融合外部知识与证据的场景图注意力网络多模态谣言检测

黄学坚[1),2)] 马廷淮[2),3)] 荣 欢[2)] 王根生[4)] 廖国琼[1)] 刘德喜[5)]

[1)](江西财经大学虚拟现实(VR)现代产业学院 南昌 330013)

[2)](南京信息工程大学计算机学院 南京 210044)

[3)](江苏海洋大学计算机工程学院 江苏 连云港 222005)

[4)](江西财经大学信息管理与数学学院 南昌 330013)

[5)](江西财经大学计算机与人工智能学院 南昌 330013)

摘 要 社交媒体上谣言的泛滥对社会造成了严重的负面影响。随着多模态内容在社交媒体中的迅速增长,多模态谣言检测受到了越来越多的关注。目前,大多数方法主要聚焦于学习各个模态的特征,并通过特征融合实现不同模态信息的互补。然而,这些方法存在两个关键问题:(1)不同特征空间之间的跨模态关联难以有效捕捉图文细粒度语义的一致性;(2)单纯依赖图文内容难以识别一些造谣者精心设计的深层语义不匹配的谣言。为此,本文提出了融合证据与知识的场景图注意力网络的多模态谣言检测方法。首先,基于预训练的语言和视觉模型,分别提取文本语义和图像视觉特征,并通过误差级别分析提取图像篡改特征;其次,构建了一种基于反事实推理的无偏场景图生成方法和微调的 Flan-T5 模型,分别将图像和文本转化为视觉场景图和文本场景图,并利用知识蒸馏从知识库中提取场景图实体的相关知识,以增强模型对场景图的深层语义理解;接着,设计了一种融合场景关系特征的场景图注意力网络,以挖掘图文间的细粒度语义匹配特征;最后,从互联网中筛选与待检验帖子相关的文本和图片证据,并通过交叉注意力机制实现证据与待检验帖子的交互对齐,提升模型对深层语义不匹配谣言的识别能力。实验表明,在 Weibo 和 Twitter 两个真实社交网络数据集上,本文提出的方法在宏准确率上比最佳基线方法分别提高了 1.6%和 2.2%,而在谣言类别的 $F1$ 值上,分别提高了 2.6%和 3.0%。实验数据和代码已在 GitHub 上开源共享(https://github.com/xuejianhuang/SGKE)。

关键词 多模态谣言检测;场景图注意力网络;图文语义匹配;多模态证据对齐;知识增强

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2025.02159

Multi-Modal Rumor Detection with Scene Graph Attention Networks Integrating External Knowledge and Evidence

HUANG Xue-Jian[1),2)] MA Ting-Huai[2),3)] RONG Huan[2)]

WANG Gen-Sheng[4)] LIAO Guo-Qiong[1)] LIU De-Xi[5)]

[1)](School of VR Modern Industry, Jiangxi University of Finance and Economics, Nanchang 330013)

[2)](School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044)

[3)](School of Computer Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu 222005)

[4)](School of Information Management and Mathematics, Jiangxi University of Finance and Economics, Nanchang 330013)

[5)](School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330013)

Abstract The proliferation of rumors on social media has had a profound negative impact on

收稿日期:2024-11-14;在线发布日期:2025-05-16。本课题得到国家自然科学基金面上项目(62372243,62272206)、江西省自然科学基金青年项目(20242BAB20074)、江西省高校人文社会科学研究项目(JC24219)资助。**黄学坚**,博士研究生,讲师,中国计算机学会(CCF)会员,主要研究领域为谣言检测、多模态学习、自然语言处理。E-mail:huangxuejian@jxufe.edu.cn。**马廷淮(通信作者)**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为社交网络隐私保护、大数据挖掘、文本情感计算。E-mail:thma@nuist.edu.cn。**荣 欢**,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为社交媒体挖掘、社交网络内容安全、知识工程。**王根生**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为数据挖掘、社交网络。**廖国琼**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据库、数据挖掘、社会网络。**刘德喜**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为社会媒体处理、自然语言处理、计算心理学。

