_{t1}$、$\boldsymbol{b}_{t2}$是偏置量；tanh是双全正切激活函数；$\boldsymbol{R}_{\mathrm{T}}$为最终得到的文本句特征向量。

2.1.2　视觉特征

针对微博中的图片，本文使用预训练的VGG-19模型，提取倒数第二层的特征向量$\boldsymbol{V}$，连接2个全连接层。为防止图像的过拟合，在全连接层与激活函数之间添加批归一化（Batch Normalization，BN）层，经过Dropout层，得到视觉特征$\boldsymbol{R}_{\mathrm{V}}$。

$$\boldsymbol{V}'=\boldsymbol{W}_{v2}\tanh(\mathrm{BN}(\boldsymbol{W}_{v1}\boldsymbol{V}+\boldsymbol{b}_{v1}))+\boldsymbol{b}_{v2} \tag{3}$$

$$\boldsymbol{R}_{\mathrm{V}}=\mathrm{Dropout}(\tanh(\mathrm{BN}(\boldsymbol{V}'))) \tag{4}$$

2.1.3　融合注意力机制的词特征和视觉特征方法

在社交媒体中，用户发布的文本和图片大多具有一定的相关性，将图片和词语进行一致性匹配有助于判断用户发布信息的真实性。借鉴人类的注意力机制，要分辨一条信息是否为谣言，会对比文本和图片，找出文本中的关键词和图片中的局部位置是否匹配，如果两者矛盾，该信息更可能为谣言。为实现这个目标，本文使用自注意力机制分别对词语和视觉神经元进行加权，按照其贡献赋予不同的权值。

针对词语，需要重点关注与图片高度匹配或矛盾的词语，使模型自动地找到其中的关联。将$\boldsymbol{R}^i_{\mathrm{W}}$（$i=1,2,\cdots,n$）分别与$\boldsymbol{R}_{\mathrm{V}}$进行串联，经过全连接层后得到$\boldsymbol{A}^i_{\mathrm{W}}$，然后将所有词语生成的$\boldsymbol{A}^i_{\mathrm{W}}$进行拼接，经过softmax层得到每个词语的注意力权重向量$\boldsymbol{A}^{i\prime}_{\mathrm{W}}$，$\boldsymbol{A}^{i\prime}_{\mathrm{W}}$与$\boldsymbol{R}_{\mathrm{W}}$点乘后得到融合视觉注意力的词向量特征$\boldsymbol{R}_{\mathrm{WA}}$，$\boldsymbol{R}_{\mathrm{WA}}$经过全连接层得到最终的词向量特征$\boldsymbol{R}'_{\mathrm{WA}}$。

$$\boldsymbol{A}^i_{\mathrm{W}}=\boldsymbol{W}_{\mathrm{w1}}[\boldsymbol{R}^i_{\mathrm{W}};\boldsymbol{R}_{V}]+\boldsymbol{b}_{\mathrm{W1}} \tag{5}$$

$$\boldsymbol{A}'_{\mathrm{W}}=\mathrm{softmax}(\boldsymbol{A}_{\mathrm{W}}) \tag{6}$$

$$\boldsymbol{R}_{\mathrm{WA}}=\sum_{i=1}^{n}\boldsymbol{R}^i_{\mathrm{W}}\,\boldsymbol{A}^i_{\mathrm{W}}{}' \tag{7}$$

$$\boldsymbol{R}'_{\mathrm{WA}}=\tanh(\boldsymbol{W}_{\mathrm{wa1}}\boldsymbol{R}_{\mathrm{WA}}+\boldsymbol{b}_{\mathrm{wa1}}) \tag{8}$$

针对视觉语义，参考文献[10]中提出的方法，对与词语相关联的视觉神经元给予更大的权值。本文将2.1.1节中得到的词特征向量$\boldsymbol{R}_{\mathrm{W}}$与视觉特征向量$\boldsymbol{R}_{\mathrm{V}}$进行匹配。将$\boldsymbol{R}_{\mathrm{W}}$通过全连接层得到与$\boldsymbol{R}_{V}$相同的维度，经过一个softmax层，归一化得到视觉特征的单词注意力权重$\boldsymbol{A}_{\mathrm{V}}$，将$\boldsymbol{A}_{\mathrm{V}}$与$\boldsymbol{R}_{\mathrm{V}}$点乘得到融合单词注意力的图像特征$\boldsymbol{R}_{\mathrm{VA}}$。将$\boldsymbol{R}_{\mathrm{VA}}$经过2个全连接层得到最终的图像特征$\boldsymbol{R}'_{\mathrm{VA}}$。

