基于多任务多模态学习的谣言检测框架

蒋方婷, 梁 刚
(四川大学网络空间安全学院, 成都 610065)

摘 要: 谣言检测是对社交网络上传播的信息内容进行真实性鉴别的任务. 一些研究表明融合多模态信息有助于谣言检测,而现有多模谣言检测方法具有以下问题:(1)只是将处于不同表示空间的单模态特征简单拼接形成多模态表示,没有考虑多模态之间的关系,难以提高模型的预测性能和泛化能力.(2)缺乏对社交网络数据组成结构的细致考虑,只能处理由文本-图像对的社交网络数据,无法处理由多幅图像组成的数据,且当其中一种模态(图像或文本)缺失时模型无法进行预测. 针对上述问题,本文提出了一种多任务多模态谣言检测框架(MMRDF),该框架由3个子网络组成:文本子网络、视觉子网络和融合子网络,通过从单模态数据中提取浅层至深层的单模特征表示,在不同的子空间中产生特征图,丰富模态内特征,并通过复合卷积结构融合生成联合多模态表示,以获得更好的预测性能. 同时该框架可以灵活地处理所有类型的推文(纯文本、纯图像、文本-图像对和多图像文本),并且没有引入造成额外时间延迟的传播结构、响应内容等数据作为输入,可以在推文发布后立即应用于谣言检测,减少辟谣的时间延迟. 在两个真实数据集上的实验结果表明,所提框架明显优于目前最先进的方法,准确率上的提升分别为7.3%和2.9%,并通过消融实验证明了各个模块的有效性.

关键词: 谣言检测; 多模态分析; 表示学习; 多任务学习; 神经网络
中图分类号: TP393 **文献标志码:** A DOI: 10.19907/j.0490-6756.2024.023004

Rumor detection framework based on multitask multi-modal learning

JIANG Fang-Ting, LIANG Gang
(School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Rumor detection is the task of identifying the veracity of the information on social networks. Previous studies have shown that fusing multimodal information can be helpful to rumor detection. However, these approaches have some limitations: (1) simply concatenated unimodal features without considering intermodality relations, resulting in limited improvement in prediction performance and generalization ability. (2) did not carefully consider the composition structure of social network data, assuming it was only composed of image-text pairs and unable to handle multi-image data or missing modalities. To address these issues, we proposed a novel framework called multitask multimodal rumor detection framework (MMRDF), which consists of three sub-networks that generate joint multimodal representation by merging features at different levels and enriching intra-modal features with feature maps from different subspaces. Moreover, the joint multimodal representation is generated by a composite convolutional fusion structure to achieve better prediction

收稿日期: 2023-02-07
基金项目: 自然科学基金联合项目(62162057); 四川省科技厅重点研发项目(2022YFG0182); 教育部地方项目(2020CDZG-18, 2021CDLZ-12, 2021CDZG-11); 达州科技局计划项目(21ZDYF0009); 四川省社会科学重点研究基地——系统科学与企业发展研究中心规划项目
作者简介: 蒋方婷(1998—), 女, 硕士研究生, 研究方向为网络信息内容处理与安全.
通讯作者: 梁刚. E-mail: lianggang@scu.edu.cn

performance. MMRDF is flexible and capable of handling various types of tweets, including pure text, pure image, image-text pairs, and text with multi-images. Additionally, the MMRDF does not require extra time-delaying data such as propagation structures and response content, allowing for immediate application to rumor detection and reducing the time delay in debunking rumors. Experimental results on two real-world datasets demonstrate that our framework outperforms the state-of-the-art methods, achieving an accuracy improvement of 7.3% and 2.9%. Ablation experiments further validate the effectiveness of each module in the proposed framework.

Keywords: Rumor detection; Multi-modal analysis; Representation learning; Multitask learning; Neural Networks

1 引　言

近年来，移动互联网技术的不断更新、媒体内容向移动平台的转变，给当今人们的生活提供了更加便捷的信息交流互动方式[1]. 根据报告（https://datareportal.com/reports/digital-021-global-overview-report）截止2021年初全球有42.2亿社交媒体用户. 随着网络社交媒体的快速发展，社交媒体上的谣言泛滥已经成为一个严重的问题，它可能会对政治、经济和社会稳定造成重大危害. 为了实现自动检测谣言并辟谣，人们做了大量的研究工作，早期的文献[2,3]大多基于手工特征构建表示，并训练浅层分类器来判定谣言的真实性. 为捕捉文本内容中更多的上下文信息和语义信息的变化，研究者们提出了各种深度学习模型[4-6]. 但由于文本表示的局限性，基于文本的检测方法对谣言识别的效果贡献受限. 为了丰富帖子的表示，一些研究开始将帖子中不同格式的数据引入到谣言检测中，如文献[7,8]结合传播结构和文本内容来捕捉评论内容中的语义演化. 然而，这些方法需要长期的观测来获取传播结构和响应信息，导致存在较大的时间延迟，且无法在推文发布后立即揭穿谣言.

图像作为推文发布时的另一种数据格式，能够从视觉角度生动地描述事件发生的情况，补充文本描述所缺失的细节，因此引起了研究者的关注. 由于社交平台的文本长度限制，越来越多的用户选择将图片与文本一起发布来描述事件. 根据文献[9]的数据，新浪微博上有超过一半的推文包含图片. 在线社交网络上的以文本为媒介的交流方式逐渐转向以文本、图像、视频为媒介的多媒体交流方式. 为吸引并误导读者，谣言制造者故意将虚假图片和欺骗性的图片添加到推文或新闻中[10]. 由于谣言利用了视觉信号，一些研究[10-12]发现视觉特征是谣言检测的重要指标，有助于提高谣言检测的性能. 然而，主流的多模态谣言检测研究大多将单峰模态表示的简单连接作为多模态表示. 这种简单的策略可能会导致可能会导致一个模态表示占主导，压制其他模态表示[13,14]. 此外，来自不同模态的信息处于不同嵌入空间中，例如，文本内容遵循时间逻辑，而视觉内容遵循空间逻辑，这意味着特征表示过程需要为每个模态建立不同的模型，在表示融合过程中需要一个融合模型. 表示融合是多模态表示学习中的一个重要问题，旨在联合来自不同模态的表示[15]. 现有的融合方法可以分为早期融合和后期融合. 早期融合在提取单模态特征后立即将其联合为多模态特征表示，而不进行任何学习处理，后期融合试图从单模态特征中单独学习语义信息，并将它们组合成最终的预测结果. 早期融合策略将各模态的低层特征进行融合以学习相关性[16]，当特征仅出现在高层抽象时，可能会造成语义信息等深层特征的不足. 对于后期融合，其通过平均或加权的方式综合考虑子模型的预测结果，而不是多模态表示学习，因此它只能对多模态的线性交互进行建模，而无法学习不同模态特征之间的相关性和互补性.