public discourse, social trust, and crisis management. The widespread use of digital platforms has further accelerated the dissemination of misinformation, particularly when multiple modalities—such as text, images, and videos—are strategically combined to increase persuasiveness or obscure deception. The rapid growth of such multimodal content has rendered cross-modal rumor detection both an urgent and intricate task. Specifically, multimodal rumor detection, which aims to determine the veracity of social media posts containing both textual and visual information, has garnered increasing attention from the research community. Existing state-of-the-art approaches predominantly focus on extracting features from each modality independently and fusing them at the feature level to achieve complementary information. However, these methods face two fundamental limitations: (1) they often struggle to capture fine-grained semantic consistency across heterogeneous modalities due to feature space misalignment, which hinders the detection of subtle manipulations or semantic inconsistencies; and (2) they rely mainly on surface-level content, making them insufficient for identifying sophisticated rumors that involve deep semantic mismatches or require external contextual knowledge for accurate interpretation. To address these challenges, we propose a novel multimodal rumor detection framework based on a Scene Graph Attention Network (SGAT), which integrates external knowledge and multi-source evidence to enhance semantic representation and reasoning capabilities. Specifically, we extract deep semantic features from text using pretrained language models and obtain rich visual representations using transformer-based vision models. Simultaneously, to identify potential image manipulation, we incorporate forensic features extracted via Error Level Analysis (ELA), which highlights signs of digital tampering such as splicing, cloning, or resampling. To model structured semantics across modalities, we introduce an unbiased scene graph generation method based on counterfactual reasoning and a fine-tuned Flan-T5 model. This transforms visual and textual inputs into structured scene graphs that explicitly encode objects, attributes, and their interrelationships, serving as interpretable intermediaries for cross-modal alignment. We further enrich semantic understanding through knowledge distillation from external knowledge graphs such as DBpedia, enabling the model to incorporate commonsense and domain-specific knowledge relevant to the entities and their interactions. Moreover, we design a Scene Graph Attention Network that integrates relational features from scene graphs and leverages cross-modal attention to achieve fine-grained semantic alignment between textual and visual modalities. This architecture allows the model to detect nuanced inconsistencies and misleading associations that may be imperceptible in raw feature spaces. To enhance factual verification, we retrieve relevant textual and visual evidence from external sources and align it with the target post using a cross-attention mechanism, thereby improving the model's ability to validate claims requiring additional context or background knowledge. Extensive experiments conducted on two real-world datasets—Weibo (Chinese) and Twitter (English)—demonstrate that our approach significantly outperforms competitive baselines. SGAT achieves improvements of 1.6% and 2.2% in macro accuracy, and gains of 2.6% and 3.0% in rumor-class F1 scores, respectively. Furthermore, our framework offers strong interpretability and robustness, making it well-suited for real-world applications such as content moderation and automated fact-checking. All code, datasets, and experimental protocols are publicly available at GitHub (https://github.com/xuejianhuang/SGKE), fostering transparency, reproducibility, and future research.

Keywords multimodal rumor detection; scene graph attention network; image-text semantic matching; multimodal evidence alignment; knowledge enhancement

1 引　言

随着移动互联网的快速发展,Twitter、Weibo和Facebook等社交媒体平台日益成为人们主要的信息交流渠道,它们在提升信息传播效率的同时,也为虚假谣言的扩散创造了条件。谣言在网络中的迅速传播不仅容易误导公众认知,干扰社会正常秩序,甚至可能对国家安全构成威胁[1]。例如,以新冠疫情为例,诸如"5G网络是病毒传播源"等虚假谣言,曾在网络广泛流传,引发了大量的误解与恐慌。为了应对这一问题,一些机构建立了专门的辟谣平台,如中国互联网联合辟谣平台,但当前大多数平台仍以人工方式进行内容核实,难以及时揭穿谣言。因此,研究有效的网络谣言自动检测方法,对于维护网络空间安全至关重要。

目前,大多数谣言检测方法主要集中在文本内容上。然而,在当今的富媒体环境中,谣言已不再局限于文本形式,而是通过图片、视频和音频等多种媒介融合呈现。这种多模态谣言利用人们"眼见为实"的心理效应,更容易被网民误信,进而在网络上迅速传播。多模态谣言检测面临更大的挑战,因为它需要评估每种模态及其组合语义的可信度。图1展示了常见的五种多模态谣言。图1(a)中的文本和图片均表明这可能是虚假的谣言信息;图1(b)中的文本没有迹象表明这是一条谣言,但图片显然是伪造的;图1(c)中的图片正常,但文本内容提示它可能是谣言;图1(d)和图1(e)中的图片和文本都不能直接证明它们是谣言信息,但图1(d)中的图片展示了洪灾,而文本描述的是地震,存在明显的语义不匹配;图1(e)中的图片和文本的浅层信息都与疫情相关,但其联合表达的深层语义却暗示COVID疫苗致死。如果能够获得关于COVID疫苗安全性的知识,以及图1(e)中的图片是来自其他新闻报道的拼接证据,则可以推断这可能是一个谣言。总体而言,图1(a)~(c)展示的谣言可通过不同模态的线索进行识别,而图1(d)则需要挖掘图文语义的不一致信息,图1(e)则需要融合背景知识和证据进行验证。