$$\boldsymbol{A}_{\mathrm{V}}=\mathrm{softmax}(\boldsymbol{W}_{\mathrm{w1}}\boldsymbol{R}_{\mathrm{W}}+\boldsymbol{b}_{\mathrm{w1}}) \tag{9}$$

$$\boldsymbol{R}_{\mathrm{VA}}=\boldsymbol{A}_{\mathrm{V}}\boldsymbol{R}_{\mathrm{V}} \tag{10}$$

$$\boldsymbol{R}'_{\mathrm{VA}}=\tanh(\boldsymbol{W}_{\mathrm{va1}}\boldsymbol{R}_{\mathrm{VA}}+\boldsymbol{b}_{\mathrm{va1}}) \tag{11}$$

2.1.4 用户特征

发布微博用户的自身特征对谣言的检测起着一定的辅助作用，比如具有拥有粉丝数多、注册天数长等特征的用户发布谣言的可能性更低。由于不同平台、不同数据集的用户特征集结构可能不同，本文设计一种统一的用户特征提取网络。针对粉丝数、关注数等数值型特征进行拼接构成特征集$\{\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_m\}$（$m$为数值型特征个数），通过Zscore标准化后得到数值型特征$\boldsymbol{R}_{\mathrm{F}}$。针对用户描述等文字特征，通过“*”拼接构成一条长文本，利用BERT捕获其语义特征，得到用户文字型特征$\boldsymbol{R}_{\mathrm{L}}$。最后将$\boldsymbol{R}_{\mathrm{F}}$与$\boldsymbol{R}_{\mathrm{L}}$进行拼接，经过一层全连接层得到用户特征$\boldsymbol{R}_{\mathrm{U}}$。

$$\boldsymbol{R}_{\mathrm{U}}=\tanh(\boldsymbol{W}_{\mathrm{u}}[\boldsymbol{R}_{\mathrm{F}};\ \boldsymbol{R}_{\mathrm{L}}]+\boldsymbol{b}_{\mathrm{u}}) \tag{12}$$

2.2 特征混合融合

上文分析了前期融合和后期融合的优、缺点，目前多数研究都是基于单一的前期融合或后期融合。为构建一个具有广泛适用性的特征融合方法，本文设计了特征的混合融合方法，如图2所示。首先将2.1节中得到的4个特征向量分别输入进一个全连接层，得到各个模态的自注意力权重$\boldsymbol{A}_{\mathrm{E}}=\{A_{\mathrm{E}}^1,A_{\mathrm{E}}^2,A_{\mathrm{E}}^3,A_{\mathrm{E}}^4\}$，进行加权求和得到前期融合特征$\boldsymbol{F}_{\mathrm{E}}$，将其输入进二元分类器得到前期融合类别概率$\boldsymbol{F}'_{\mathrm{E}}$。