现有谣言检测研究的另一个局限性是所提出的模型未能处理社交网络上所有类型的推文. 仅考虑推文中的图像和文本内容，社交网络上的推文有4种类型，分别是：(1)纯文本推文；(2)纯图像推文；(3) 文本-图像对（文本推文配一幅图像）；(4)文本推文配多幅图像. 大多数方法[14]只能处理文本-图像对形式的数据，而忽略了纯文本推文和纯图像推文. 此外，以往的研究仅限于处理推文中的单张图像，这种忽略除第一幅图像外的其他图像的数据处理方式，导致视觉信号和图像语义信息不完整.

为了应对上述挑战，本文提出了一种多任务多模态谣言检测框架MMRDF. 该框架包含两个分别用于处理文本和视觉模态的深度神经子网络，以及一个将文本特征和视觉特征投影到同一

多模态空间的融合子网络.每个子网络后附加分类器,通过全局损失反向传播约束以协调每个模态中的表示,从而可以有效地处理各种图文内容.具体而言,我们的工作主要贡献如下:(1)本文提出了一种新的多任务多模态谣言检测框架,该框架由3个子网络组成,即文本模态子网络、视觉模态子网络和融合子网络.为了减少谣言揭穿过程中的时间延迟,该框架没有将传播结构、响应内容等造成额外时间延迟的数据作为输入,以实现在推文发布时即可实现谣言鉴别.(2)针对谣言检测中简单拼接策略导致的联合表示效果不佳的问题,本文提出了一种中间融合方法来融合多模态中不同层次的特征,通过学习单模态不同层次的边缘表示和复杂的跨模态关联特征,进而生成联合表示.从该模型中学习到的联合多模态表示进一步提升了预测性能和泛化能力.(3)由于以往的方法仅限于处理文本-图像对的推文,本文通过从多任务中学习到的三个分类器来处理上述所有类型的推文,并且在其中一种模态缺失时仍然可以运行.此外,为了保持图像语义的完整性、丰富视觉表征,通过数据增强的方法将含有多张图片的推文转换为文本-图像对,而不是默认忽略除第一张图像外的其他图像.(4)在两个真实社交网络数据集上,本文进行了对比实验、定性消融实验、定量消融实验以及减少训练数据实验,结果显示本文模型具有更优的预测性能,比现有基线在准确率上提升了7.3%和2.9%,增强了泛化能力,在少量数据上的训练效果仍能保持领先.

2　相关工作

2.1　基于单模态的自动谣言检测方法

基于文本内容的单模谣言检测方法主要被提出于谣言识别研究的早期,大量研究[2,17,18]根据文本内容中的鉴别特征(如词汇特征,句法特征,主题特征,时序特征等)构建文本表示并训练浅度分类器以鉴别谣言.如2011年,Castillo等[2]从文本内容中提出了基于内容的特征和基于主题的特征,以通过支持向量机、J48决策树、决策规则和贝叶斯网络来评估推文的可信度水平.2015年,Zhao等[19]提出社交网络中不实信息推文发布后会产生与其主题相关的查询推文与质疑推文,并通过这一特征来发现网络中正在传播的不实信息.在上述基于手工构建特征的方法中,有效特征的提取和选择都是手工的,这是一种代价高昂的劳动密集型方法,并且会给模型带来人为误差,导致模型的性能依赖于手工特征工程的质量.因此,一些研究者考虑使用深度学习方法来克服上述问题.Ma等[4]提出了一个基于RNN的模型,该模型学习连续的隐藏表示,捕捉上下文信息的变化以识别谣言.Yu等[20]提出了一种基于卷积神经网络的方法,该方法提取分散在输入序列中的局部-全局特征,并形成重要特征之间的深层交互,以助于识别错误信息.由于用户评论中包含了丰富的信息,如人们的立场和观点,一些研究试图分析推文的内容和用户评论之间的潜在关系.Ma等[21]提出了自底向上和自顶向下的树结构来表示推文和用户回复,并利用递归神经网络进行预测.Rao等[22]采用双层粒度的基于注意力掩码的BERT来处理帖子内容和评论,从而发现与推文文本内容相关的评论.Shu等[23]利用新闻内容和部分评论之间的语义相似性实现可解释的谣言检测模型,通过一个共同注意子网络来捕获Top-k个可解释的句子-评论对.然而这个框架需要长文本的新闻和评论,并不适用于社交媒体中的短文本数据.如前所述,由于社交媒体中文本内容的简短性质和非正式表达,基于文本的模型在谣言识别方面的效果有限.

相较于基于文本内容的检测方法,仅利用图像的单模谣言检测方法较少.Choudhary等[10]采用5个基于CNN的预训练模型和多数投票策略的分类器来发现多媒体特征,从而进行基于图像的假新闻检测.Qi等[24]提出了一种多域视觉神经网络,将频域的物理信息和像素域的语义信息融合在一起,用于区分虚假新闻图像和真实新闻图像.由于不同的人对同一幅图像可能有不同的解读,很难使用图像单独表示发布者的理解或意图.鉴于文本内容在谣言检测中起着重要的作用,推文内容的表示应同时考虑文本内容和视觉内容.

2.2　基于多模态的自动谣言检测方法

尽管单模态的谣言检测方法在一定程度上发挥了作用,但在简短、嘈杂、充斥非正式用语的社交网络数据中挖掘谣言仍然是一项艰巨的任务.幸运的是,社交媒体平台上有着大量丰富的多模态信息,如文本内容,传播结构信息,图像内容等.

图像作为一种多模态数据,与纯文本相比更具有吸引力,更容易引起人们的关注.为了吸引和误导读者,越来越多的虚假图片和欺骗性图片被故意添加到推文或新闻中[10].深度学习神经网络由于其强大的特征提取能力,可以捕捉到复杂的

模式分布，以缓解传统统计特征方法的局限性．一些研究尝试使用深度学习模型提取单峰特征，并将其融合以获得丰富的表征，用于进一步的特定任务．对于文本表示，最近的文献[11,25,26]使用基于Transformer的模型（例如BERT，Encoder-Decoder模型）和基于CNN的模型以更好地捕获语义和上下文信息．对于视觉表示，大多数文献[25,27]使用基于CNN的模型（例如VGG16，VGG19）来获取视觉特征．然而前向研究很少关注发现模态内的相关性和复杂的跨模态关系．多模态视图下谣言检测的大多数研究[11,25,28]都是采用简单的拼接方式将不同嵌入空间中的多模态表示组合在一起，不同模态的信息在不同的嵌入空间中表示，具有不同的噪声结构，例如，文本内容遵循时间逻辑，而视觉内容则遵循空间逻辑，即在特征表示过程中需要针对每种模态训练不同的模型，再将多模特征融合表示为一个联合表示才能进行更好的预测．但现有的这种简单策略可能会导致一种表示压制另一种表示的表现，并且不能在跨模态中提取更多有效的联合边缘特征[29,30]．

