基于证据增强和局部语义交互的多模态虚假新闻检测

钟 将 高晋鹏 黄敬旺 杨钰铭

(重庆大学计算机学院 重庆 401331)

摘 要 多模态虚假新闻检测的目标是判断新闻中图像和文本内容的真实性。现有虚假新闻检测方法主要存在以下两种问题:(1)现有方法通常从整体语义角度融合图文特征,忽略了图文局部语义之间的联系,导致模型不能有效捕捉图文局部语义差异性;(2)新闻的真实性往往基于可靠的证据和事实,现有方法仅依赖新闻本身的图像和文本难以判断其真假。鉴于此,本研究提出了一种基于证据增强和局部语义交互的多模态虚假新闻检测模型。针对新闻缺乏事实依据的问题,该模型引入证据文本并设计了一种证据增强方法,该方法通过证据文本筛选网络,剔除证据文本中的冗余信息,并利用自注意力模块实现新闻文本的证据增强。同时,为了增强图像语义信息,该模型先从图像块中提取局部特征,再通过双向 GRU 图像语义增强网络,捕获图像序列特征的上下文关系,并利用自注意力模块将图像中嵌入的文字作为新闻背景信息融入图像特征。最后,针对局部语义信息交互问题,该模型使用交叉注意力模块,学习证据增强后的文本特征和语义增强后的图像特征之间的互补信息,增强细粒度的局部语义交互,实现多模态虚假新闻的精确检测。在 Weibo 数据集与 MR2 中英文数据集上的实验结果表明,本文提出的模型性能优于基线方法,在各数据集的准确率上分别提高了 0.8%、2.4%、4.9%。此外,在 IKCEST 第五届"一带一路"国际大数据竞赛中,使用该模型指定的方案从全球 3809 个方案中取得第一的成绩,证实了该方案的有效性。

关键词 多模态虚假新闻检测;证据增强;局部语义交互;证据文本筛选;图像语义增强

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2025.000556

Multimodal Fake News Detection Based on Evidence Enhancement and Local Semantic Interaction

ZHONG Jiang GAO Jin-Peng HUANG Jing-Wang YANG Yu-Ming

(College of Computer Science, Chongqing University, Chongqing 401331)

Abstract With the continuous development of information technology and the widespread popularity of social media in recent years, a large amount of multimodal information is generated on the Internet every day, among which false news is widely exposed and spread through social networks. Effective false news detection methods can reduce the harm caused by false news to society. Current multimodal false news detection methods can obtain more prior semantic knowledge through pre-training models and use the overall semantics between images and texts to guide model decisions. Although these methods can detect false news with large semantic differences between images and texts, they cannot distinguish the local semantic differences between images and texts well. In addition, shallow overall semantic fusion cannot fully tap the prior semantic knowledge of each modality, and deep local semantic interaction is required to effectively capture the semantic differences between different modalities. At the same time, news usually focuses on what is happening at the moment. Without relevant evidence content, the authenticity of news

收稿日期:2024-07-13;在线发布日期:2024-12-13。本课题得到国家自然科学基金(62176029)、重庆市科技创新与应用发展专项基金(CSTB2023TIAD-KPX0064、CSTB2022TIAD-KPX0206)的部分资助。**钟 将**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为大数据分析与挖掘、自然语言处理、云网融合、网络安全等。E-mail: zhongjiang@cqu.edu.cn。**高晋鹏**,硕士研究生,主要研究方向为多模态虚假新闻检测。**黄敬旺**,硕士研究生,主要研究方向为多模态情感分析。**杨钰铭**,博士研究生,主要研究方向为定制化多模态生成。

reports cannot be judged. In fact, external evidence can provide different perspectives and viewpoints to the model to assist the model in judging the credibility of news. Inspired by this, this paper proposes a cross-modal deep local semantic interaction multimodal fake news detection model based on additional evidence. In response to the problem that news lacks factual basis, the model introduces evidence text and designs an evidence enhancement method. The introduction of evidence text verifies the authenticity of news content in multiple ways. In response to the problem of local semantic information interaction, the fine-grained interaction of local semantics of images and texts is achieved through the cross-attention mechanism, which improves the effectiveness and rationality of fake news detection. Specifically, the model first uses a multimodal feature extraction network to represent the semantic information of image data and text data respectively to maintain the structural consistency of image features and text features. Then, the evidence text screening network is used to learn more relevant evidence information and remove redundant information in the evidence text. Then, the long-term dependencies of the image feature sequence are enhanced through the visual semantic residual network to understand the local semantics of the image in a more fine-grained way. Finally, in the cross-modal local semantic information fusion and detection network, the image OCR text is used to enhance image semantic information, the evidence text is used to enhance the news text semantic information, and the cross-attention mechanism is used for cross-modal feature interaction. Through the interaction mechanism between evidence text and local semantics, the model can achieve more accurate fake news detection in an open domain environment. Experimental results on Weibo datasets and MR2 Chinese and English datasets show that the proposed model outperforms the baseline method, and the accuracy on each dataset is improved by 0.8%, 2.4%, and 4.9%, respectively. In addition, in the fifth "Belt and Road" International Big Data Competition of IKCEST, the model stood out among 3809 schemes worldwide and won the first place, which confirmed the effectiveness of the scheme.

Keywords　multimodal fake news detection; evidence enhancement; local semantic interaction; evidence text screening; enhance image semantic

1　引　言

自 Web2.0 时代以来，新闻创造和传播方式发生了变革，用户不仅是信息的消费者，更是信息的创作者[1]。用户在社交媒体上以多种媒体形式发布信息，包括文字、图片、视频、音频等，在加强了信息的传播效果和提升影响力的同时，也导致许多虚假新闻容易迅速传播。虚假信息[2]是指制造者故意误导读者，并能够通过一些其他来源证实其结果为假的信息。在我国，虚假新闻问题严峻，公安部于 2023 年组织全国公安机关进行网络谣言专项整治，截至目前已侦办了 4800 余起虚假新闻案件，查处了 6300 余名违法人员，关停了 3.4 万个违法违规账号。2023 年的众多网络谣言，如“河南三门峡高速货车侧翻众人哄抢物品”“中日友好医院是日本人建的医院”等网络谣言，破坏地域团结，严重影响政府和卫生健康行业的公信力。因此，社交媒体中的虚假新闻检测对增强社会信任和改善信息生态具有重要意义。

按照输入数据的形式，可以将虚假新闻检测方法分为两类：基于社交上下文的方法和基于新闻内容的方法。

(1) 基于社交上下文的虚假新闻检测方法。该方法通过分析社交媒体上的用户行为和新闻在社交网络中的传播路径等信息来识别虚假新闻。现有工作[3-4]通过评估新闻发布用户的行为可信度，对其所发布的新闻进行虚假检测。然而，用户的行为和社交环境会随时间发生动态性变化，从而对检测带来挑战。部分研究[5-6]利用新闻的传播方式构建谣言传播时间线、谣言传播树和谣言传播图，挖掘新闻在社交网络中的时空特征进行虚假新闻检测，但是虚假新闻的传播方式具有多样性，因不同社交网络、平台和用户而异，仅依赖传播信息可能无法全面捕捉虚假新闻。

(2)基于新闻内容的虚假新闻检测方法。该方法利用新闻文本及图像的特征来判断新闻的真实性,该类方法对新闻内容特征的质量有着较高要求。其中以循环神经网络(RNN)为代表[7-8]的单模态检测方法一般使用预训练模型得到文本的词嵌入特征,再利用 RNN 学习文本的语义信息用于分类。对于新闻图片来说,出现了以卷积神经网络(CNN)为代表[9-10]的单模态检测方法,这类方法利用残差网络(ResNet)、视觉几何组(VGG)等基于 CNN 的图像预训练模型提取视觉特征用于分类。然而,单模态方法不能全面地捕获不同数据模式之间的关联,所以出现了结合文本和图像的多模态虚假新闻检测方法。

现有的多模态虚假新闻检测方法主要包括基于多模态融合的方法、基于图文一致性的方法和基于跨模态增强的方法。(1)基于多模态融合[11]的方法先使用预训练模型对图像和文本进行表征,然后拼接图像特征和文本特征用于分类,但是简单的拼接无法充分利用不同模态数据间的互补性;(2)基于图文一致性[12]的方法将虚假新闻检测任务转化为图文匹配任务,利用图文语义相似性检测文本和图像描述的事件是否一致,然而图文特征间存在语义鸿沟,单纯的特征匹配可能无法捕捉到复杂的语义关系;(3)基于跨模态增强的方法[13-14]先单独抽取各模态特征,再使用注意力机制让各模态信息相互增强,但由于文本与图像特征在语义空间上并未对齐,限制了模型跨模态语义融合的能力。

上述方法主要从整体语义层面把握图文相似性,忽略了图文局部语义理解,导致模型不能有效捕获图文之间的差异性,从而影响虚假新闻检测效果。此外,仅依赖新闻内容的方法缺乏外部信息的交叉验证,无法捕捉到与新闻事件相关的背景信息,限制了模型的准确性。图 1 展示了两个虚假新闻案例,图 1(a)中图像描述的是一只猴子抢游客物品,而新闻文本是“这下可以放心去峨眉山了”,二者在语义表达上呈对立状态,可以初步怀疑其新闻的真实性,并且图像光学字符识别(OCR)中的“猴子被判刑”进一步表明其可能是虚假新闻,最后证据文本直接证明了其是谣言。此外,有些虚假新闻的文本和图像描述一致,但是也无法判断其内容的真实性,如图 1(b)所示,图像展示了闪电击中了建筑,文本描述了“华科大宿舍被雷劈了”,图文所表达内容一致,如果不引入外部信息,无法判断新闻的准确性。