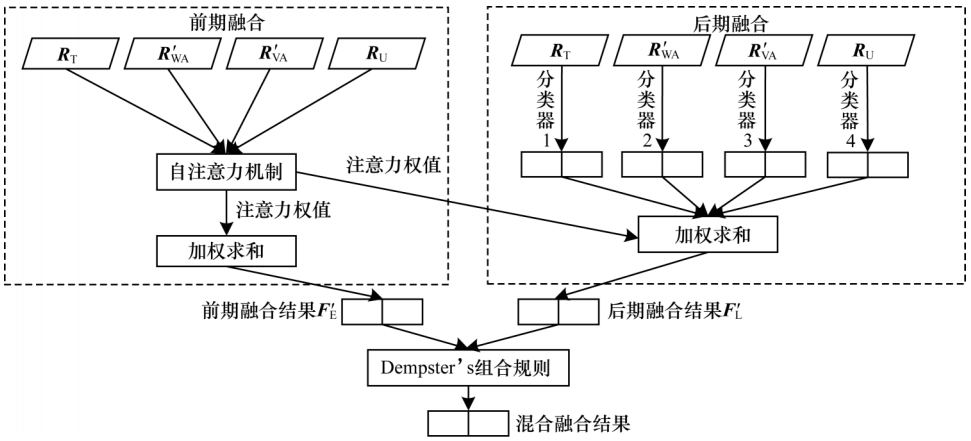

图2 多模态特征混合融合方法

Fig.2 Method of multi-modal feature hybrid fusion

一般的后期融合将各个分类器结果的均值作为最终分类结果，没有考虑各个模态的重要性，造成分类误差。针对该问题，本文提出一种基于特征自注意力权值的后期融合方法。将2.1节中每个特征向量分别输入进各自的分类器得到类别概率集合$\boldsymbol{F}_{\mathrm{L}}=\{F_{\mathrm{L}}^1,F_{\mathrm{L}}^2,F_{\mathrm{L}}^3,F_{\mathrm{L}}^4\}$（集合内元素分别为句特征、融合注意力机制的词特征、融合注意力机制的视觉特征和用户特征的单模态分类器输出结果），利用前期融合过程中的注意力权重$\boldsymbol{A}_{\mathrm{E}}$进行加权求和，得到后期融合概率$\boldsymbol{F}'_{\mathrm{L}}$。

最后，使用Dempster's组合规则对2种融合方式进行结合。Dempster's组合规则是D-S证据理论[17]中的重要组成部分，本文将前期融合和后期融合的输出结果定义为证据理论中的证据），基本概率分配（Basic Probability Assignment，BPA）为类别概率$\boldsymbol{F}'_{\mathrm{E}}$和$\boldsymbol{F}'_{\mathrm{L}}$，利用Dempster组合规则求得最终的谣言和非谣言概率分配$\boldsymbol{m}_d$：

$$\boldsymbol{m}_d=1/(1-K)\sum_{X\cap Y=B}\boldsymbol{m}_{\mathrm{E}}(X)\boldsymbol{m}_{\mathrm{L}}(Y) \tag{13}$$

$$K=\sum_{X\cap Y\neq B}\boldsymbol{m}_{\mathrm{E}}(X)\boldsymbol{m}_{\mathrm{L}}(Y) \tag{14}$$

其中：空间$\boldsymbol{B}\in\{$谣言，非谣言$\}$；$\boldsymbol{m}_{\mathrm{E}}$和$\boldsymbol{m}_{\mathrm{L}}$分别对应前期融合和后期融合证据；$K$代表代表证据之间的冲突程度；$1/(1-K)$为归一化因子。

3 实验与结果分析

3.1 数据集

为评估本文方法在开放环境中对谣言的检测效果，本文使用文献[10]中真实的中文Weibo数据集和外文Twitter数据集，数据集涵盖了中文和外文，覆盖目前社交媒体上各种类型的谣言，能在一定程度

上模拟真实的开放环境。每个数据集均含有文本、对应的图像和粉丝数、关注数、注册天数等用户属性。在训练集和测试集中,谣言和非谣言比例均接近1:1,总体数据量统计如表1所示。

表1　数据量统计

Table 1　Statistics of data volume

数据集	训练集	测试集
Weibo	6 202	1 759
Twitter	11 636	1 490

3.2　实验设置

对于文本词特征,词维度为64,删除词频小于2的词语,考虑所有文本长度和运行效率,每条文本最多取196个单词,长度不足196的文本进行补0。对于视觉特征,取VGG-19倒数第2层4 096维的特征向量,Dropout丢弃率取0.5。对于少量灰度图像,进行通道复制,模拟RGB图像的3个通道。

为减少模型训练时间,在训练整个网络时固定BERT和VGG-19预训练模型的参数,只训练本文提出的网络结构。在后续工作中可以构建丰富的谣言语料库对预训练模型进行微调以达到更好的效果。

模型中隐藏层参数采取glorot_uniform初始化,并通过参数L2正则化防止过拟合,正则化参数取0.001。在训练过程中,训练批次大小设置为128,学习率为0.001,采用Gradient Descent优化器,均根据早停法训练100轮。

3.3　对比实验模型及评价指标

本文对比了4种单模态特征检测方法、3种特征融合方法和2个最新的多模态谣言检测模型,采用准确率(accuracy)和F1值(F1-score)作为实验的评价指标。

Text_BERT:将2.1.1节中提取的句特征向量输入到全连接层中,训练一个谣言检测模型。

Text_Glove_GRU:将Glove提取的64维词特征向量输入到一层GRU,训练一个谣言检测模型。

Image_VGG-19:将2.1.2节中提取的视觉特征向量输入到逻辑回归模型中,得到谣言与非谣言的分类结果。

User-feature:用2.1.4节中提取的用户模态特征进行分类。

att-RNN[10]和EANN[18]:由于本文使用数据集(包括训练集与测试集的划分)与上述数据集相同,因此直接使用了文中的实验结果进行对比。att-RNN融入了注意力机制,生成融合注意力机制的视觉向量;EANN加入了对抗网络以提高模型的准确性。两者均只涉及了文本和视觉2个模态,并且对多模态特征进行了简单拼接。本文考虑模态间的相关性,对视觉和文本进行了双相匹配,增加了用户模态,对多模态特征进行了混合融合。