现有谣言检测研究的另一个局限性是现有模型未能处理社交网络上所有类型的推文．绝大多数关于多模态谣言检测的研究[28,31]只能处理文本-图像对（由文本信息和推文中的一个图像组成），而忽略了单模态的文本/图像推文，当其中一个模态缺失时就无法工作．此外，主流的研究[32-34]难以处理推文中的多个视觉信号，模型只从推文中的第一张图像中提取视觉特征，忽略了剩余图像，破坏了视觉信息的完整性．处理每张图像将有助于生成更丰富的数据表示，并有助于谣言检测任务．

3　本文模型

3.1　问题定义

给定一个帖子$P=\{T,V\}$包括文本内容T和视觉内容V．文本表示定义为$T=\{w_1,w_2,\ldots,w_n\}$，其中$w_i\in R^{d_w}$是第i个字向量，n是总字数，d_w是字嵌入的维数．图像表示定义为$V=\{v_1,v_2,\ldots,v_m\}$，其中$v_i\in R^{d_v}$是图像的第i个补丁嵌入，m是图像分割块的个数，d_v是视觉嵌入的维数．本文将谣言检测归结为一个二元分类任务，即P分为谣言和非谣言．

如图1所示，MMRDF包括三个组成部分：文本子网络、视觉子网络和融合网络．为了减少谣言检测的时间延迟，该框架将推文中包括文本和图像的原始信息作为输入，以便在推文发布后立即高效工作．该框架对一条推文所附的多张图片进行预处理，将图片横向粘贴到一个白色背景中，形成文本-图像对的数据格式．文本和图像内容分别输入至文本子网络和视觉子网络中，提取单模态特征并进行预测．融合网络以文本初始嵌入、图像初始嵌入和两个单峰子网络的隐藏状态作为输入，合并跨模态相关的单模态表示．融合网络从两单模态中提取浅层至深层的单模特征表示，在不同的子空间中产生特征图，丰富模态内特征，并通过复合卷积结构挖掘多模态表征．达到从不同模态中提取的内容相互补充并增强多模表示的目的，从而有助于谣言检测．

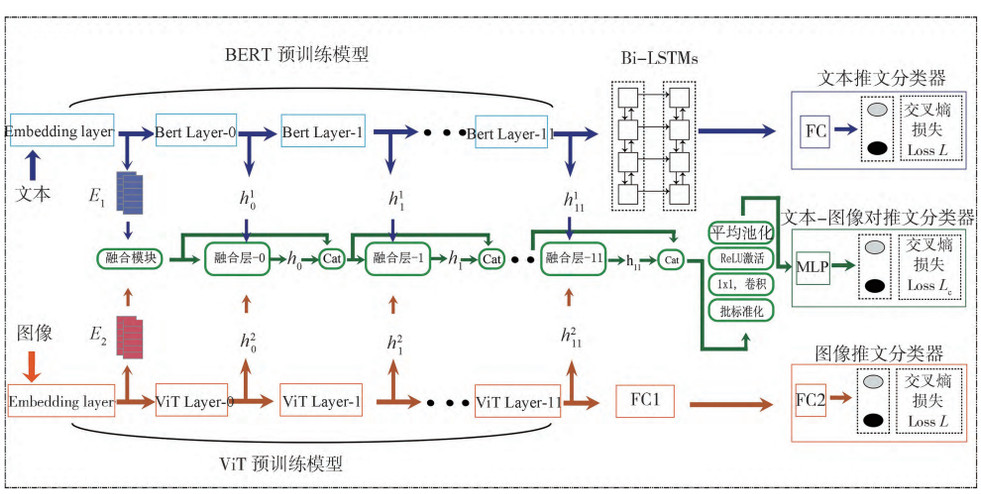


图1　多任务多模态谣言检测框架的网络结构

Fig. 1　Network architecture of the proposed Multitask Multimodal Rumor Detection Framework

3.2　文本子网络

文本子网络处理推文中的文本内容以提取文本表示特征. 部分谣言检测研究利用词袋模型和word2vec嵌入方法来表示文本内容,但这些方法生成的嵌入不能提供单词的位置,并且独立于上下文信息. 预训练BERT模型被提出对大量未标记文本语料库进行预训练,以获得深度双向的语义表示,从而有效地微调各类自然语言处理任务. 在文本子网络中,初始嵌入表示E_1是令牌嵌入、位置嵌入和段嵌入的总和,在早期融合阶段将作为浅层特征输入融合网络中,接着将初始嵌入表示传入12层编码器的预训练BERT模型中,每个编码器层使用多头自注意力机制并行地从不同的表示子空间中提取信息,其中自注意力机制定义为

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}V\right) \tag{1}$$

式(1)中,Q是一个查询向量;K是一个键向量;V是一个值向量;d是向量的维度. 多头注意力如下式进行并行计算后拼接L个特征子空间的词表示,并将多头表示传递给归一化层和含有GELU激活函数的前馈神经网络,得到当前编码器的输出.

$$\mathrm{MultiHead}(Q,K,V)=\mathrm{concate}(head_1,head_2,\dots,head_L)W^O \tag{2}$$

$$head_i=\mathbf{Attention}(QW_i^Q,KW_i^K,VW_i^V),\quad i\in[1,L] \tag{3}$$

式(2)和式(3)中,W_i^Q,W_i^K和W_i^V是查询矩阵、键矩阵和值矩阵的权重;W^O为前馈神经网络的参数. 经过12层编码器的处理,每个单词转换为768维的词向量w_i. 每一层编码器的输出构成模型的隐藏状态$H_1=\{h_0^1,\ h_1^1,\cdots,h_{11}^1\}$. 为了进一步理解上下文并缓解RNN中的长程依赖问题,文本子网络采用堆叠的Bi-LSTM单元来捕获文本特征R_1,将BERT的最后一个隐藏状态h_{11}^1输入到Bi-LSTM中,从文本序列的前向和后向双方向上提取上下文特征. 我们将正向LSTM单元和反向LSTM单元的隐藏层输出$\overrightarrow{s_t^l}$和$\overleftarrow{s_t^l}$连接起来,以得到每个时间点Bi-LSTM的隐藏层输出s_t:

$$s_t=\overrightarrow{s_t^l}\oplus\overleftarrow{s_t^l} \tag{4}$$

式(4)中,$\oplus$是连接操作符;$[s_1,\ s_2,\cdots,s_n]$是文本序列的双向特征表示. 在文本推文分类器中,最后一层的双向特征表示s_t经过密集连接层,通过softmax函数计算预测$[y_{t_1},\ y_{t_2},\cdots,y_{t_N}]$,最后采用交叉熵损失函数计算文本子网络的损失$\mathrm{Loss}_1$,如下式.

$$\mathrm{Loss}_1=-\sum_{i=1}^{N}(y_i*\log(y_{t_i})+(1-y_i)*\log(1-y_{t_i})) \tag{5}$$

式(5)中,N为预测推文总数;y_i为第i个推文的真实标签.