新闻文本:这下可以放心去峨眉山了
OCR:峨眉山猴子被判刑了,这下可以放心去了
证据文本:景区辟谣:不属实,“人猴分离”2020年已施行

(a)图文语义差异性

新闻文本:华科大宿舍被雷劈了还好没听到其他坏消息
证据文本:华科大有宿舍遭遇雷击?校方辟谣:天花板松动掉落与打雷无直接关联_手机新浪网

(b)缺乏证据支撑

图 1 虚假新闻示例

针对以上问题,本文提出了一种基于证据增强和局部语义交互的虚假新闻检测方法。在增强图像与文本的语义信息的基础上进行跨模态局部语义信息交互。具体操作流程如下:

该方法可以被分为特征提取、证据文本筛选、跨模态局部语义融合和分类四个阶段。

在特征提取阶段,对于文本信息,该方法使用预训练语言模型提取相关语义信息;对于图像信息,该方法使用多模态预训练模型提取相关图像语义信息,并设计双向门控循环单元(GRU)图像增强网络模块对图像块(patch)进行序列建模,捕获图像 patch 的上下文关系。在证据文本筛选阶段,该方法计算新闻文本和证据文本的注意力评分,最后保留评分大于指定阈值的词元(token)特征向量,从而达到去除冗余信息的效果。在跨模态局部语义融合阶段,将图像特征与图像 OCR 特征拼接形成图像增强特征,文本特征与证据特征拼接形成文本增强特征,然后,利用自注意力机制分别对图像侧和文本侧特征进行语义线索感知,挖掘出图像 OCR 中能补充图像语义的线索,以及能够支持验证虚假新闻成立的证据语义线索。最后,将增强后的图像特征和文本特征送入双向交叉注意力模块,从局部语义层面实现图文模态相互增强。在分类阶段将相互增强的图像特征和文本特征分别用于判断虚假新闻,最终取二者的平均值作为预测结果。

本文的主要贡献包括三个方面:

(1)提出了一种基于证据增强的虚假新闻检测方法,通过引入网页证据提供更多事实依据,并利用

证据筛选网络去除网页证据中的冗余信息，最后将证据文本建模为隐向量参与模型训练。

(2)设计了一种细粒度图文局部语义交互模块，利用视觉语义残差网络强化图像序列特征，再分别使用图像 OCR 与证据文本增强图文语义信息，最后使用交叉注意力模块进行 token 级的细粒度图文局部语义交互。

(3)与现有基线方法相比，该模型展示出了更优性能，并且在 IKCEST 第五届“一带一路”国际大数据竞赛中使用该方案从全球 3809 个竞赛团队中取得第一名，证实了该方案的有效性。

2　相关工作

2.1　基于单模态的虚假新闻检测方法

早期的互联网世界中，文本是新闻内容的重要载体，新闻通常以大量文本夹杂少量图片的形式在互联网中传播。因此，很多研究分别以文本和图片来检测新闻的真实性。

在深度学习引入虚假新闻检测领域之前，研究人员先通过特征工程提取文本特征[15]，如词频、词向量，再使用机器学习算法对特征进行分类或回归从而实现检测虚假新闻的目的。随着算力的提高，Ma 等人[16]首次将深度学习技术应用到虚假新闻检测中，用 RNN 及其变体长短期记忆人工神经网络(LSTM)、GRU 自动学习文本特征，最终将新闻信息表示为隐层向量并进行分类。由于社交媒体中的新闻有时需要结合背景知识才能判断真实性，仅凭新闻本身的文本难以满足检测需求，所以需要向模型引入外部知识。Dun 等人[17]先将文本中的实体与知识图谱中的实体对齐，再利用注意力机制将知识图谱中的实体上下文融入新闻文本信息。Xu 等人[18]利用图结构捕获证据文本内部的长距离依赖关系，从而剔除证据文本中冗余信息。Shu 等人[19]从微观和宏观层面构建了一个新闻分层传播网络，并从传播网络中提取有效特征，提升虚假新闻检测有效性。Liao 等人[20]设计了一个多步证据检索增强的虚假新闻检测框架，通过段落检索和关键证据选择收集现有证据，模拟人在网上验证新闻的搜索行为。Huang 等人[21]探索了如何利用 ChatGPT 中的潜在背景知识增强虚假新闻检测的有效性。

部分虚假新闻通过篡改图像信息混淆事实，基于文本的虚假新闻检测方法无法表示图像特征，因此出现了基于图像的虚假新闻检测方法。Cao 等人[22]将图片特征分为空域特征与频域特征，使用频域信息判断图片伪造痕迹，用空域特征判断图像语义信息是否经过篡改，最终将空域与频域特征拼接起来送入分类器。Singh 等人[23]使用基于 CNN 的图像预训练模型提取视觉特征，并结合文本线索，搭建了一套虚假图片检测框架。尽管这类方法能够检测到部分图像篡改内容，但是互联网上的图像多样且复杂，在真实应用中存在较大局限性。

2.2　基于多模态信息的虚假新闻检测方法

随着社交媒体的普及，利用多模态信息进行虚假新闻检测已成为了一种重要的研究方向。传统的单模态方法无法充分利用丰富的多模态信息来区分真实新闻与虚假新闻，引入多模态信息可以提供更全面的视角和更准确的判断。基于多模态信息的虚假新闻检测方法主要分为多模态融合方式、图文一致性对比方式、跨模态增强方式。

常见的多模态融合方式先使用 VGG 提取图像特征，再使用文本特征提取器对文本进行表征，最后将图像特征与文本特征拼接起来用于分类。Wang 等人[24]使用基于文本的卷积神经网络(Text-CNN)提取文本特征，将图像特征与文本特征拼接在一起作为新闻表示，并设计事件鉴别子任务，通过消除各事件特征之间的差异，学习事件之间共享特征。Singhal 等人[25]使用预训练双向编码器表示模型(BERT)提取文本特征，在不使用辅助任务的情况下极大提高了虚假新闻检测性能，其原因在于预训练语言模型包含更多先验背景知识，能更好挖掘出新闻文本中的语义信息。但这类方法只是将图像特征作为文本特征的补充，并没有充分利用不同模态中的丰富信息。

虚假新闻常用的一种造假方法是在新闻内容插入一些关联性不大的图片用于展示虚假新闻内容，因此可以用图文一致性来判断新闻内容的真假，当图文不一致时则为假新闻。Shang 等人[26]构造对象感知视觉编码器提取新闻图像中的显著对象特征，并利用生成式网络来评估图文特征一致性，从而判断图像是否与文本内容相匹配。Zhong 等人[27]根据图像与文本单词的相关性计算出线索注意力矩阵，从而让模型自适应地融合共享语义特征和非共享语义特征。Xue 等人[28]使用 BERT 提取文本语义特征，ResNet 提取视觉语义特征，通过计算图文整体语义相似度判断图文一致性。这类方法通过比

较不同模态特征之间的整体差异性，缺乏细粒度的具体语义交互。

人们在阅读新闻时通常利用图片理解文字，同时也利用文本理解图片，基于这种启发式思想，一些研究探索利用图像特征帮助模型理解文本特征，以及利用文本特征帮助模型理解图像特征，从而挖掘出多模态数据中的丰富信息，更好地识别虚假新闻。Jin 等人[29]提出了一种带有注意力机制的循环神经网络，融合文本特征和社交上下文特征后，利用线性注意力机制增强图像特征。Wu 等人[30]提出了一种多模态交叉注意力网络，按照人看一眼图像再看一眼文字的阅读方法，使用交叉注意力机制融合多模态特征，学习不同模态之间的依赖关系。Xiao 等人[31]提出了一种基于历史帖子的多模态虚假新闻检测方法，使用自注意力机制融合历史上下文特征来评估帖子的真实性，并使用注意力机制[32]融合各模态特征。这类方法都使用注意力机制让各模态信息相互增强，但由于文本与图像特征在语义空间上并未对齐，限制了模型跨模态语义融合的能力。

2.3 多模态学习

近年来，多模态机器学习在视觉问答[33]、图文检索[34]、图像分类[35]等任务上取得了许多优秀的研究成果，其中以对比语言-图像预训练（CLIP[36]）为基础的模型展现出了优异的性能。CLIP 是一个关联图像和文本的多模态学习模型，通过对图像-文本对的训练学习，能够同时理解视觉内容和语言描述，在多模态表征上体现出了巨大潜力。CLIP 是双编码架构，分别使用图像编码和文本编码器将图文特征嵌入到同一语义空间，再通过对比学习损失函数优化模型，最大化正确图文对的相似性，同时最小化错误图文对的相似性，从而让模型更好地理解图像内容和文本描述之间的关系。

由于对图文数据具备丰富的理解能力和强大表征能力，CLIP 在多个下游任务中发挥重要作用，Rao 等人[37]把 CLIP 应用于稠密预测任务中，通过隐式和显式地利用 CLIP 的先验知识，将图像-文本匹配任务转化为像素-文本匹配任务，实现更细粒度的语义理解。Luo 等人[38]将 CLIP 对图像的理解能力迁移到视频上，分别使用 CLIP 的图像编码器和文本编码器对视频和文本进行编码，然后对视频特征和文本特征做相似度匹配，从而实现视频与文本的相互检索。此外，CLIP 也应用于虚假新闻检测领域，Zhou 等人[39]通过 CLIP 提取新闻的图文特征，并将图文特征串联起来作为多模态特征，引入模态注意力模块来自适应的聚合特征，从而得到新闻最终的特征表示。但是仅通过线性加权融合各模态特征会忽略局部语义信息，导致模型难以充分挖掘图文之间的互补信息。

针对当前研究方法的不足，本文提出一种基于证据增强和局部语义理解的多模态虚假新闻检测方法。该方法利用证据文本与图像 OCR 增强文本语义信息与图像语义信息，并对图文语义信息进行深度交互，从而提升虚假新闻检测效果。

3 虚假新闻检测方法

本文提出了一种基于证据增强和局部语义理解的多模态虚假新闻检测方法。模型框架如图 2 所示，检测方法主要包括 4 个部分：(1) 多模态特征提取网络，先对新闻内容进行预处理得到图像 OCR