为了应对多模态谣言检测的挑战,一些研究者开始探索相关方法,利用如BERT[2]、GPT[3]、ResNet[4]和ViT[5]等预训练语言和视觉模型,从文本和图像中分别提取特征,并对不同模态的特征进行融合。相比于单模态方法,这些多模态方法充分利用了不同模态之间的互补信息,从而提高了图1(a)~(c)这类谣言检测的准确率。然而,由于图文特征空间的异构性存在跨模态语义鸿沟,这些跨模态的特征关联方法难以有效捕捉图1(d)所示的图文语义不一致的谣言信息。因此,文献[6-9]提出了基于图文语义相似度的多模态谣言检测方法,将异构的文本与图像特征投射到一个共享的潜在语义空间,并

(a) 文本和图片都是虚假的　(b) 带有正常文本的虚假图片　(c) 带有正常图片的虚假文本

(d) 图文浅层语义不匹配　(e) 图文深层语义不匹配

图1　多模态谣言的常见形式

利用空间距离的约束促进图文之间的语义对齐，从而识别两者语义不一致的谣言内容。然而，该方法所采用的对齐策略属于跨模态的粗粒度语义匹配，难以精准建模图文之间更细致的语义对应关系，且对于图 1(e) 这类精心设计的深层语义不匹配的谣言，难以从图文表层语义相似度的角度进行判断，而需要借鉴人类的判断思维，融合上下文相关的背景知识和证据进行推理。

因此，本文设计了一种结合外部知识与证据信息的多模态谣言检测方法。首先，利用 BERT 和 Swin Transformer 分别提取文本语义和图像视觉特征，并通过错误级别分析(Error Level Analysis，ELA)挖掘图片篡改特征。其次，构建基于反事实推理的无偏场景图生成方法和微调的 Flan-T5 模型，分别将图像和文本转化为视觉场景图和文本场景图，并利用知识蒸馏技术从知识库中提取相关背景知识，以强化对场景图的深层语义理解。接着，为了实现对场景图的有效学习，设计了一种融合关系特征的场景图注意力网络，以挖掘图文细粒度语义匹配特征。然后，从互联网中筛选与待检验帖子相关的文本和图片证据，并通过交叉注意力实现证据与待检验帖子的交互对齐，以挖掘证据验证特征。最后，将不同的特征进行融合，并输入多层感知机(Multi-layer Perceptron，MLP)进行分类。实验表明，该方法在 Twitter 和 Weibo 两个真实社交网络数据集上的表现优于所有对比的基线方法。本文的主要贡献如下：

(1)针对图文深层语义不匹配的谣言，提出了一种融合知识增强和证据验证的多模态谣言检测方法。利用知识蒸馏提取背景知识，增强模型对图文的语义理解，并通过从互联网中筛选文本和图像证据，提高模型的可解释性。

(2)针对图文异构特征空间的语义鸿沟导致难以有效捕捉图文细粒度语义一致性的问题，设计了一种基于场景图的图文细粒度语义对齐方法。同时，提出了一种融合关系特征的场景图注意力网络，以增强场景图之间的深层语义交互。

(3)在两个真实的社交网络数据集上进行了充分的实验分析。结果表明，本文提出的方法在准确率上比最先进的基线方法分别提高 1.7%和 2.8%。此外，本文方法不依赖于随传播时间变化的社交上下文信息(如元数据、转发和评论)，能够实现谣言的早期检测。

本文第 2 节介绍社交网络谣言检测的研究现状，第 3 节给出问题的定义，第 4 节阐述融合知识与证据的场景图注意力网络的多模态谣言检测模型，第 5 节通过对比实验、消融实验、误差分析和案例分析验证所提模型的有效性，第 6 节总结全文并指出方法所面临的挑战和不足。

2 相关工作

早期的网络谣言检测主要采用基于特征工程的机器学习方法，研究焦点集中在设计与谣言相关的特征，例如符号、情感词、文本长度和用户信息等[10]。然而，这些手工设计的特征受限于先验知识，因此在捕捉谣言的深层语义方面存在一定局限，缺乏全面性和灵活性。深度学习技术的不断演进促使研究者开始借助其自动提取谣言的语义特征，从而推动谣言检测向以数据驱动为核心的方向发展[11]。目前，大部分研究主要集中在单模态的文本内容，但在当今的富媒体环境中，图文并茂的谣言已成为主流[12]。因此，多模态谣言检测成为研究热点。单模态谣言检测主要侧重于处理单一类型的数据，如文本或图像，而多模态谣言检测则结合文本、图像和视频等多种数据类型，以捕捉更加丰富和准确的信息。接下来，本节将对单模态和多模态谣言检测的研究现状进行分析和总结。