3.3　视觉子网络

在Transformer成为自然语言处理领域的基准架构之时,其在计算机视觉领域的应用也被提出:Vision Transformer(ViT)[35]. 与BERT类似,ViT也由Transformer中的编码器组成. ViT将图像分割为16×16大小的块,然后使用密集层将展平的块映射到具有可训练线性投影的固定块嵌入,并添加位置嵌入,共同组成嵌入向量E_2输入至Transformer编码器中. 在MMRDF的视觉子网络中,我们使用在1400万张图像上预训练得到的ViT模型来获取视觉特征,12层编码器的输出组成了模型的隐藏状态$H_2=\{h_0^2,\ h_1^2,\cdots,h_{11}^2\}$,为了捕获视觉特征$R_2$,视觉子网络采用了一个含有激活函数的密集连接层:

$$R_2=\sigma(W_{FC_1}\cdot h_{11}^2) \tag{6}$$

式(6)中,W_{FC_1}为全连通层的权值矩阵;h_{11}^2为ViT模型的最后一个隐藏状态. 与文本子网络相同,σ为激活函数. 视觉分类器再通过softmax函数计算得到视觉子网络的预测$[y_{v_1},\ y_{v_2},\cdots,y_{v_N}]$,计算交叉熵损失$\mathrm{Loss}_2$:

$$\mathrm{Loss}_2=\sum_{i=1}^{N}(y_i*\log(y_{v_i})+(1-y_i)*\log(1-y_{v_i})) \tag{7}$$

3.4　融合子网络

多模态谣言检测常用方法是分别学习单模态表示,并简单地将它们连接组成多模序列中,即:

$$S=[w_1,w_2,\cdots,w_n,v_1,v_2,\cdots,v_m] \tag{8}$$

文本表示是$S_t=[w_1,w_2,\cdots,w_n]$和可视化表示$S_v=[v_1,\ v_2,\cdots,v_m]$. 然而这种简单的连接策略可能会导致一个模态表示占主导,压制其他模态表示. 同时,由于文本表示和视觉表示属于不同的表示空间,简单的拼接不能将不同的嵌入空间映射到同一个公共空间.

卷积神经网络被认为是一种有效的特征提取器,同时它也被用于多模态数据的融合,如人脸、虹膜、指纹、RGB图像、骨骼数据等. 由于BERT在预训练期间未能关注短文本(如短语),基于CNN的模型可以通过获取文本的局部信息来解决

这个问题．本文构建了一个基于CNN的融合网络，以获取模态内的相关性（如局部表示）和跨模态表示．该融合网络包含12个融合层，依次融合来自文本子网络和视觉子网络的浅层至深层的单模态特征，以产生多模态表征．如图2所示，每个融合层包含两个主要处理块：多模态表示的模态内融合模块和挖掘跨模态表征的模态间关联模块，第i个融合层的输入是前一层融合层的输出O_{i-1}、文本模态的隐藏表示h_i^1和视觉模态的隐藏表示h_i^2．如图3所示，文本的隐藏表示h_i^1和图像的隐藏表示h_i^2首先通过通道堆叠，得到192x768x2的表示，考虑到后续卷积操作的计算复杂度和计算内存需求，卷积层前加入一个全连接层将768维度的嵌入表示降至128维（D=128）．四类具有不同大小的卷积核在输入图上滑动以捕获n-gram特征，由于在Transformer预训练模型产生的表示中，每一行代表一个离散的单词或是被分割的图像区域，为获取完整的表示特征，所有卷积核的长度应与词嵌入的维数相同，高度分别为n=[2,3,4,5]．每种不同大小的卷积核都设置16个，从而在不同的子空间中产生特征图，进而丰富特征．在输入图$A\in R^{s\times D}$上的卷积计算过程如下：

$$y_n^t=\|_{i=1}^{N-n+1}\sigma(w_n^t\bullet x_{i:i+n-1}+b_n^t) \tag{9}$$

$$Y_n=\mathrm{concat}\left(y_n^1,y_n^2,..,y_n^t,\cdots,y_n^{16}\right) \tag{10}$$

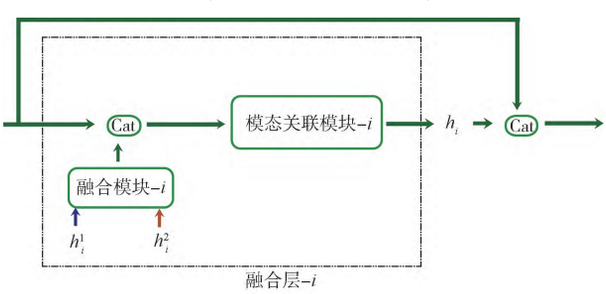


图2　第i个融合层结构

Fig. 2　The network structure of the i-th fusion layer

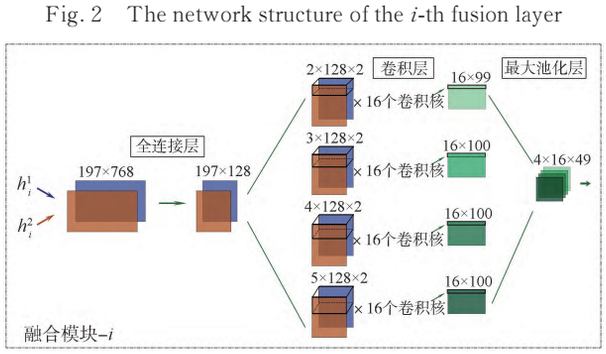


图3　第i个融合模块结构

Fig. 3　The structure of fusion block-i

其中，w_n^t是高度为n的第t个卷积核的权重矩阵；$x_{i:i+n-1}$是输入图A中第i行至第i+n-1行的表示；b_n^t是高度为n的第t个卷积核中的偏置项；σ是ReLU激活函数．特征映射Y_n表示的n-gram特征是在具有相同大小（n=[2,3,4,5]）同类卷积核的输出上进行特征维度级拼接得到的．如图3中右侧矩阵所示，2-gram的特征图Y_2尺寸为16×99，3-gram，4-gram和5-gram的特征图尺寸均为16×100，接着四个特征图通过一个1×3最大池化层，而不是Text-CNN[36]模型中的1-max池化策略．最大池化层从每个特征映射中提取一个标量，这有助于将每个特征映射到固定长度的向量，但会丢失特征映射中的大部分标量．因此，本文中的池化层包含一个步长为2的1×3核，以下采样得到更多信息的特征，同时将特征图映射到相同的大小．池化后的4个特征图在通道层面上拼接在一起得到局部语义信息作为模态内特征．

接着为获取跨模态关联，本文构建复合卷积层（包含1×1卷积层和3×3卷积层）的模态关联模块，沿着通道方向进一步聚合特征．如图4的模态关联模块中，与常见的BN-ReLU-Conv（3×3）网络不同，本文在3×3卷积之前附加1×1卷积构成BN-ReLU-Conv（1×1）-BN-ReLU-Conv（3×3）复合卷积网络结构，1×1卷积层在保留特征相关信息的同时，提供了过滤池化和降维功能，提高了计算效率，使得整个复合卷积网络加强了模态间信息的相互作用．结合图2和图4所示，第i个模态关联模块的输入是第i-1层融合层的输出O_{i-1}和第i层融合模块的输出在通道维度上的拼接，经过批标准化（归一化层），ReLU激活层，1×1卷积层，批标准化（归一化层），ReLU激活层，3×3卷积层得到当前融合层的隐藏表示h_i.