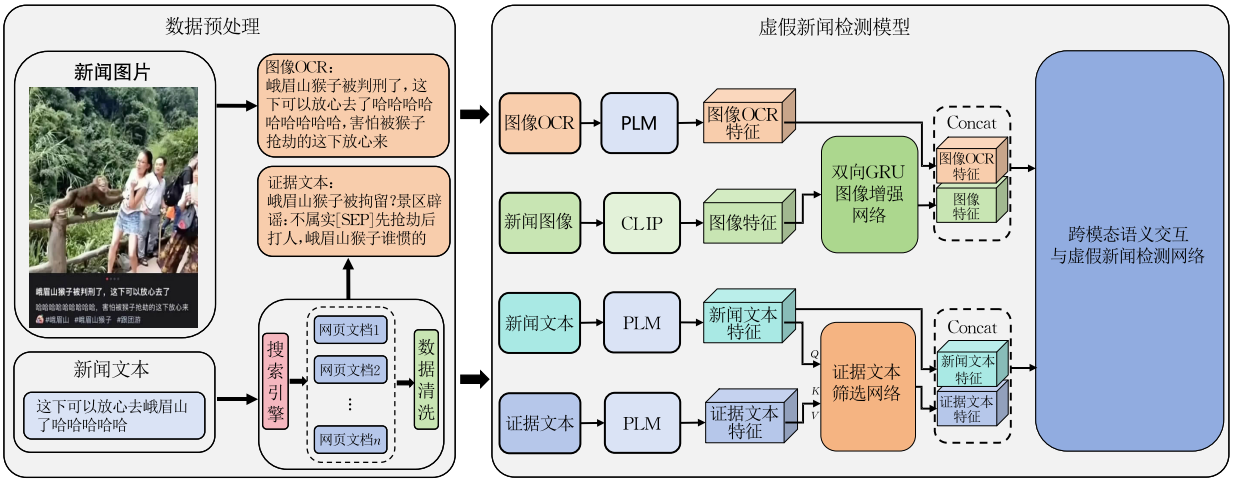


图 2 基于证据增强和局部语义交互的多模态虚假新闻检测框架图

数据和证据文本数据，再分别利用文本预训练模型和图像预训练模型对多模态数据表征得到图像特征 $\boldsymbol{V}$、新闻文本特征 $\boldsymbol{D}$、图像 OCR 特征 $\boldsymbol{O}$ 以及证据文本特征 $\boldsymbol{N}$；(2) 双向 GRU 图像增强网络，利用门控神经网络单元对图像特征进行序列建模，增强模型对于图像特征的语义理解能力，并将图像特征映射到文本语义空间对齐文本特征；(3) 证据文本筛选网络，通过注意力机制计算证据文本对于新闻文本的特征重要性权重，并设置筛选阈值，保留特征重要性大于阈值的证据 token，从而剔除证据文本中的冗余信息；(4) 跨模态语义交互与虚假新闻检测网络，先通过自注意力机制融合图像特征与图像 OCR 特征得到图像侧特征，同理，融合新闻文本特征与证据文本特征得到文本侧特征。再利用交叉注意力机制对图像侧特征和文本侧特征进行跨模态增强；下面对其进行详细介绍。

3.1　多模态特征提取网络

3.1.1　数据预处理

在虚假新闻检测中，充分挖掘多模态新闻内容中的信息是提高检测效果的关键。除了新闻内容中原始的图片 $\boldsymbol{P}$ 与文本数据之外，在新闻图片之中嵌入的文本也很重要，它通常包含图像的背景信息与新闻的细节信息，本文使用 PaddleOCR[40] 提取图像中的嵌入文本 OCR，记为 $\boldsymbol{W}_O=[w_1,w_2,\cdots,w_{m_O}]$，其过程如下：

$$\boldsymbol{W}_O=[w_1,w_2,\cdots,w_{m_O}]=\mathrm{PaddleOCR}(\boldsymbol{P}) \quad (1)$$

其中，$\boldsymbol{P}$ 表示新闻图片，m_O 表示 OCR 文本的长度。

此外，仅依赖新闻内容本身，缺乏外部信息的交叉验证，容易受到单一信息源的误导，无法有效验证新闻的真实性。所以在预处理阶段，本文通过基于规则的方法对检索到的外部网页文档进行初步数据清洗，并提取出网页文档中的标题与摘要作为证据文本，最后拼接所有网页文档中的证据文本将其作为相关证据，增强对新闻真实性的判断，记为 $\boldsymbol{W}_N=[w_1,w_2,\cdots,w_{m_N}]$，其中 m_N 表示证据文本的长度。

3.1.2　多模态特征抽取

该网络模型需要输入图像和文本两种模态的数据。对于图像数据本文使用 CLIP 的图像编码器 vision transformer 对图像进行表征，首先，一张图片会被分为 n 个 patch 并加上一个分类头，记为 $\boldsymbol{P}=[cls,p_1,p_2,\cdots,p_n]$，可类比为 $n+1$ 个 token。然后将所有 token 输入图像编码器，得到带有语义信息的图像特征 $\boldsymbol{V}$，具体过程如下：

$$\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{n+1}]=\mathrm{CLIP}(\boldsymbol{P}) \quad (2)$$

其中，$\boldsymbol{v}_i$ 是 CLIP 图像编码器中对应 patch 的输出层隐藏状态。

对于文本来说，除了表征原始新闻文本 $\boldsymbol{W}_D=[w_1,w_2,\cdots,w_{m_D}]$ 之外，还需要对预处理得到的图像 OCR 与证据文本进行编码，本文使用预训练语言模型(PLM)对文本进行表征，其过程如下：

$$\begin{aligned}\boldsymbol{D}&=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_{w_D}]=\mathrm{PLM}(\boldsymbol{W}_D),\\ \boldsymbol{O}&=[\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_{w_O}]=\mathrm{PLM}(\boldsymbol{W}_O),\\ \boldsymbol{N}&=[\boldsymbol{n}_1,\boldsymbol{n}_2,\cdots,\boldsymbol{n}_{w_N}]=\mathrm{PLM}(\boldsymbol{W}_N)\end{aligned} \quad (3)$$

其中，$\boldsymbol{D}$ 表示新闻文本特征，$\boldsymbol{O}$ 表示图像 OCR 特征，$\boldsymbol{N}$ 表示证据文本特征，$\boldsymbol{d}_i$，$\boldsymbol{o}_i$，$\boldsymbol{n}_i$ 是文本编码器中对应 token 的输出层隐藏状态。

3.2　双向 GRU 图像增强网络

由于 CLIP 模型通过对比学习的方式将图像与文本的分类头进行匹配，其训练过程中主要关注的是图像的整体语义信息，因此在局部细节的捕捉上存在一定不足。为了弥补这一缺陷，本文设计了双向 GRU 图像增强网络模块对图像的 patch 序列进行处理，如图 3 所示。双向 GRU(BiGRU)的优势在于其能够捕捉序列数据中的上下文信息，它不仅能从每个 patch 中提取出有效的局部语义信息，还能通过其门控机制与顺序处理能力，捕捉不同 patch 之间的时序依赖和上下文联系，使得局部细节与全局语义相互补充，从而提升图像特征的语义表达能力。

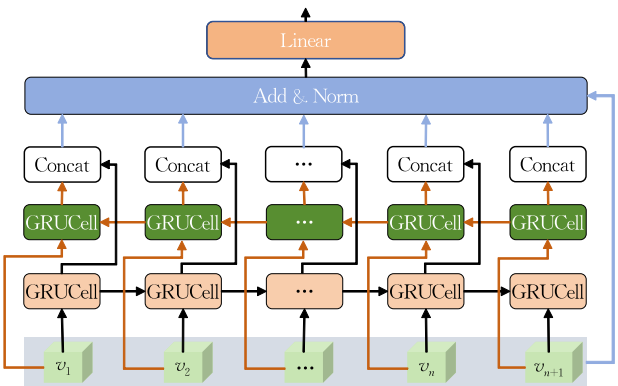


图 3　双向 GRU 图像增强网络

具体来说，该模块先利用 BiGRU 对图像特征 $\boldsymbol{V}$ 进行序列建模，最后输出语义增强后的图像特征 $\boldsymbol{V}^*$。BiGRU 需要分别处理图像特征 $\boldsymbol{V}=[\boldsymbol{v}_1,\cdots,\boldsymbol{v}_t,\cdots,\boldsymbol{v}_{n+1}]$ 的正向和反向序列，设 t 时刻 BiGRU 的正向隐藏状态为 $\overrightarrow{\boldsymbol{h}}_t$，反向隐藏状态为 $\overleftarrow{\boldsymbol{h}}_t$，其计算过程如下：

$$\begin{aligned}\overrightarrow{\boldsymbol{h}}_t&=GRUCell_{\text{正向}}(\boldsymbol{v}_t,\overrightarrow{\boldsymbol{h}}_{t-1}),\\ \overleftarrow{\boldsymbol{h}}_t&=GRUCell_{\text{反向}}(\boldsymbol{v}_t,\overleftarrow{\boldsymbol{h}}_{t+1})\end{aligned} \quad (4)$$

其中，$GRUCell_{正向}$表示正向 GRU 单元，$GRUCell_{反向}$表示反向 GRU 单元，$\boldsymbol{v}_t$表示 t 时刻的图像特征，$\overrightarrow{\boldsymbol{h}}_{t-1}$表示 $t-1$ 时刻的正向隐藏状态，$\overleftarrow{\boldsymbol{h}}_{t+1}$表示 $t+1$ 时刻的反向隐藏状态。然后将正向和反向隐藏状态拼接作为 BiGRU 的输出，记为 $\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{n+1}]$，其中 $\boldsymbol{h}_t=Concat(\overrightarrow{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t)$，$t\in[1,n+1]$。

此外，为了防止图像特征中原有的全局语义信息的丢失，本文对 BiGRU 的输出使用残差连接，最后使用线性层将语义强化后的图像特征映射到文本语义空间对齐文本特征，该过程如下：

$$\begin{aligned}\boldsymbol{V}^{Res}&=Layernorm(\boldsymbol{H}+\boldsymbol{V}),\\ \boldsymbol{V}^{*}&=\boldsymbol{w}_1\boldsymbol{V}^{Res}+b_1\end{aligned}\tag{5}$$