2.1 单模态谣言检测

目前，单模态谣言检测主要依赖深度学习方法，从文本或图像中提取有价值的深层特征，以提升检测性能[13]。由于文本能够传递更丰富的信息，因此大多数研究致力于从文本内容中捕捉语义[14-15]、情感[16]和意图[17]等特征。例如，Alkhodair 等人[18]联合 Word2vec 与 LSTM 学习文本内容的语义特征，以检测和新兴主题相关的突发新闻谣言；Luvembe 等人[19]则提出了一种基于深度归一化的注意力机制，用于提取谣言的双重情感特征；苏兴等人[20]提出了一种基于层次门控交互融合网络的谣言检测方法。该方法首先利用一阶门控对原帖和评论的语义特征及情感特征进行增强，随后利用二阶门控对增强特征进行跨语义融合，以解决特征融合过程中由于不同特征之间差异引入噪声的问题；王友卫等人[21]针对当前大多数研究忽略了语法、情感及语言等特征的问题，设计了一种基于事件-词语-特征异质图的谣言检测模型，挖掘微博事件中蕴含的情感、语言和语法特征，从而提高微博事件表示的准确性。

造谣者常常模仿真实信息的表达方式,加之社交媒体上的帖子通常非常简短,导致仅依赖文本语义内容进行检测的方式在实际应用中效果受限[22]。为了解决这一挑战,部分研究开始在文本基础上融合其他补充性特征,如用户评论[23]、传播结构[24-25]和发文用户特征[26]等。例如,Wang 等人[27]提出了一种评论-内容编码网络,用于探索关键评论与新闻文本之间的内在关系,以揭示虚假新闻的关键评论;杨延杰等人[28]为有效利用转发结构信息和源帖的信息,提出了一种融合源信息与门控图神经网络的谣言检测方法;杨帆等人[29]则引入用户属性信息,以补充微博文本内容特征和传播结构特征,从而提高谣言检测的准确率。尽管引入用户评论和传播特征能够提高检测准确性,但这些特征通常需要谣言在社交平台上获得一定范围的传播后才能收集,因此无法实现谣言的早期检测[30]。

近年来,以 ChatGPT 为代表的大语言模型展现出了强大的自然语言理解和生成能力,在众多自然语言处理任务中取得了卓越的表现。在谣言检测领域,相关研究人员也开始探索大语言模型的潜力。例如,Yang 等人[31]利用 ChatGPT 作为文本增强、实体链接和文本概念化的辅助手段,成功展示了该方法在谣言检测任务中的有效性;柯婧等人[32]利用现有生成式大语言模型的总结与推理能力,提出了一种基于大语言模型隐含语义增强的虚假新闻检测方法;Zhang 等人[33]通过上下文学习检验大语言模型在虚假新闻验证方面的性能,发现仅通过 4 个示范性示例,一些提示方法的性能可以与先前的监督模型相媲美;Caramancion 等人[34]通过黑盒测试评估了 ChatGPT 3.5、ChatGPT 4.0、LaMDA 和 Bing AI 等几个知名大模型在判断新闻真实性方面的能力,研究结果显示,所有大模型都表现出中等水平的能力,其在理解新闻信息中的微妙语义差异和上下文方面仍存在一些不足,尚无法完全替代当前基于监督学习的谣言检测方法。

2.2　多模态谣言检测

在当今的富媒体环境中,社交媒体上传播的谣言已从简单的文本形式演变为融合文本、图像和视频的多模态形式[35]。由于这些多模态谣言具有更强的视觉冲击力,它们的传播速度更快、影响范围更广、危害程度更深。因此,如何自动检测多模态谣言已成为当前研究的热点[36]。目前,多模态谣言的检测主要侧重于结合文本和图像信息[37],通过预训练的语言和视觉模型分别提取文本和图像特征,然后对不同模态的特征进行交互与融合[38]。