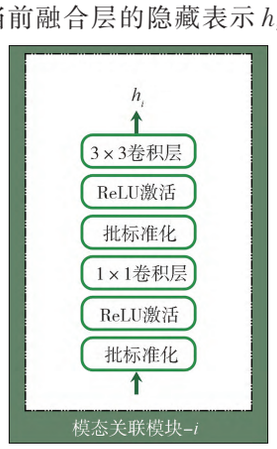


图4　第i个模态关联模块

Fig. 4　The structure of fusion correlation block-i

此外,本文还设计了一个密集连接,将当前融合层的隐藏状态h_i与前一个融合层的输出O_{i-1}连接起来. Huang等[37]提出了一种密集连接的卷积神经网络DenseNet,它将前向层以前馈方式连接到后向层,改善了信息在网络中的流动,缓解了梯度消失问题. 受DenseNet的启发,我们将当前融合层的隐藏状态h_i与前一层融合层的输出O_{i-1}连接起来,以增强特征传播和特征重用. 这样当前融合层输出O_i表示为:

$$O_i = O_{i-1} \oplus h_i \tag{11}$$

由于在第一个融合层中不存在先前的融合隐藏表示,因此将浅层的单模态特征,即文本初始嵌入E_1和视觉初始嵌入E_2输入到额外的融合块中,得到多模态初始表示h_{init}. 第i个融合层的输出O_i表示为

$$O_i = h_{init} \oplus h_0 \oplus h_1 \oplus \cdots \oplus h_{i-1} \oplus h_i \tag{12}$$

如图1所示,通过12个融合层计算的融合表示输入至包含归一化层、1×1卷积层(减少特征图的数量)、激活函数和平均池化层的过渡层. 再通过多层感知器(MLP)进行谣言分类,得到预测$[\widehat{y_1}, \widehat{y_2}, \cdots, \widehat{y_N}]$,最后交叉熵损失计算为$Loss_c$.

$$\mathrm{Loss_c} = -\sum_{I=1}^{N} \left(y_i * \log\left(\widehat{y_i}\right) + \left(1 - y_i\right) * \log\left(1 - \widehat{y_i}\right) \right) \tag{13}$$

结合文本分类器损失和视觉分类器损失,多任务全局损失$Loss$定义为:

$$\mathrm{Loss} = \mathrm{Loss_c} + \mathrm{Loss_1} + \mathrm{Loss_2} \tag{14}$$

包括单峰损失在内的全局损失通过多任务学习来帮助模型泛化. 本模型采用AdamW的随机梯度下降优化单峰子网络和融合网络中的可学习权值和参数.

4　实验与结果

4.1　数据集

目前多模谣言检测数据集较少,如表1所示,仅有三个公开可用的数据集,MediaEval、Weibo[14]和DataFountain(https://www.datafountain.cn/competitions/422/datasets). MediaEval是一个基于Twitter的数据集,包含MediaEval2015[28]和MediaEval2016[38],用于识别Twitter上包括文本和图像的多媒体内容,其包含20657条至少有一张图片的推文,然而在这个数据集中有很多重复使用的图片和重复的文本内容. 删除具有重复文本内容的推文后,只剩下64%的推文. 在这13242条推文中,不同图片的数量为544,这意味着多个不同的推文包含相同的视觉信息. 训练数据的重复会使拟合模型产生偏差,导致模型过拟合. 若删除具有重用图像的推文,此数据集将太小,无法支持所提出框架中的实验. 所以MMRDF在Weibo和DataFountain数据集上进行训练.

表1　数据集描述表

Tab. 1　Statistics of datasets for rumor detection

Statistics	Weibo	dataFountain	MediaEval
Fake post	4749	9494	6695
Real post	4779	9529	6547
Total images	38824	34044	544
Average images/post	4.1	1.8	0.04

Weibo数据集包括4 749条谣言推文和4 779条非谣言推文,每条微博都附有至少一张对应的图片. 在DataFountain数据集中,我们删除非常小或无法显示的图像,删除具有重复文本内容的微博,接着将有关相同事件的微博聚类在一起,并确保来自相同事件的微博不会同时出现在训练集和测试集中. 最终得到9 494个谣言推文和9 529个非谣言推文,共计34 044张图片. 在数据预处理阶段,本文采用一种数据增强的方法,将属于同条微博的多张图片沿水平方向粘贴到一个白色背景的画布上,从而将多图像的微博转换为文本-图像对.

4.2　基线模型

为了验证MMRDF在谣言检测任务中的有效性,本文将其与基于单模态和基于多模态的方法进行了比较. (1) Text-cnn[36]:文本单模态模型,使用预先训练好的文本内容词嵌入权重初始化文本嵌入阶段的参数,并利用Text-cnn网络提取文本特征. 使用具有softmax激活函数的全连接层来预测真实性. (2) MVNN[24]:图像单模态模型,由三个子网络组成:基于CNN的频域子网络,基于Bi-GRU的像素域子网络,融合频域和像素域视觉信息的融合子网络. (3) att-RNN[14]:该模型利用LSTM网络联合表示文本内容和社会上下文,将VGG-19提取的图像特征与文本和社会上下文的联合特征进行融合,并利用注意力机制进行特征对齐. (4) EANN[11]:提出了事件对抗神经网络(EANN)来推导事件不变特征,使用事件鉴别器负责删除所有特定于事件的特性,利用多模态特征提取器从帖子中提取文本和视觉特征学习判别假新闻. 为了进行公平的比较,在DataFountain数

据集上的实验中,我们使用其不包括事件鉴别器的变体EANN-作为对比模型.(5) MVAE[27]:提出一个端到端的多模态变分自编码器来分类假新闻.(6) CLIP[39]:该模型是在4亿文本-图像对上进行预训练的视觉-语言对齐模型,为了将其运用于谣言检测任务,本文微调CLIP将生成的文本表示和图像表示结合起来,随后添加一个带有激活函数的全连接层来进行预测.(7) SpotFake[26]:该模型在假新闻检测方面是最先进的,其利用预先训练好的BERT模型学习文本特征,利用预先训练好的VGG-19学习图像特征.

4.3　实验结果

本节将我们的模型和现有基线模型在两个数据集上进行了比较.表2和表3是各模型的准确率(Accuracy)、精确度(Precision)、召回率(Recall)和*F*1值.MMRDF框架在Weibo数据集和DataFountain数据集上都给出了最先进的结果.这表明我们的框架可以有效地验证谣言,并在不同数据集上具有泛化能力.