其中，$\boldsymbol{V}^{Res}$表示残差连接后的图像特征，$Layernorm$表示层归一化，加快模型收敛速度的同时保持模型鲁棒性，$\boldsymbol{V}^{*}$是视觉语义残差网络的输出结果，参数$\boldsymbol{w}_1$和b_1分别是线性层可学习的权重和偏置。

3.3 证据文本筛选网络

为了减少证据文本 $\boldsymbol{N}$ 的冗余，本文设计了基于注意力机制的证据文本筛选网络，如图 4 所示，该网络利用注意力评分衡量证据文本 $\boldsymbol{N}$ 的重要性，剔除评分小于指定阈值的 token，从而得到经过筛选的证据文本 $\boldsymbol{N}^{*}$。首先，将新闻文本的特征序列 $\boldsymbol{D}$ 映射到查询空间 $\boldsymbol{Q}_D$，证据文本特征序列 $\boldsymbol{N}$ 分别映射到键空间 $\boldsymbol{K}_N$和值空间 $\boldsymbol{V}_N$，再对 $\boldsymbol{Q}_D$、$\boldsymbol{K}_N$进行矩阵运算得到新闻文本与证据文本的注意力评分矩阵 $\boldsymbol{\alpha}$，具体如下：

$$\begin{aligned}\boldsymbol{Q}_D&=\boldsymbol{W}_Q\boldsymbol{D},\\ \boldsymbol{K}_N&=\boldsymbol{W}_K\boldsymbol{N},\\ \boldsymbol{V}_N&=\boldsymbol{W}_V\boldsymbol{N}\end{aligned}\tag{6}$$

其中，$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$是可学习的参数矩阵。

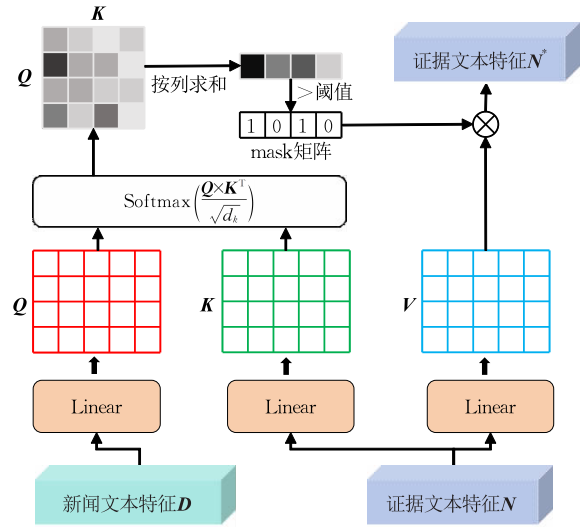


图 4 证据文本筛选网络

然后对 $\boldsymbol{\alpha}$ 按列求和得到证据文本每个 token 的重要性权重：

$$\begin{aligned}\boldsymbol{\alpha}&=\mathrm{Softmax}(\boldsymbol{Q}_D\boldsymbol{K}_N^{\mathrm{T}}/\sqrt{d_k}),\\ \boldsymbol{\beta}&=\Big[\sum_{i=1}^{m_D}\alpha_{ij}\Big]_{j=1}^{m_N}\end{aligned}\tag{7}$$

其中，d_k表示文本特征的维度，m_D表示新闻文本序列长度，m_N表示证据文本序列长度，α_{ij}表示证据文本中第 j 个 token 对新闻文本中第 i 个 token 的重要性，$\boldsymbol{\beta}$表示证据文本中每个 token 对于整个新闻文本的重要程度。

最后，通过设置阈值筛选出 $\boldsymbol{V}_N$中特征重要性大于阈值的 token，从而剔除证据文本中的冗余信息，具体如下：

$$\begin{aligned}\boldsymbol{N}^{*}&=[\boldsymbol{V}_{N_i}*mask(\beta_i)]_{i=1}^{n},\\ mask(x)&=\begin{cases}1, & \text{如果 } x\geq threshold\\ 0, & \text{其他}\end{cases}\end{aligned}\tag{8}$$

其中，$\boldsymbol{N}^{*}$表示经过筛选后的证据文本特征，$mask$ 表示阈值筛选函数，β_i表示证据文本中第 i 个 token 对于整个新闻文本的重要程度，$threshold$ 表示阈值。

3.4 跨模态语义交互与检测网络

为了缩小语义鸿沟，增强图像与文本的局部语义交互，更细粒度捕捉各模态语义信息的异同，本文搭建跨模态语义信息交互网络，其结构如图 5 所示，该网络结构包含图像分支和文本分支，每条分支可分为基于自注意力子层(Self-Att)的信息融合阶段、基于交叉注意力子层(Cross-Att)的跨模态语义交互阶段和虚假新闻检测阶段。

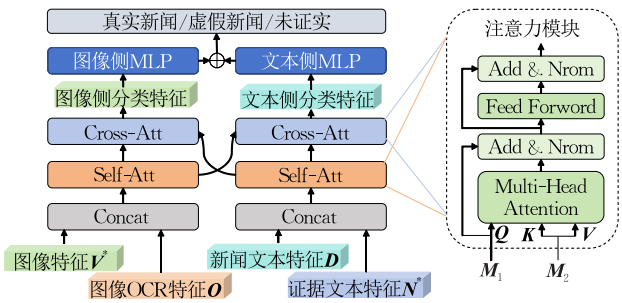


图 5 跨模态语义交互与检测网络

具体来说，Self-Att 和 Cross-Att 的结构如图 5 右侧的注意力模块所示，其包含一个多头注意力函数和一个前馈神经网络，二者后面都有一个残差连接和归一化模块，其输入数据 $\boldsymbol{M}_1$会映射到查询空间 $\boldsymbol{Q}$，$\boldsymbol{M}_2$会映射到键空间 $\boldsymbol{K}$ 和值空间 $\boldsymbol{V}$，其过程如下：

$$\begin{aligned}&\boldsymbol{Q}=\boldsymbol{W}_{att}^{Q}\boldsymbol{M}_1,\ \boldsymbol{K}=\boldsymbol{W}_{att}^{K}\boldsymbol{M}_2,\ \boldsymbol{V}=\boldsymbol{W}_{att}^{V}\boldsymbol{M}_2,\\ &\boldsymbol{M}_1^{*}=Laryernorm(MA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{M}_1),\\ &\boldsymbol{M}_{att}=Laryernorm(FF(\boldsymbol{M}_1^{*})+\boldsymbol{M}_1^{*})\end{aligned}\tag{9}$$

其中，$\boldsymbol{W}_{att}^{Q}$、$\boldsymbol{W}_{att}^{K}$、$\boldsymbol{W}_{att}^{V}$是可学习的参数矩阵，MA 表示多头注意力函数，FF 表示前馈神经网络，$\boldsymbol{M}_{att}$表示注意力模块的输出。

在信息融合阶段，图像分支的输入是经过BiGRU增强后的图像特征序列 $\boldsymbol{V}^{*}$ 和图像 OCR 特征序列。文本分支的输入是新闻文本特征序列 $\boldsymbol{D}$ 和筛选后的证据文本特征序列 $\boldsymbol{N}^{*}$。各分支在信息融合之前，需要拼接各特征，其过程如下所示：

$$\begin{aligned}\boldsymbol{V}_F&=Concat(\boldsymbol{V}^{*},\boldsymbol{O}),\\ \boldsymbol{D}_F&=Concat(\boldsymbol{D},\boldsymbol{N})\end{aligned}\tag{10}$$

其中，$Concat$ 表示拼接操作，$\boldsymbol{V}_F$序列长度为 $\boldsymbol{V}^{*}$与 $\boldsymbol{O}$ 序列长度之和，$\boldsymbol{D}_F$序列长度为 $\boldsymbol{D}$与 $\boldsymbol{N}$ 序列长度之和。

然后，$\boldsymbol{V}_F$和 $\boldsymbol{D}_F$分别进入不同的 Self-Att 模块，利用自注意力学习图像与图像 OCR 之间的互补知识，以及学习新闻内容与证据之间的内部关联性，得到经过 OCR 增强的图像特征 $\boldsymbol{V}_{FF}$和经过证据增强的文本特征 $\boldsymbol{D}_{FF}$。该过程如下：

$$\begin{aligned}\boldsymbol{V}_{FF}&=SelfAtt_V(\boldsymbol{V}_F,\boldsymbol{V}_F),\\ \boldsymbol{D}_{FF}&=SelfAtt_D(\boldsymbol{D}_F,\boldsymbol{D}_F)\end{aligned}\tag{11}$$

其中，$SelfAtt$ 中的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 都是由同一映射而来，$SelfAtt_V$是图像分支的自注意力模块，$SelfAtt_T$是文本分支的自注意力模块。

在跨模态语义交互阶段，该网络使用交叉注意力子层加强模型对于图文特征局部语义的理解，在图像分支中使用图像特征指导模型学习文本局部语义信息，在文本分支中使用文本特征指导模型学习图像局部语义信息，从而实现细粒度的语义交互，具体过程如下所示：

$$\begin{aligned}\widetilde{\boldsymbol{V}}&=CrossAtt_{V\to D}(\boldsymbol{V}_{FF},\boldsymbol{D}_{FF}),\\ \widetilde{\boldsymbol{D}}&=CrossAtt_{D\to V}(\boldsymbol{D}_{FF},\boldsymbol{V}_{FF})\end{aligned}\tag{12}$$

其中，在过程 $V\to D$ 中，$\boldsymbol{V}_{FF}$被映射到查询空间 $\boldsymbol{Q}_{V\to D}$空间，$\boldsymbol{D}_{FF}$ 被分别映射到键空间 $\boldsymbol{K}_{V\to D}$和值空间 $\boldsymbol{V}_{V\to D}$。在过程 $D\to V$ 中，$\boldsymbol{D}_{FF}$被映射到查询空间 $\boldsymbol{Q}_{D\to V}$，$\boldsymbol{V}_{FF}$ 被分别映射到键空间 $\boldsymbol{K}_{D\to V}$和值空间 $\boldsymbol{V}_{D\to V}$。输出 $\widetilde{\boldsymbol{V}}$、$\widetilde{\boldsymbol{D}}$ 分别表示经过跨模态局部语义交互之后的图像侧特征和文本侧特征。