Early-Fusion1:将3个模态产生的4个特征向量拼接成一个最终的特征向量,并输入到分类器中。

Early-Fusion2:使用2.2节中自注意力机制进行前期融合,得到融合后的特征向量,并输入到分类器中。

Late-fusion:使用各个模态基分类器预测概率的均值作为最终的预测结果。

EWLF(Early Weight Late Fusion):本文提出基于特征自注意力权值的后期融合方法,实现特征的后期融合。

DHF(Dempster Late Fusion):本文提出的谣言检测混合融合方法,通过Dempster组合规则将Early-Fusion2与EWLF产生的2个预测结果进行融合。

3.4　结果分析

表2为3.3节中所有方法模型的实验结果。

表2　各方法实验结果对比

Table 2　Comparison of experimental results for each method

方法	Weibo		Twitter	
	准确率	F1-score	准确率	F1-score
Text_BERT	0.852 2	0.847 6	0.782 6	0.829 8
Text_Glove_GRU	0.654 3	0.667 0	0.617 2	0.650 6
Image_VGG-19	0.618 5	0.604 6	0.593 5	0.588 3
User-feature	0.893 7	0.899 3	0.540 3	0.538 0
att-RNN[10]	0.788 0	0.785 5	0.682 0	0.682 5
EANN[18]	0.827 0	0.829 0	0.715 0	0.719 0
Early-Fusion1	0.943 1	0.941 7	0.877 5	0.910 4
Early-Fusion2	0.973 2	0.972 3	0.896 6	0.924 9
Late-fusion	0.878 3	0.873 7	0.891 9	0.920 9
EWLF(本文方法)	0.925 0	0.921 4	0.904 7	0.931 7
DHF(本文方法)	0.974 4	0.973 3	0.923 5	0.943 7

通过实验结果可以发现,先进的BERT模型的表现较好,仅用BERT提取的文本模态特征用于检测谣言就能达到较高的准确率和F1值,甚至超过了最新的2个多模态谣言检测模型,这主要得益于BERT强大的自注意力机制和Masked LM方法,使其能较好地捕获文本的双向特征。使用Glove+GRU的词嵌入方法的准确率会比BERT低15%~20%,这主要是因为在人工对文本进行预处理时,针对特定任务难以提取出有效的特征。例如在分词中一般需要去除特殊符号和停用词,而对于谣言检测,

某些特殊符号或停用词都可以作为重要特征，谣言会包含更多的“!”“…”等符号，“不仅”“竟然”等词语也常会出现，因此，使用传统的分词方法会对分类结果产生负面影响。尤其在外文Twitter数据集中，语言种类繁多，很难高效地进行分词和词干提取，使用分词方法的准确率相对中文Weibo数据集要更低。

单一视觉模态特征（Image_VGG-19）针对2个数据集的谣言检测准确率较低，仅有50%~60%的准确率。这主要是因为在谣言检测任务中，仅依靠图片难以区分谣言和非谣言，也很容易造成过拟合。尤其在Twitter数据集中，由于不同的图片只有几百张，往往在几轮训练后准确率迅速下降，本文采取了添加Batch Normalization层、添加Dropout层、早停法等多种方法缓解过拟合，但仍很难提升其准确率。

用户模态特征（User-feature）在Twitter数据集中对谣言检测几乎没有作用，而在Weibo数据集中则能较好地区分出谣言与非谣言，这说明在Weibo数据集中谣言与垃圾用户拥有较强的关联性（可能有专门用于发布谣言的垃圾账号）。社交媒体上用户的属性对谣言检测具有不容忽视的贡献，未来的开放谣言检测系统完全有必要将发布该消息用户的特征考虑在内。

对于多模态谣言检测，目前使用最多的是特征拼接（Early-Fusion1）。特征拼接聚集了所有模态的信息，相比任何单模态模型都有一定的提升。而在多模态检测中，有的模态具有很大的权重，而有的模态信息冗余，甚至对检测结果产生一定的误导性。在大量信息聚集的情况下，全连接层权重的隐式调整很难调整各个模态的重要性，使用结合自注意力机制的前期融合（Early-Fusion2），减少特征冗余，提高有用模态的权重，准确率相比特征拼接有2~3个百分点的提升。