表2　Weibo数据集实验结果

Tab. 2　Experiment results on Weibo dataset

Model	Accuracy	Precision	Recall	*F*1
Text-cnn	0.79	0.79	0.79	0.79
MVNN	0.82	0.81	0.81	0.81
CLIP	0.785	0.78	0.78	0.78
att-RNN	0.788	0.79	0.78	0.78
EANN	0.809	0.81	0.81	0.81
MVAE	0.824	0.82	0.82	0.82
Spotfake	0.892	0.874	0.810	0.835
MMRDF	**0.965**	**0.965**	**0.965**	**0.965**

表3　dataFountain数据集实验结果

Tab. 3　Experiment results on dataFountain dataset

Model	Accuracy	Precision	Recall	*F*1
Text-cnn	0.78	0.78	0.78	0.78
MVNN	0.84	0.84	0.84	0.84
EANN-	0.83	0.82	0.82	0.82
Spotfake	0.90	0.90	0.90	0.90
CLIP	0.927	0.927	0.927	0.927
MMRDF	**0.956**	**0.956**	**0.955**	**0.955**

在两个数据集上视觉单模态模型MVNN比文本单模态模型Text-CNN具有更好的性能.文献[35]认为一个图像可被表示为16*16个单词,图片可能比文字更有效地传达内容.虽然单模态方法具有一定效果,但它们的性能比大多数多模态模型差.基线中SpotFake模型是现有谣言检测开源模型之中表现最优的模型,这可能归功于预先训练的BERT模型捕获的潜在语义和上下文内容.从表2中可以看出在Weibo数据集上,我们的模型MMRDF与SpotFake相比在准确率指标上有7.3%的显著提高,在*F*1值指标上有13%的显著提高.在dataFountain数据集上,由于MVEA方法在生成一个巨大的全局词嵌入矩阵时发生了内存错误,且该数据集缺少att-RNN模型需要的社会信息,本文将MMRDF模型与其余五个模型进行比较.视觉单模态模型MVNN的性能不仅优于文本单模态模型还优于多模态模型EANN-.这一结果可能是由于EANN-失去事件鉴别器,未能删除特定于事件的特征.MMRDF模型比视觉-语言对齐模型CLIP在准确度指标和*F*1值指标都提高了2.9%.其主要原因是,与使用简单连接策略进行联合表示多模态的基线模型不同,本文模型MMRDF从不同模态的预训练模型中提取浅层至深层的单模特征表示,在不同的子空间中产生特征图,丰富模态内特征,并通过复合卷积结构挖掘多模态表征,从而提高了模型性能.

表4　消融实验的准确率结果

Tab. 4　Accuracy results on ablation experiments

Model	Textual network	Visual network	Fusion network
MMRDF	0.861	0.896	0.965
MMRDF-text	0.853	Nan	Nan
MMRDF-image	Nan	0.897	Nan
MMRDF-cat	0.815	0.776	0.836
MMRDF-w/o dense	0.841	0.898	0.953

表5　消融实验的*F*1-score结果

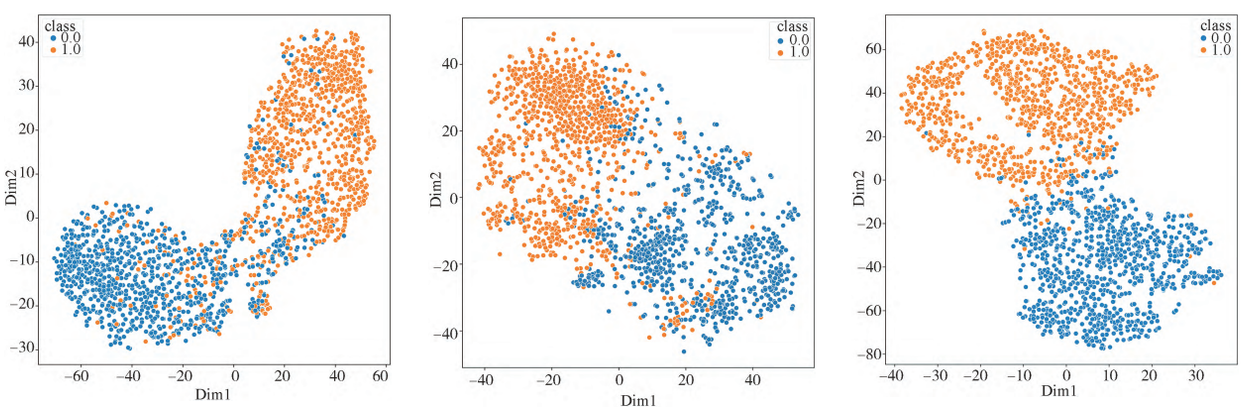
Tab. 5　*F*1-score results on ablation experiments

Model	Textual network	Visual network	Fusion network
MMRDF	0.860	0.896	0.965
MMRDF-text	0.851	Nan	Nan
MMRDF-image	Nan	0.897	Nan
MMRDF-cat	0.813	0.775	0.834
MMRDF-w/o dense	0.838	0.898	0.953

为了验证文本子网络、视觉子网络、融合子网络和其他网络组件的有效性,本文设计了几种MMRDF的变体进行消融实验:MMRDF-text是MMRDF中的单模态文本子网络;MMRDF-image是MMRDF中的单模态视觉子网络;MMRDF-w/

o dense 是去除融合子网络中密集连接的 MMRDF 变体;MMRDF-cat 是用简单连接策略联合文本子网络和视觉子网络表示的变体.

4.3.1　定量消融比较　如表 4 和表 5 所示,在准确率度量上,MMRDF 融合网络比 MMRDF-text 提高了 11.2%,比 MMRDF-image 提高了 6.8%. 在 F1-score 度量上,MMRDF 融合网络比 MMRDF-text 提高了 11.4%,比 MMRDF-image 提高了 6.8%. 结果表明,文本特征和视觉特征在谣言检测中都发挥着重要作用,MMRDF 通过丰富多模态表示有效地提高了多模态融合网络的性能. 为了说明本文融合策略所带来的改进,我们增加了一个实验来比较 MMRDF 和简单拼接融合模型 MMRDF-cat 的性能. 在简单拼接融合模型中,我们将来自文本子网络和视觉子网络的文本表示和视觉表示进行拼接,并将拼接后的表示传递给 MMRDF 融合网络中的同一分类器. 结果显示 MMRDF-cat 中两单模态子网络比 MMRDF 中两单模子网络在准确率上分别降低 4.6% 和 12%,融合网络降低了 13.1%. 说明了 MMRDF 从浅层至深层的多模态特征融合策略比单模态简单拼接的融合方式更有效. 为了验证密集连接在 MMRDF 中的有效性,我们去掉了每个融合层中的密集连接,并将模型命名为 MMRDF-w/o dense. 从表 4 可以看出,失去密集连接后,文本子网络和融合子网络的预测准确率都有所下降,证明密集操作有助于提高信息在融合网络中的流动.