在虚假新闻检测阶段，使用图像侧特征 $\widetilde{\boldsymbol{V}}$ 和文本侧特征 $\widetilde{\boldsymbol{D}}$ 中特殊字符“CLS”对应的特征向量 $\boldsymbol{V}_{cls}$和 $\boldsymbol{D}_{cls}$作为分类特征，分别经过两个全连接层获取图像分支预测虚假新闻的概率分布 y_V和文本分支预测虚假新闻的概率分布 y_T，最后，取两个分支的均值作为最终结果 $\bar{y}$，该过程如下：

$$\begin{aligned}y_T&=\mathrm{Softmax}(MLP_T(\boldsymbol{D}_{cls})),\\ y_V&=\mathrm{Softmax}(MLP_V(\boldsymbol{V}_{cls})),\\ \bar{y}&=(y_T+y_V)/2\end{aligned}\tag{13}$$

本文的虚假新闻检测任务被建模为一个多分类问题。在 MR2 数据集中，新闻需要被分类为“虚假”、“真实”和“无法证实”三类；而在 Weibo 数据集中，新闻则被分为“虚假”和“真实”两类。为优化模型性能，本文采用了交叉熵损失函数，通过最小化预测标签与真实标签之间的差异，来提高模型的分类准确度：

$$L_{CE}=-\sum_{i=1}^{c}y_i\log(\bar{y}_i)\tag{14}$$

其中，L_{CE}表示模型的交叉熵损失，c 表示标签的类别数，y_i表示真实标签的第 i 个类别，$\bar{y}_i$ 表示预测标签的第 i 个类别的概率。

4　实验与分析

4.1　数据集

本文使用 2 个从真实社交媒体采集的 MR2、Weibo 数据集对虚假新闻检测模型进行评估。数据集的详细信息统计如表 1 所示。

表 1　数据集详细统计信息

数据集		虚假新闻数量	真实新闻数量	无法证实数量
MR2 中文	训练集	1426	2000	2980
	验证集	201	209	304
	测试集	190	202	187
MR2 英文	训练集	966	1592	2220
	验证集	190	202	202
	测试集	201	209	319
Weibo	训练集	3395	2807	—
	测试集	924	835	—

MR2 数据集[41]是 Hu 等人构建的一个多模态检索增强虚假新闻数据集，含有不同领域声明的社交媒体多模态数据。该数据集包含中文和英文两种语言，作者将数据分为虚假新闻、真实新闻和无法证实三种类型。该数据集中的每条数据包含声明、声明对应的图片、标签以及根据声明和图片检索到的网页证据内容。

Weibo 数据集[29]包含了国内主流社交媒体的新闻数据，其中真实新闻来源于中国权威媒体平台新华社，虚假新闻数据则来自微博官方辟谣系统在 2012 年 5 月至 2016 年 1 月期间核实的虚假新闻记录。由于该数据集中部分图片链接已失效，本文已删除缺失图片的相关数据。此外，Weibo 数据集只

包含新闻声明与新闻图片，不包含网页证据内容，所以本文遵循 MR2 数据集中的方法爬取网页证据，该过程如下：

网页证据内容由新闻的声明和图片检索得到，首先，将新闻的声明作为查询，并使用 Google 可编程搜索引擎将检索到的前 5 个图像及其对应网页作为新闻图片的外部证据。然后，将新闻的图片作为查询，并使用 Google 反向图像搜索引擎将检索到的前 5 个图像及其对应网页作为新闻声明的外部证据。

4.2 实验设置

在数据预处理部分需要截断文本，新闻内容的最大长度为 64，新闻证据的最大长度为 128，图像 OCR 的最大长度是 200。在特征抽取部分，使用预训练语言模型 RoBERTa-Base 提取文本特征，词向量的维度为 768。CLIP 提取图像特征，维度为 512。BiGRU 隐藏层向量维度为 512。跨模态语义信息融合模块输入输出维度为 768。训练过程中使用 Adam 优化器更新参数。初始学习率设置为 10^{-5}，使用线性学习率预热调整学习率，批大小为 64，迭代轮次为 10。

在评估指标选择上，采用准确率(Accuracy)和 $F1$ 值($F1$ Score)对实验结果进行评估。

4.3 对比实验

4.3.1 对比方法

为了验证本文所提出方法的有效性，本文对虚假新闻检测的各类经典方法作对比。

(1) Bi-LSTM[16]：使用双向长短期循环神经网络表征新闻文本，将学习到的特征向量经过线性层得到虚假新闻的概率分布，是一种文本侧的单模态检测方法。

(2) Bert[42]：使用预训练语言模型提取文本特征，将特殊字符“CLS”对应的特征向量经过线性层分类，是一种文本侧的单模态检测方法。

(3) VGG-19[43]：使用预训练图像模型提取图像特征，并将特征经过线性层分类，是一种图像侧的单模态检测方法。

(4) CLIP[36]：使用 CLIP 的图像编码器提取图像特征，将特殊字符“CLS”对应的特征向量经过线性层分类，是一种图像侧的单模态检测方法。

(5) Att-RNN[29]：Att-RNN 是多模态循环神经网络，使用 LSTM 融合新闻文本和社交上下文信息，VGG 提取图片信息，并使用注意力机制进行模态融合。

(6) SpotFake[25]：SpotFake 是一种多模态虚假新闻检测框架，使用 Bert 和 VGG 分别提取文本和图像特征，摆脱了对子任务的依赖。

(7) LIIMR[44]：LIIMR 是一种基于模态内和模态间关系的多模态虚假新闻检测方法，使用预训练模型提取细粒度图像和文本特征，并建立模态内关系，最后使用乘法多模态融合方法捕获模态间关系。

(8) FND-CLIP[39]：FND-CLIP 是一种用 CLIP 指导的多模态虚假新闻检测方法，使用多模态预训练模型 CLIP 计算图文相似性，指导模型跨模态相似性加权融合，最后将提取的特征送入分类器。

(9) CCN[45]：CCN 是一种多模态循环一致性检验方法，通过比较新闻文本与文本证据、新闻图片与图片证据以及新闻文本与新闻图片之间的关系，挖掘多模态证据与新闻内容的一致性关系。

(10) END[46]：END 是一种端到端的多模态事实核查框架，使用 CLIP 分别对新闻文本、图像证据、文本证据进行编码，图文证据特征分别与新闻文本特征串联，最后送入分类器判断新闻真实性。

(11) MR2[41]：MR2 是一种基于检索的多模态虚假新闻检测方法，使用 Bert 和 ResNet 分别提取文本和图像特征，并利用注意力机制融合证据特征，最后将文本特征和图像特征拼接后送入分类器。

(12) GPT-4[47]：调用 OpenAI 的 GPT-4 模型接口，使用基于原因感知的推理模板[21]，通过文本内容判断新闻的真实性。

(13) GPT-4V[48]：调用 OpenAI 的 GPT-4V 模型接口，使用基于原因感知的推理模板，通过文本内容与图片内容判断新闻的真实性。

4.3.2 结果分析

表 2 列出了本实验方法与其他基线方法在 MR2 中、英文数据集与 Weibo 数据集上的性能比较结果，经过观察分析得到以下结果：

(1) 本文提出的方法大部分指标优于其他对比方法。在准确率上，本文方法优于 MR2 中英文数据集和 Weibo 数据集上的所有对比方法，分别为 90.8%、88.5%、91.9%，超出对比方法中最好结果 2.4%、4.9%、0.8%，有效提升了虚假新闻检测性能。此外，本文方法在所有测试中的 $F1$ 值均排名第一，表明了本文提出的模型综合考虑了虚假新闻检测的准确性和完整性，表现得更加稳健。