均值后期融合（Late-fusion）在Weibo数据集中的检测效果略低于用户单模态检测，因为直接对各个分类器的结果求均值会放大某个弱分类器的误差，同时又减小某个强分类器对正确分类结果的贡献。本文采取的后期融合方法（EWLF）明显优于均值后期融合方法，其中，Twitter数据集的准确率会相比前期融合高1个百分点左右，而Weibo数据集的准确率要低5个百分点左右。这主要因为在Twitter数据集中，上文提到的视觉模态单分类器过拟合对融合结果产生了较大的负面影响，而后期融合能有效缓解某一模态过拟合问题，因此准确率有所提升；针对Weibo数据集，后期融合造成单模态分类器误差的累计和放大，带来的负面影响远大于缓解过拟合对融合结果的贡献，和已经达到了很高准确率的前期融合相比，准确率有所下降。这也说明，前期融合和晚期融合没有一般性的好坏，其效果会随着具体数据集而定。

本文设计的DHF混合融合模型求得前期融合和晚期融合2个证据的联合信度，解决了2种融合方式的选择问题。实验结果表明，DHF在2个数据集上均比早、晚期融合有一定的提升。

3.5 模型参数与结构分析

为进一步研究所提出模型中参数及结构对实验结果的影响，本文在3.2节的基础上，通过修改模型参数或结构进行实验对比。

首先研究模态特征（$\boldsymbol{R}_{\mathrm{T}}$，$\boldsymbol{R}'_{\mathrm{WA}}$，$\boldsymbol{R}'_{\mathrm{VA}}$，$\boldsymbol{R}_{\mathrm{U}}$）的维度对模型准确度的影响，结果如图3所示。在Weibo数据集中，特征维度取128、256或512时都能达到较高的准确率。在Tiwtter数据集中，特征取64、128和256维时模型准确率较高。从图3可以看出，维度过低会造成特征中的信息损失，维度过高有可能造成信息的冗余，两者都可能造成模型准确率的降低。在考虑模型性能的前提下，本文将特征向量的维度定为128。

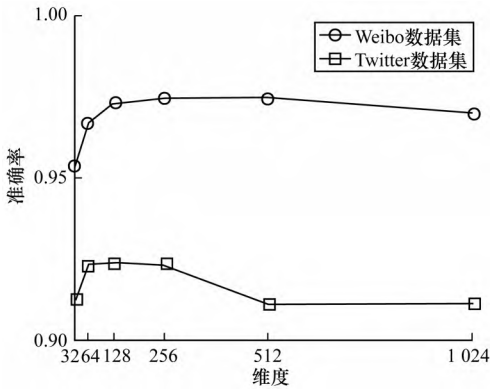

图3 不同特征维度对结果的影响

Fig.3 Influence of different feature dimensions on results

本文对比了现有的注意力计算方式对检测模型的影响，实验结果如表3所示。可以看出，本文通过全连接层和softmax层产生的自注意力权重效果最佳。使用余弦相似度或矩阵点乘的方式虽然算法可解释性较高，但是其计算方式相对固定，而神经网络层具有更多的可自动调节参数，随着对整个网络训练，注意力权值可以得到更准确的修正。

表3 不同注意力计算方式实验结果对比

Table 3 Comparison of experimental results of different attention calculation approaches

方法	Weibo		Twitter	
	准确率	F1-score	准确率	F1-score
Addtive[6]	0.974 4	0.973 3	0.923 0	9.942 1
Scaled Dot-Product[7]	0.972 3	0.973 0	0.917 8	0.939 0
Cosine[19]	0.954 1	0.963 5	0.910 3	0.928 5
Dot-Product[20]	0.972 3	0.972 5	0.917 7	0.939 0
Self-Att(本文)	0.974 4	0.973 3	0.923 5	0.943 7

4　结束语

本文针对真实社交媒体中信息模态的多样性问题，设计一种融合文本、视觉和用户的多模态谣言检测方法。对词语和视觉神经元进行双向注意力加权，在提取文本特征时，除传统的词向量模型外，引入BERT模型生成句向量特征。实验结果表明，混合融合方法能较好地结合特征前期融合和后期融合的优点，并具有良好的可扩展性和普适性，可以作为多模态融合的一般性方法，该方法在2个真实数据集上的检测效果均优于基准方法和2个先进的多模态方法。目前多模态谣言数据集稀缺，本文方法虽然在2个具有代表性的数据集上分类效果较好，但在真实环境中的性能仍需验证，后续需要不断收集各种类型的实时谣言，扩大语料库，提升模型的泛化能力，以验证本文方法在开放环境下的有效性。由于谣言的风格、内容会随着时间的推移发生变化，如何高效检测时事类谣言也是未来工作的一大挑战。

参考文献

[1] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on Twitter [C]//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India: [s. n.], 2011: 675-684.