(a) MMRDF-text　(b) MMRDF-image　(c) MMRDF

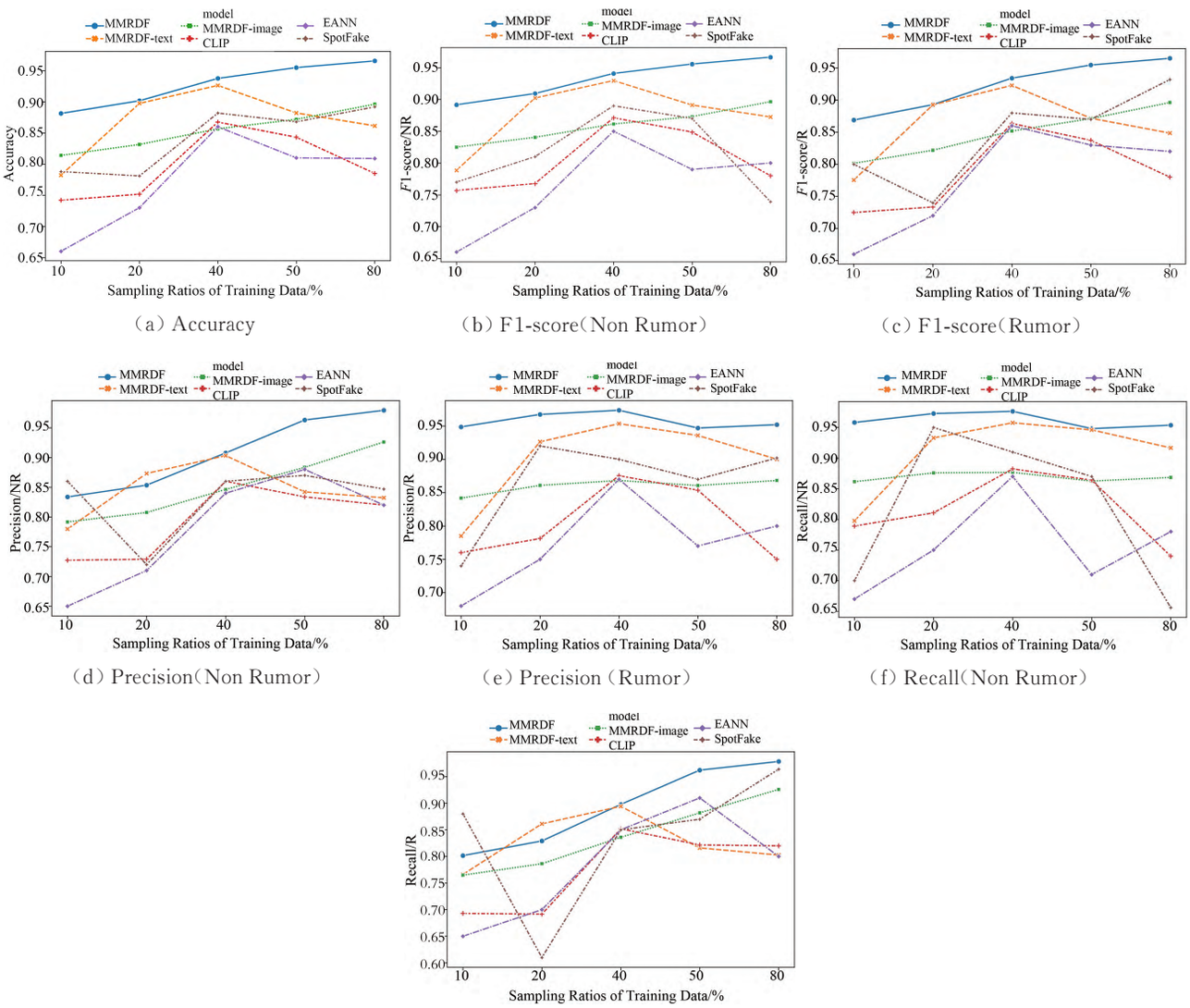
图 5　MMRDF 及其变体模型特征表示的可视化

Fig. 5　Visualizations of learned latent feature representations from MMRDF and variants

4.3.2　定性消融比较　为了进一步分析融合网络的有效性,如图 5 所示,本文应用 t-SNE 算法对测试数据上模型产生的潜在特征表示进行定性可视化. 图 5a 是 MMRDF-text 模型文本特征的可视化结果,图 5b 是 MMRDF-image 模型视觉特征的可视化结果,图 5c 为 MMRDF 模型多模特征的可视化结果,其中蓝色圆点代表谣言,橙色圆点代表非谣言. 在图 5a 和图 5b 中,橙点和蓝点大致位于两个不同的区域,但正样本和负样本的特征表示空间有很多重叠区域. 而在图 5c 中,正样本和负样本的边界较为明显,重叠样本较少. 这表示 MMRDF 模型学习到的特征表示的可分离性大于 MMRDF-text 或 MMRDF-image 模型,从而证明所提出的融合模型通过融合策略学习到更可区分的多模态特征表示,从而获得更好的预测性能.

4.3.3　减少训练数据　为测试模型的泛化性能,本文降低训练集的在总数据集上的占比,分别采用 10%、20%、40%、50% 和 80% 的数据作为训练集,剩余数据作为测试集进行模型训练. 图 6a 显示在准确率指标上,本文采用 40% 的数据训练得到的 MMRDF 模型,测试结果已经优于使用 80% 数据的所有基线模型. 在谣言和非谣言的精确度度量上,本文更关注于非谣言的精确度,即预测为非谣言的样本中实际为非谣言的比例. 因为当谣言被错误预测为非谣言时,将不能及时揭露谣言,造成谣言的进一步扩散传播. 所以需要尽可能提高预测非谣言的精确度,降低被预测为非谣言的谣言比例,如图 6d 中,MMRDF 模型随着训练数据的增多,对非谣言的预测能力相较于其他对比模型

更加稳定地上升，并在50%数据训练结果上优于所有基线模型．在谣言和非谣言的召回率度量上，本文更关注于谣言的召回率，即谣言被预测为谣言的样本占总体谣言样本的比例，需要尽可能提高谣言的召回率，尽可能预测出更多的谣言．如图6g中，MMRDF模型仍然在50%训练数据和80%训练数据上领先于所有基线模型．总体来看，MMRDF模型在图6中7种度量指标上采用少量数据训练学习特征表示，仍能展现较强的预测能力，表明本文模型能够通过少量的数据训练出更具有泛化性的模型．



（a）Accuracy　（b）F1-score（Non Rumor）　（c）F1-score（Rumor）

（d）Precision（Non Rumor）　（e）Precision（Rumor）　（f）Recall（Non Rumor）

（g）Recall（Rumor）

图6　不同训练集测试集比例的实验结果

Fig. 6　Experiment results on different train-test ratio

5　结　论

本文提出了MMRDF多任务多模态谣言检测框架，该框架融合文本和图像模态中浅层至深层的单模态特征表示，在不同的子空间中产生特征图，丰富模态内特征信息，使模型在测试阶段的准确率、精确度、召回率和$F1$值评估指标上的表现都有显著提高；并通过复合卷积结构学习多模态表征，增强泛化能力，使模型在少量数据上的训练效果仍能保持领先．同时该框架可以灵活地处理所有类型的推文（纯文本、纯图像、文本-图像对和多图像文本），并且没有引入造成额外时间延迟的传播结构、响应内容等数据，从而可以在推文发布后立即应用于谣言检测，减少辟谣的时间延迟．本文通过对比实验、定性消融实验、定量消融实验以及减少训练数据实验证明了算法的有效性和泛化能

力.在两个真实数据集上的实验结果表明,本文模型比目前的基线模型在准确率上分别提升7.3%和2.9%.