表 2　对比实验性能比较

数据集	方法	准确率	虚假新闻			真实新闻			无法证实		
			精确率	召回率	*F*1 值	精确率	召回率	*F*1 值	精确率	召回率	*F*1 值
MR2 中文	Bi-LSTM	0.730	0.678	0.589	0.630	0.892	0.821	0.855	0.640	0.780	0.703
	Bert	0.791	0.914	0.505	0.651	0.773	**0.980**	0.864	0.752	0.877	0.810
	VGG-19	0.734	0.839	0.631	0.720	0.808	0.792	0.800	0.609	0.775	0.682
	CLIP	0.820	0.870	0.742	0.801	0.881	0.886	0.883	0.724	0.828	0.773
	Att-RNN	0.775	0.727	0.673	0.699	**0.934**	0.772	0.845	0.699	0.882	0.780
	SpotFake	0.830	0.850	0.657	0.741	0.871	0.970	0.918	0.772	0.855	0.811
	LIIMR	0.826	0.935	0.611	0.739	0.890	0.965	0.926	0.708	0.893	0.790
	FND-CLIP	0.884	**0.967**	0.779	0.863	0.903	0.970	0.935	0.804	**0.898**	0.848
	CCN	0.855	0.892	0.784	0.835	0.899	0.921	0.910	0.780	0.856	0.816
	END	0.839	0.838	0.763	0.799	0.882	0.886	0.884	0.798	0.866	0.831
	MR2	0.850	0.764	0.755	0.759	0.864	0.937	0.899	**0.883**	0.837	0.859
	GPT-4	0.520	0.564	0.419	0.481	0.469	0.516	0.491	0.538	0.601	0.568
	GPT-4V	0.531	0.587	0.455	0.513	0.468	0.511	0.489	0.552	0.607	0.578
	Ours	**0.908**	0.942	**0.847**	**0.892**	0.916	0.975	**0.945**	0.870	**0.898**	**0.884**
MR2 英文	Bi-LSTM	0.729	0.551	0.373	0.445	0.811	0.885	0.846	0.745	0.852	0.795
	Bert	0.759	0.614	0.388	0.476	0.780	0.866	0.821	0.797	0.924	0.856
	VGG-19	0.624	0.528	0.328	0.405	0.664	0.626	0.644	0.633	0.808	0.710
	CLIP	0.758	0.601	0.601	0.601	0.944	0.732	0.825	0.762	0.874	0.814
	Att-RNN	0.733	0.561	0.383	0.455	0.922	0.794	0.853	0.713	0.962	0.819
	SpotFake	0.802	0.670	0.577	0.620	0.867	0.846	0.856	0.829	0.915	0.870
	LIIMR	0.764	0.680	0.348	0.460	0.888	0.914	0.901	0.720	0.928	0.811
	FND-CLIP	0.789	0.657	0.562	0.606	0.861	0.742	0.797	0.814	0.962	0.882
	CCN	0.829	0.688	0.701	0.694	0.970	0.785	0.868	0.842	0.937	0.887
	END	0.819	0.655	**0.736**	0.693	0.953	0.775	0.855	0.862	0.900	0.881
	MR2	0.836	0.682	0.730	0.705	0.913	0.897	0.905	**0.865**	0.846	0.855
	GPT-4	0.638	0.750	0.149	0.249	0.678	0.856	0.757	0.602	0.803	0.688
	GPT-4V	0.663	0.726	0.267	0.390	0.600	0.962	0.739	0.713	0.718	0.715
	Ours	**0.885**	**0.870**	0.697	**0.774**	**0.980**	**0.938**	**0.959**	0.840	**0.969**	**0.900**
Weibo	Bi-LSTM	0.736	0.752	0.745	0.748	0.720	0.728	0.724	—	—	—
	Bert	0.845	**0.957**	0.738	0.833	0.769	**0.963**	0.855	—	—	—
	VGG-19	0.635	0.630	0.706	0.666	0.641	0.559	0.597	—	—	—
	CLIP	0.771	0.769	0.778	0.773	0.772	0.764	0.768	—	—	—
	Att-RNN	0.795	0.925	0.644	0.759	0.724	0.947	0.821	—	—	—
	SpotFake	0.903	0.917	0.887	0.904	0.889	0.920	0.902	—	—	—
	LIIMR	0.905	0.904	0.909	0.906	0.907	0.902	0.904	—	—	—
	FND-CLIP	0.911	0.927	0.893	0.910	0.895	0.929	0.912	—	—	—
	CCN	0.901	0.950	0.849	0.897	0.861	0.955	0.906	—	—	—
	END	0.894	0.933	0.850	0.890	0.861	0.938	0.898	—	—	—
	MR2	0.899	0.896	0.904	0.900	0.902	0.894	0.898	—	—	—
	GPT-4	0.744	0.897	0.521	0.659	0.686	0.946	0.795	—	—	—
	GPT-4V	0.762	0.913	0.551	0.687	0.701	0.952	0.807	—	—	—
	Ours	**0.919**	0.892	**0.955**	**0.922**	**0.951**	0.883	**0.916**	—	—	—

注：粗体数字表示最优结果；其他表格同理。

(2) 从表 2 中可以发现，无论是通过直接拼接特征还是跨模态增强的方式，将其他模态信息引入单模态方法，都能提升模型检测虚假新闻的能力，这表明其他模态的语义信息可以补充或增强单一模态的语义表达，从而丰富模型对数据的语义理解，提高模型对复杂多变的虚假新闻的识别能力。

(3) 相比于将各模态特征串联融合的 CCN、END、MR2 方法，FND-CLIP 的表现不够稳定，其利用图文一致性从整体语义层面指导模型融合多模态语义信息，忽略了图文的局部语义理解，当图像与文本的相似性较低时，模型无法深度挖掘图像特征中的语义信息，限制了模型的表达能力。此外，CCN、END、MR2 方法虽然设计了特定模块处理证据信息，但并未过滤证据数据中的冗余信息，导致其性能可能不如直接拼接证据文本的 FND-CLIP，而本文提出的方法，不仅使用证据筛选模块剔除冗余数据，还使用基于交叉注意力机制的跨模态语义交互模块学习图文之间的局部语义信息，提升虚假新闻检测的准确性与稳定性。

(4) 在大模型方法中，基于纯文本的 GPT-4 方

法和基于多模态的 GPT-4V 方法在 MR2 与 Weibo 数据集上的虚假新闻检测能力较弱，尽管 GPT-4 和 GPT-4V 在理解和生成语言时表现出较强的能力，但其“幻觉”现象在面对复杂且模糊的语境时，可能偏离真实信息，导致模型难以正确识别细微的虚假信息。相比之下，本文提出的方法更加稳健，推理成本也低于基于大模型的虚假新闻检测方法。

4.4 消融实验

4.4.1 对比方法

为了评估模型中每个组件的有效性，本节依次去除每个组件，并分别对模型进行消融分析，模型的变体如下：

(1) 去除证据筛选网络：不对证据文本做处理，直接送入模型训练。

(2) 去除证据增强网络：不使用证据文本增强文本语义信息。

(3) 去除 OCR 增强网络：不使用图像 OCR 增强图像语义信息。

(4) 去除跨模态增强网络：使用线性层替换跨模态局部语义信息交互网络。

(5) 去除双向 GRU 图像增强网络：不使用双向 GRU 增强图像语义。

4.4.2 结果分析

消融实验结果如表 3 所示，观察得到以下结论：

(1) 移除模型中的任意一个组件，都会对模型的性能产生不同程度的影响，这表明了模型的各个组成部分对于模型整体性能的重要性。

(2) 当模型移除了证据增强网络后，MR2 中英文数据集中虚假新闻的召回率都大幅下降，说明证据信息能帮助模型有效鉴别虚假新闻。通过证据筛选模块的消融实验可以发现，证据内容中的冗余信息会影响模型判断，导致模型不能很好地学习证据文本中的有效信息。证据筛选网络通过注意力机制判断证据与新闻文本的相关性，引导模型聚焦学习相关性较高的证据文本，从而剔除与新闻无关的冗余信息，增强了模型对于虚假新闻的鉴别能力。

(3) 图像中的 OCR 数据是图像语义信息的重要组成部分，当模型移除了 OCR 增强网络后，仅依靠图像特征无法获取图像中完整的语义信息，这会削弱模型检测虚假新闻的能力。通过自注意力机制将图像中的 OCR 信息融入图像特征，能够有效地补充图像语义信息，增强了模型的可解释性，这说明了模态的语义表达能力对虚假新闻检测起重要作用。

(4) 通过对比去除各组件对模型性能的影响程度可以发现，去除跨模态增强网络后模型性能下降最明显，表明跨模态局部语义交互网络是模型提升虚假新闻检测能力的重要组件。去除跨模态交互网络后，仅仅融合文本和图像两个单模态的检测结果无法深度挖掘多模态特征的内在语义联系，导致在中英文两个数据集上都表现不佳。这说明本文所提出的跨模态语义交互模块能够细粒度地感知文本和图像模态在语义上的差异，并且能够利用多模态来增强单一模态的语义表达能力，让文本和图像模态能够有效互补、相互增强。

表 3 消融实验性能比较

数据集	方法	准确率	虚假新闻			真实新闻			无法证实		
			精确率	召回率	*F*1 值	精确率	召回率	*F*1 值	精确率	召回率	*F*1 值
MR2 中文	去除证据筛选网络	0.893	0.967	0.789	0.869	0.892	0.990	0.938	0.835	0.893	0.863
	去除证据增强网络	0.860	**0.984**	0.652	0.784	0.884	0.990	0.934	0.767	**0.930**	0.841
	去除OCR增强网络	0.872	0.950	0.800	0.869	0.896	0.985	0.938	0.842	0.887	0.864
	去除跨模态增强网络	0.823	0.912	0.605	0.727	0.879	0.975	0.925	0.724	0.887	0.797
	去除双向GRU图像增强网络	0.884	0.929	0.831	0.877	0.866	**0.995**	0.926	0.864	0.818	0.840
	本文方法	**0.908**	0.942	**0.847**	**0.892**	**0.916**	0.975	**0.945**	**0.870**	0.898	**0.884**
MR2 英文	去除证据筛选网络	0.846	**0.891**	0.527	0.662	0.957	**0.947**	0.952	0.777	0.981	0.867
	去除证据增强网络	0.833	0.794	0.557	0.655	0.893	0.880	0.886	0.814	0.975	0.887
	去除OCR增强网络	0.819	0.709	0.607	0.654	0.947	0.775	0.852	0.811	0.981	0.888
	去除跨模态增强网络	0.786	0.626	0.557	0.589	0.961	0.722	0.825	0.789	0.971	0.871
	去除双向GRU图像增强网络	0.827	0.770	0.562	0.650	0.978	0.842	0.905	0.781	**0.984**	0.871
	本文方法	**0.885**	0.870	**0.697**	**0.774**	**0.980**	0.938	**0.959**	**0.840**	0.969	**0.900**

4.5 证据筛选阈值定量分析

本文在证据筛选网络中通过设置阈值的方式过滤证据中的冗余信息，但不同的阈值对冗余信息的过滤效果不同。本节中阈值指的是相对阈值，例如 0.4 表示剔除注意力评分位于后 40% 的 token。为了探究阈值大小对模型的影响，本节设置了不同大小的阈值，对模型进行定量分析，实验结果如图 6、图 7 所示，通过观察与分析可以得到以下结论：