[2] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks [C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2016: 3818-3824.

[3] BHATT G, SHARMA A, SHARMA S, et al. Combining neural, statistical and external features for fake news identification [C]//Proceedings of International Conference of World Wide Web. Lyon, France: ACM Press, 2018: 1353-1357.

[4] ZHIYUAN Z, YUAN Q, ZHENG L, et al. A C-GRU neural network for rumors detection [C]//Proceedings of the 5th International Conference on Cloud Computing and Intelligence Systems. Nanjing, China: [s. n.], 2018: 704-708.

[5] QI P, CAO J, YANG T, et al. Exploiting multi-domain visual information for fake news detection [C]//Proceedings of 2019 International Conference on Data Mining. Beijing, China: [s. n.], 2019: 518-527.

[6] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [C]// Proceedings of 2015 International Conference on Learning Representations. San Diego, USA: ACM Press, 2015: 1104-1116.

[7] VASWANI A, SHAZEER N. Attention is all you need [C]// Proceedings of Advances in Neural Information Processing Systems. Long Beach, USA: [s. n.], 2017: 5998-6008.

[8] 何俊，张彩庆，李小珍，等. 面向深度学习的多模态融合技术研究综述[J]. 计算机工程，2020，46(5)：1-11.
HE J, ZHANG C Q, LI X Z, et al. Survey of research on multimodal fusion technology for deep learning [J]. Computer Engineering, 2020, 46(5): 1-11. (in Chinese)

[9] 冯耀功，蔡国永. 融合多层语义的跨模态检索[J]. 计算机科学，2019，46(3)：227-233.
FENG Y G, CAI G Y. Cross-modal retrieval fusing multilayer semantics [J]. Computer Science, 2019, 46(3): 227-233. (in Chinese)

[10] JIN Z W, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs [C]//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM Press, 2017: 795-816.

[11] DEVAMANYU H, SRUTHI G. Self-attentive feature-level fusion for multimodal emotion detection [C]//Proceedings of International Conference on Multimedia Information Processing and Retrieval. Miami, USA: ACM Press, 2018: 196-201.

[12] 卢良锋，谢志军，叶宏武. 基于RGB特征与深度特征融合的物体识别算法[J]. 计算机工程，2016，42(5)：186-193.
LU L F, XIE Z J, YE H W. Object recognition algorithm based on RGB feature and depth feature fusing [J]. Computer Engineering, 2016, 42(5): 186-193. (in Chinese)

[13] GENG Y, LIN Z, FU P, et al. Rumor detection on social media: a multi-view model using self-attention mechanism [C]//Proceedings of International Conference on Computational Science. Berlin, Germany: Springer, 2019: 339-352.

[14] PENNINGTON J, SOCHER R, CHRISTOPHER D. GloVe: global vectors for word representation [C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Qatar, Doha: Association for Computational Linguistics, 2014: 1532-1543.

[15] DEVLIN J, CHANG M W. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2020-09-01]. https://arxiv. org/abs/1810. 04805v1.

[16] MIKOLOV T, CHEN K. Efficient estimation of word representations in vector space [C]//Proceedings of International Conference on Learning Representations. New York, USA: ACM Press, 2013: 1-8.

[17] DEMPSTER A P. A generalization of Bayesian inference [J]. Journal of the Royal Statal Society, 1968, 30(2): 205-232.

[18] WANG Y Q, MA F L. EANN: event adversarial neural networks for multi-modal fake news detection [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK: [s. n.], 2018: 849-857.

[19] ALEX G, GREG W, IVO D. Neural Turing machines [EB/OL]. [2020-09-01]. https://arxiv. org/abs/1410. 6247.

[20] THANG L, HIEU P, CHRISTOPHER D. Effective approaches to attention-based neural machine translation [C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: [s. n.], 2015: 1412-1421.

编辑　索书志