参考文献:

[1] Peng S Q, Zhou A M, Liao S, *et al*. Forecast method of public opinion evolution based on graph attention network [J]. J Sichuan Univ (Nat Sci Ed), 2022, 59: 013004.[彭思琪, 周安民, 廖珊, 等. 基于图注意力网络的舆情演变预测研究[J]. 四川大学学报(自然科学版), 2022, 59: 013004.]

[2] Castillo C, Mendoza M, Poblete B. Information credibility on twitter [C]//Proceedings of the 20th International Conference Companion on World Wide Web. New York: Association for Computing Mac inery, 2011: 675.

[3] Liu X, Nourbakhsh A, Li Q, *et al*. Real-time rumor debunking on Twitter [C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: Association for Computing Machinery, 2015: 1867.

[4] Ma J, Gao W, Mitra P, *et al*. Detecting rumors from microblogs with recurrent neural networks [EB/OL]. [2023-01-08]. https://www. ijcai. org/Proceedings/16/Papers/537. pdf.

[5] Ajao O, Bhowmik D, Zargari S. Sentiment aware fake news detection on online social networks [C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 2507.

[6] Chen T, Li X, Yin H, *et al*. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection [C]//Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, Melbourne. Australia: [s. n.], 2018: 40.

[7] Zhong L, Cao J, Sheng Q, *et al*. Integrating semantic and structural information with graph convolutional network for controversy detection [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2020: 515.

[8] Lin H, Ma J, Cheng M, *et al*. Rumor detection on twitter with claim-guided hierarchical graph attention networks [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: Dominican Republic, 2021: 10035.

[9] Jin Z, Cao J, Zhang Y, *et al*. Novel visual and statistical image features for microblogs news verification [J]. IEEE T Multimed, 2017, 19: 598.

[10] Choudhary A, Arora A. ImageFake: An ensemble convolution models driven approach for image based fake news detection [C]//Proceedings of the 7th International Conference on Signal Processing and Communication. Noida, India: [s. n.], 2021: 182.

[11] Wang Y, Ma F, Jin Z, *et al*. 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. New York: ACM, 2018: 849.

[12] Shu K, Sliva A, Wang S, *et al*. Fake news detection on social media: a data mining perspective [J]. ACM SIGKDD Explor Newslett, 2017, 19: 22.

[13] Singhal S, Dhawan M, Shah R R, *et al*. Inter-modality discordance for multimodal fake news detection [C]//ACM Multimedia Asia 2021. New York: Association for Computing Machinery, 2021: 1.

[14] Jin Z, 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. New York: ACM, 2017: 795.

[15] Baltrusaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy [J]. IEEE T Pattern Anal, 2018, 41: 423.

[16] Gadzicki K, Khamsehashari R, Zetzsche C. Early vs late fusion in multimodal convolutional neural networks [C]//Proceedings of the 23rd International Conference on Information Fusion. [S. l.]: IEEE, 2020: 1.

[17] Qazvinian V, Rosengren E, Radev D R, *et al*. Rumor has it: Identifying misinformation in microblogs [C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland: [s. n.], 2011: 1589.

[18] Yang F, Liu Y, Yu X, *et al*. Automatic detection of rumor on Sina Weibo [C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. New York: ACM, 2012: 1.

[19] Zhao Z, Resnick P, Mei Q. Enquiring minds: Early detection of rumors in social media from enquiry posts categories and subject descriptors detection problems in social media [C]//Proceedings of the 24th International Conference on World Wide Web. New York: WWW, 2015: 1395.

[20] Yu F, Liu Q, Wu S, *et al*. A convolutional approach

for misinformation identification [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: [s. n.], 2017: 3901.

[21] Ma J, Gao W, Wong K F. Rumor detection on twitter with tree-structured recursive neural networks [C]//Proceedings of the ACL 2018-56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018, 1: 1980.

[22] Rao D, Miao X, Jiang Z, *et al*. STANKER: Stacking network based on level-grained attention-masked bert for rumor detection on social media [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 3347.

[23] Shu K, Cui L, Wang S, *et al*. Defend: Explainable fake news detection [C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2019: 395.

[24] Qi P, Cao J, Yang T, *et al*. Exploiting multi-domain visual information for fake news detection [C]//Proceedings of the IEEE International Conference on Data Mining. [S. l.]: IEEE, 2019: 518.

[25] Giachanou A, Zhang G, Rosso P. Multimodal multi-image fake news detection [C]//Proceedings of the 7th International Conference on Data Science and Advanced Analytics. [S. l.]: IEEE, 2020: 647.

[26] Singhal S, Shah R R, Chakraborty T, *et al*. SpotFake: A multi-modal framework for fake news detection [C]//Proceedings of the 7th Fifth International Conference on Multimedia Big Data (BigMM). Singapore: IEEE, 2019: 39.

[27] Khattar D, Gupta M, Goud J S, *et al*. MvaE: Multimodal variational autoencoder for fake news detection [C]//Proceedings of the World Wide Web Conference. [S. l.]: WWW, 2019: 2915.

[28] Tanwar V, Sharma K. Multi-model fake news detection based on concatenation of visual latent features [C]//Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing. [S. l.]: IEEE, 2020: 1344.

[29] Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge [J]. IEEE T Pattern Anal, 2016, 39: 652.

[30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//Proceedings of the 3rd International Conference on Learning Representations. [S. l.: s. n], 2015: 1.

[31] Zhang H, Fang Q, Qian S, *et al*. Multi-modal knowledge-aware event memory network for social media rumor detection [C]//Proceedings of the 27th ACM International Conference on Multimedia. [S. l.]: ACM, 2019: 1942.

[32] Ma J, Gao W, Wei Z, *et al*. Detect rumors using time series of social context information on microblogging [C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: [s. n.], 2015: 1751.

[33] Jaiswal R, Singh U P, Singh K P. Fake news detection using BERT-VGG19 multimodal variational autoencoder [C]//Proceedings of the 8th Uttar Pradesh Section International Conference on Electrical. Uttarakhand: [s. n.], 2021: 1.

[34] Qian S, Hu J, Fang Q, *et al*. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection [J]. ACM T Multim Comput, 2021, 17: 1.

[35] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at scale [EB/OL]. [2020-10-22]. https://arxiv.org/abs/2010.11929.

[36] Kim Y. Convolutional neural networks for sentence classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: IEEE, 2014: 1746.

[37] Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4700.

[38] Boididou C, Papadopoulos S, Dang-Nguyen D T, *et al*. Verifying multimedia use at MediaEval 2016 [C]//Proceedings of the CEUR Workshop. Hilversum, Netherlands: MediaEval 2016 Workshop, 2016: 4.

[39] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision [C]//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021: 8748.