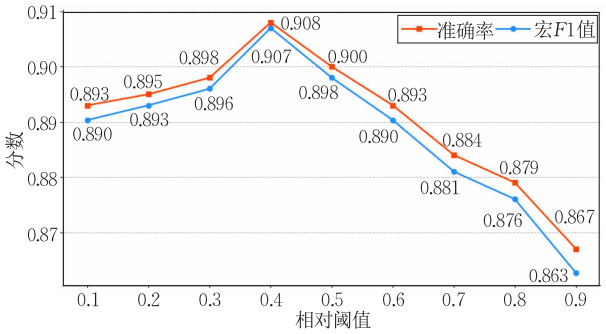


图 6　MR2 中文数据集上的证据筛选阈值定量分析

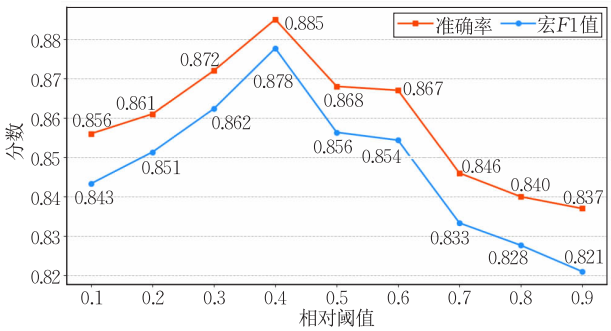


图 7　MR2 英文数据集上的证据筛选阈值定量分析

（1）通过图 6、图 7 两个折线图可以观察到，阈值设置过大或过小均会对模型的虚假新闻检测性能产生负面影响。当阈值设置过低时，模型无法有效剔除证据文本中的冗余信息，从而干扰模型的判断能力。而当阈值设置过高时，则可能导致关键的证据信息被过度剔除，削弱模型对证据信息的整体理解，最终影响其检测性能。

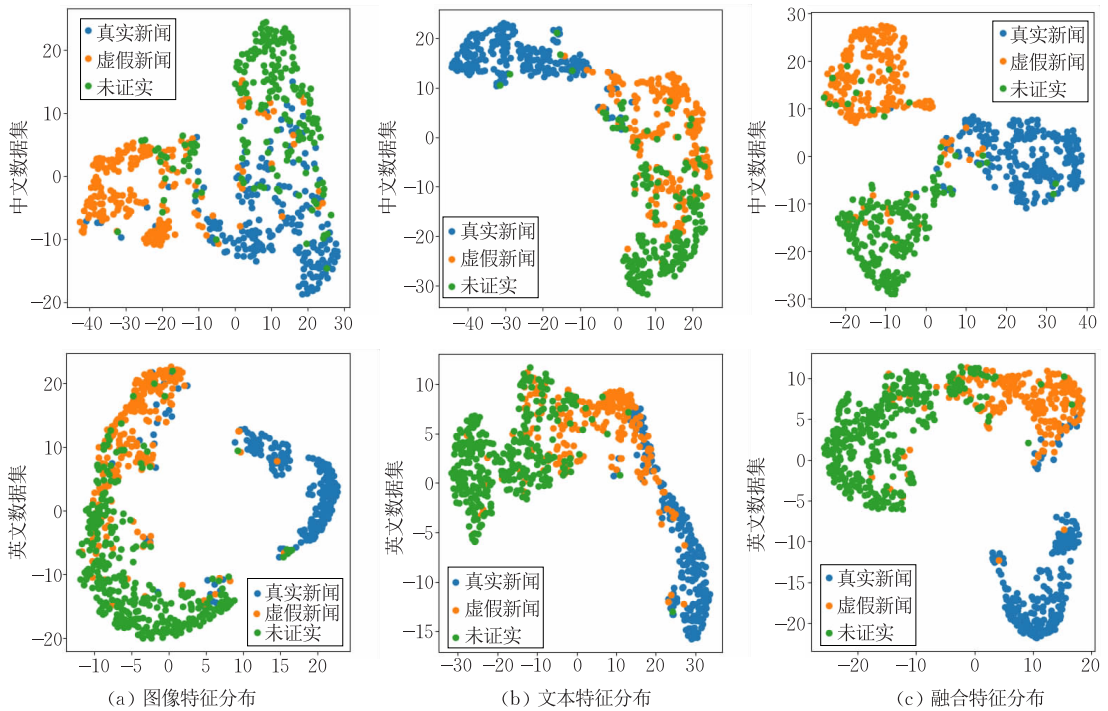
（2）通过设置不同的阈值对模型性能进行对比实验，可以发现当阈值设定为 0.4 时，模型的预测效果达到最佳。在此阈值下，模型的准确率和宏 $F1$ 值均位居第一。随着阈值从 0 逐步增加至 0.4，宏 $F1$ 值逐步提升，这表明适度剔除证据中的冗余信息有助于增强模型对虚假新闻的识别能力。然而，当阈值进一步增大至 0.8 时，宏 $F1$ 值逐渐下降。这一趋势表明，过高的阈值会导致重要证据信息的丢失，从而削弱模型的性能。因此，本文最终将阈值设定为 0.4，以实现冗余信息的有效过滤，同时保留关键的有效信息，达到性能的最佳平衡。

4.6　T-SNE 可视化

图 8 展示了文本分类特征、图像分类特征与融合特征在 MR2 中英文数据中的特征分布 t-SNE 二维可视化结果。其中相同颜色的点表示属于同一标签，通过观察与分析可以得到以下结论：

（1）中文数据集训练出的特征区分度高于英文训练集的特征区分度，这表明模型更容易学到中文数据集中的新闻语义信息，其原因在于中文数据的图片中存在更多的 OCR 数据，也说明了 OCR 数据能有效补充图像语义信息。

（2）对比图像侧和文本侧特征分布情况，可以直观发现文本侧特征学习到了更多可区分度高的语义信息，这表明模型鉴别虚假新闻主要以文本语义



（a）图像特征分布　（b）文本特征分布　（c）融合特征分布

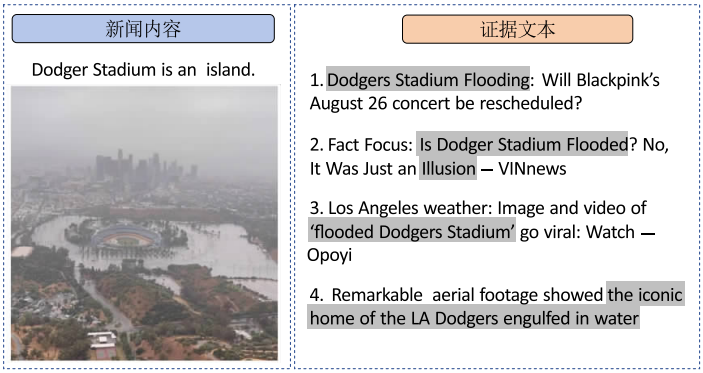
图 8　MR2 中英文数据集各特征的 t-SNE 二维可视化结果

为主,尽管图像侧特征的区分度低于文本侧特征,但是也为模型提供了丰富的语义信息,最终模型结合文本侧分类特征与图像侧分类特征共同检测虚假新闻,并表现出良好的性能。

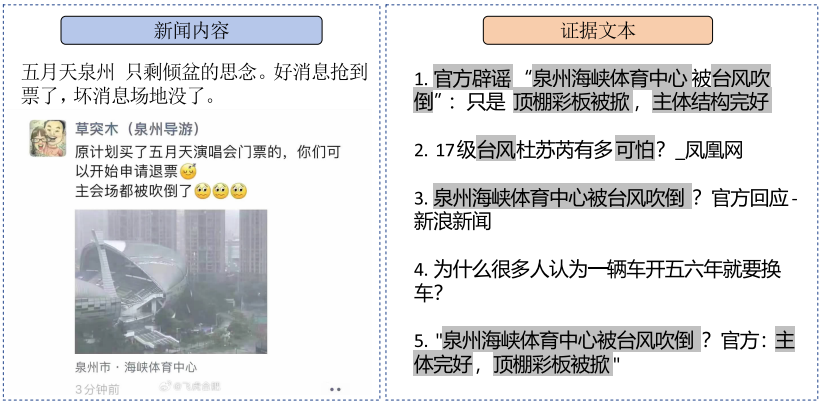
(3)从整体来看,跨模态增强后的单一文本特征与图像特征能正确区分大部分样本,但是不能有效处理各标签的边界问题。融合了图像和文本信息的特征各类别标签边界清晰度更高,表明先通过跨模态增强各模态语义信息,再进行多模态融合能够有效整合不同模态的互补信息,从而提高特征区分度。

4.7 案例分析

为了直观展示模型对于证据文本的筛选效果,本文从 MR2 中英文数据集中各选取了一个代表性的虚假新闻进行案例分析。图 9 展示了模型对于证据筛选的效果,证据文本中的深色阴影区域表示与新闻文本相关度较高的文本信息。



(a) MR2英文数据集代表案例



(b) MR2中文数据集代表案例

图 9 多模态虚假新闻案例分析示意图

通过观察图 9(a)的英文代表案例可以发现,新闻内容中的文本与图像所表达的信息语义一致,若只通过新闻内容会将其判断为真实新闻。而将证据文本通过证据筛选网络筛选后,模型能够重点关注证据文本中的关键信息,比如模型识别到了图 9(a)的第四条证据的"the iconic home of the LA Dodgers engulfed in water",表明道奇体育场被水淹没了,并不是新闻文本中说的"道奇体育场是一个岛屿",证明了该新闻是虚假新闻。

从图 9(b)的中文代表案例可以发现模型也识别出了证据文本中的关键信息,虽然新闻内容图文语义信息一致,但是从证据文本中筛选出的"官方辟谣""主体结构完好"等关键信息可以判断该新闻为虚假新闻。

5 总 结

针对现有的多模态虚假新闻检测方法无法充分挖掘图文之间的互补语义信息,并且缺乏证据对新闻进行多方验证的问题,本文提出了一种基于证据增强和局部语义理解的多模态虚假新闻检测方法。该方法使用 CLIP 提取包含细粒度语义信息的图像序列特征,并通过双向 GRU 图像增强网络进一步强化图像语义特征。同时,设计证据筛选网络剔除证据文本中的冗余信息。最后,在跨模态语义交互与检测网络利用自注意力机制对图像和文本分别进行 OCR 增强和证据增强,并通过交叉注意力机制细粒度强化文本特征与图像特征的局部语义信息。

实验结果证明了本文所提出方法在虚假新闻检测任务中的有效性和网络的合理性。

除了文本证据外，新闻内容中还可能包含图像证据，而本文提出的方法主要基于文本证据进行增强，未能充分利用网页中的图像信息。因此，本文模型的局限性在于只关注了证据中的文本信息而无法有效筛选和利用证据中的图像信息。为此，未来的工作将重点探索如何将图像证据有效整合进虚假新闻检测模型。首先，相关网页中往往存在大量与新闻内容无关的图像，如何有效筛选与新闻相关的图像是一个挑战。为此，可以设计一个图片检索模块，以提升图像与新闻内容之间的相关性。其次，证据图片中不同区域包含的语义信息量差异较大，因此考虑引入一个图像筛选模块，针对图像区域进行语义过滤，去除语义信息量较低的区域，从而突出图像中更具辨识度的关键信息。最后，进一步探索如何利用经过筛选的图像特征与文本特征进行融合，以增强模型的多模态信息处理能力，从而提升虚假新闻检测的整体性能。

参　考　文　献

[1] Li K, Guo B, Liu J, et al. Dynamic probabilistic graphical model for progressive fake news detection on social media platform. ACM Transactions on Intelligent Systems and Technology, 2022, 13(5): 1-24

[2] Zhang Zhi-Yong, Jing Jun-Chang, Li Fei, et al. Survey on fake information detection, propagation and control in online social networks from the perspective of artificial intelligence. Chinese Journal of Computers, 2021, 44(11): 2261-2282(in Chinese)
(张志勇，荆军昌，李斐等. 人工智能视角下的在线社交网络虚假信息检测、传播与控制研究综述. 计算机学报，2021，44(11)：2261-2282)

[3] Lu Y J, Li C T. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648, 2020

[4] Dou Y, Shu K, Xia C, et al. User preference-aware fake news detection//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021: 2051-2055

[5] Cui C, Jia C. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(1): 73-81

[6] Kang Z, Cao Y, Shang Y, et al. Fake news detection with heterogenous deep graph convolutional network//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2021: 408-420

[7] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 2020, 404: 132306

[8] Ma J, Gao W, Wong K F. Detect rumors on twitter by promoting information campaigns with generative adversarial learning//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 3049-3055

[9] Cao J. Exploring the role of visual content in fake news detection//Shu K, Wang S, Lee D, et al, eds. Disinformation, Misinformation, and Fake News in Social Media. Cham: Springer International Publishing, 2020: 141-161

[10] Zhou P, Han X, Morariu V I, et al. Learning rich features for image manipulation detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1053-1061

[11] Khattar D, Goud J S, Gupta M, et al. MVAE: Multimodal variational autoencoder for fake news detection//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 2915-2921

[12] Zhou X, Wu J, Zafarani R. SAFE: Similarity-aware multi-modal fake news detection//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, 2020: 354-367

[13] Qian S, Wang J, Hu J, et al. Hierarchical multi-modal contextual attention network for fake news detection//Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2021: 153-162

[14] Chen Y, Li D, Zhang P, et al. Cross-modal ambiguity learning for multimodal fake news detection//Proceedings of the ACM Web Conference. Texas, USA, 2022: 2897-2905

[15] Castillo C, Mendoza M, Poblete B. Information credibility on twitter//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 675-684

[16] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 3818-3824

[17] Dun Y, Tu K, Chen C, et al. KAN: Knowledge-aware attention network for fake news detection//Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1): 81-89

[18] Xu W, Wu J, Liu Q, et al. Evidence-aware fake news detection with graph neural networks//Proceedings of the ACM Web Conference 2022. Texas, USA, 2022: 2501-2510

[19] Shu K, Mahudeswaran D, Wang S, et al. Hierarchical propagation networks for fake news detection: Investigation and exploitation//Proceedings of the International AAAI Conference on Web and Social Media. Atlanta, USA, 2020: 626-637

[20] Liao H, Peng J, Huang Z, et al. MUSER: A multi-step evidence retrieval enhancement framework for fake news detection//Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. California, USA, 2023: 4461-4472

[21] Huang Y, Shu K, Yu P S, et al. From creation to clarification: ChatGPT's journey through the fake news quagmire//Companion Proceedings of the ACM on Web Conference 2024. Singapore, 2024: 513-516

[22] Qi P, Cao J, Yang T, et al. Exploiting multi-domain visual information for fake news detection//Proceedings of the IEEE International Conference on Data Mining. New York, USA, 2019: 518-527

[23] Singh B, Sharma D K. Predicting image credibility in fake news over social media using multi-modal approach. Neural Computing and Applications, 2022, 34(24): 21503-21517

[24] Wang Y, Ma F, Jin Z, et al. EANN: Event adversarial neural networks for multi-modal fake news detection//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 849-857

[25] Singhal S, Shah R R, Chakraborty T, et al. SpotFake: A multi-modal framework for fake news detection//Proceedings of the IEEE 5th International Conference on Multimedia Big Data. New Delhi, India, 2019: 39-47

[26] Shang L, Kou Z, Zhang Y, et al. A duo-generative approach to explainable multimodal COVID-19 misinformation detection//Proceedings of the ACM Web Conference. Lyon, France, 2022: 3623-3631

[27] Zhong Shan-Nan, Peng Shu-Juan, Liu Xin, et al. Multimodal fake news detection via two-branch deep clue perception and adaptive collaborative optimization. Chinese Journal of Computers, 2023, 46(12): 2612-2625(in Chinese)
(钟善男, 彭淑娟, 柳欣等. 双分支线索深度感知与自适应协同优化的多模态虚假新闻检测. 计算机学报, 2023, 46(12): 2612-2625)

[28] Xue J, Wang Y, Tian Y, et al. Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management, 2021, 58(5): 102610

[29] Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs//Proceedings of the 25th ACM International Conference on Multimedia. California, USA, 2017: 795-816

[30] Wu Y, Zhan P, Zhang Y, et al. Multimodal fusion with co-attention networks for fake news detection//Proceedings of the Association for Computational Linguistics and Asian Federation of Natural Language Processing. Bangkok, Thailand, 2021: 2560-2569

[31] Xiao T, Guo S, Huang J, et al. HiPo: Detecting fake news via historical and multi-modal analyses of social media posts//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham, UK, 2023: 2805-2815

[32] Tan H, Bansal M. LXMERT: Lcross-modality encoder representations from transformers//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hong Kong, China, 2019: 5100-5111

[33] Song H, Dong L, Zhang W N, et al. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. arXiv preprint arXiv:2203.07190, 2022

[34] Gao P, Geng S, Zhang R, et al. CLIP-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 2024, 132(2): 581-595

[35] Li J, Li D, Savarese S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the International Conference on Machine Learning. Hawaii, USA, 2023: 19730-19742

[36] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. 2021: 8748-8763

[37] Rao Y, Zhao W, Chen G, et al. DenseCLIP: Language-guided dense prediction with context-aware prompting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18082-18091

[38] Luo H, Ji L, Zhong M, et al. CLIP4Clip: An empirical study of CLIP for end to end video Clip retrieval and captioning. Neurocomputing, 2022, 508: 293-304

[39] Zhou Y, Yang Y, Ying Q, et al. Multimodal fake news detection via CLIP-guided learning//Proceedings of the IEEE International Conference on Multimedia and Expo. Brisbane, Australia, 2023: 2825-2830

[40] Du Y, Li C, Guo R, et al. PP-OCR: A practical ultra light-weight OCR system. arXiv preprint arXiv:2009.09941, 2020

[41] Hu X, Guo Z, Chen J, et al. MR2: A benchmark for multi-modal retrieval-augmented rumor detection in social media//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, China, 2023: 2901-2912

[42] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018

[43] Sengupta A, Ye Y, Wang R, et al. Going deeper in spiking neural networks: VGG and residual architectures. Frontiers in Neuroscience, 2019, 13: 95

[44] Singhal S, Pandey T, Mrig S, et al. Leveraging intra and inter modality relationship for multimodal fake news detection//Proceedings of the Web Conference 2022. Lyon, France, 2022: 726-734

[45] Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14940-14949

[46] Yao B M, Shah A, Sun L, et al. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taiwan, China, 2023: 2733-2743

[47] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023

[48] Yang Z, Li L, Lin K, et al. The dawn of LMMs: Preliminary explorations with GPT-4V (ision). arXiv preprint arXiv: 2309.17421, 2023, 9(1): 1

ZHONG Jiang, Ph. D., professor. His research interests include big data analysis and mining, natural language processing, cloud network convergence, network security and so on.

GAO Jin-Peng, M. S. candidate. His research interest is multimodalfake news detection.

HUANG Jing-Wang, M. S. candidate. His research interest is multimodal sentiment analysis.

YANG Yu-Ming, Ph. D. candidate. His research interest is customized multimodal generation.

Background

With the rapid development of the Internet and mobile technology, social media has become an important platform for information dissemination. Compared with traditional text news, news with pictures and texts in social media is more likely to mislead users, causing false news to be widely spread, thus having a negative impact on society. In order to cope with changes in social media content, false news detection methods have developed from single modality to multimodality.

At present, multimodal false news detection methods usually use pre-trained models to represent text features and image features in multimodal data, and then design downstream networks to mine feature clues that can judge false news. However, current methods only use the overall semantics between images and texts to guide model decisions, ignoring the local semantic interactions between images and texts, resulting in the model not being able to capture the semantic differences between images and texts well. In addition, it is difficult to identify the authenticity of news by relying solely on news content and without relying on external information. In order to solve the above problems, this paper proposes a multimodal fake news detection model based on evidence enhancement and local semantic understanding, introduces evidence text to verify the authenticity of news content from multiple parties, and realizes fine-grained interaction of local semantics between images and text through a cross-attention mechanism. The model proposed in this paper is compared with seven representative fake news detection baseline methods on Chinese and English datasets. The experimental results show that the model outperforms the baseline models in both precision and $F1$ value, proving the effectiveness and interpretability of the proposed model.

The research is part of a project supported by the National Natural Science Foundation of China (No. 62176029) and the Chongqing Science and Technology Bureau (Nos. CSTB2023-TIAD-KPX0064, CSTB2022TIAD-KPX0206). These projects are mainly for natural language processing in different fields, including social fields and cross-language fields.