在多模态交互方面,目前的方法主要通过交叉注意力机制促进文本和视觉特征的对齐[39],以探索文本和图像内容之间的语义一致性。例如,Zheng 等人[40]将文本、视觉和社交图特征融入多模态框架,并构建了注意力网络,以实现特征对齐交互,在多模态谣言检测中取得了良好效果;Wu 等人[41]通过堆叠多个交叉注意力层,分层学习不同模态之间的相互依赖关系;Hu 等人[42]设计了一个图像-文本匹配感知的交叉注意力网络,用于捕捉图像和文本的对齐,以实现更好的多模态交互。为了实现细粒度的模态交互,Xu 等人[43]将文本和图像转换为由单词、实体和图像块组成的异构图,利用分层图注意网络(Hierarchical Graph Attention Network, HAGT)捕捉模态内和跨模态的细粒度语义交互。针对已有模型通常从全局角度学习图文间的跨模态关联而忽视内部的局部差异,钟善男等人[44]构建了一种基于双分支线索感知与自适应协同优化机制的多模态虚假新闻识别模型。

在多模态融合方面,目前的方法主要通过不同的预训练模型提取文本和视觉信息,然后将这些信息融合在一起。常用的融合策略包括拼接[45]、相加[46]和门控[47]。例如,刘金硕等人[48]提出了一种融合图像、图像内嵌文本和配文内容的多模态网络谣言检测方法,利用 VGG-19 网络提取图像特征,DenseNet 提取图像内嵌文本特征,以及 LSTM 提取文本特征,最终通过拼接的方式将不同模态的特征整合在一起;Qian 等人[49]则分别利用 BERT 和 ResNet 提取文本和图像特征,采用多模态上下文注意力网络实现语义空间的映射,并通过拼接方式将不同模态特征融合;Zhou 等人[50]设计了一种对比式的文本-图像预训练模型,通过相加方式和模态特定注意力融合图像和文本特征。除了在特征层进行早期融合外,还有研究者探索了决策层的晚期融合。例如,Meel 等人[51]通过分析文本语义特征、图像篡改特征和图文相似度特征,分别进行虚假信息判断,并利用最大投票法集成所有判断结果。此外,一些方法[52-53]采用多任务学习作为融合策略,以推动不同类型特征之间的关联学习。

外部证据在事实核验中发挥着至关重要的作用[54]。因此,相关研究人员探索了基于证据检索的谣言检测方法[55]。例如,Abdelnabi 等人[56]通过互联网收集了文本和视觉两种模态的证据,并基于记忆网络结构设计了一致性检查网络。该网络能够学

习并验证待检验帖子文本与文本证据、帖子图片与视觉证据之间的一致性，从而有效地评估图文上下文不匹配(Out-of-Context，OOC)的虚假信息。为了提高模型的可解释性，Qi 等人[57]针对 OOC 虚假信息的检测，设计了一种多模态大型语言模型 SNIFFER。通过两阶段指令调优，SNIFFER 使现有的通用 MLLM 能够更好地适应 OOC 虚假信息的检测。同时，SNIFFER 通过证据检索，成功建模了内部的图像-文本线索和外部的声明-证据线索，从而实现了虚假信息的检测与结果解释的有机结合。为了在判断新闻真实性的同时揭示多模态虚假新闻中的潜在欺骗模式，Dong 等人[58]提出了一种神经符号潜在模型，将每种欺骗模式视为一个二值可学习的潜在变量，并通过基于符号逻辑规则的弱监督和变分推断方法进行学习。此外，考虑到仅通过新闻内容难以检测 OOC 虚假信息，他们利用图像反向搜索技术从网络中检索图像的上下文信息，作为补充证据。

综上所述，尽管多模态谣言检测已经取得了一定的进展，但仍面临一些亟待解决的问题。例如，在多模态交互方面，现有方法主要关注图文的表层语义，未能有效融合外部知识和证据，导致模型难以识别图文深层语义不匹配的谣言信息。在多模态融合方面，由于文本和图像特征空间的异构性，现有的跨模态融合方法难以有效捕捉图文细粒度语义的一致性。此外，当前基于证据检索的方法主要针对 OOC 虚假信息，未能充分考虑多模态谣言类别的多样性，从而限制了模型的应用范围和鲁棒性。为此，本文提出了一种融合外部知识与证据的多模态谣言检测方法。该方法通过整合外部知识和证据，增强模型识别深层语义不匹配谣言的能力，并将图像和文本转化为同构的场景图，通过改进的图注意力网络实现场景图对象间的深层语义交互，从而提升图文细粒度语义一致性学习的能力。

3 问题定义

谣言的定义可以分为广义和狭义两种。广义的谣言是指未经验证的信息，而狭义的谣言则是与事实不符的虚假信息。为了综合这两种定义，本文将待检验的帖子划分为未验证的谣言(根据现有信息无法验证真假)、谣言(虚假信息)和非谣言(真实信息)。根据检测对象的不同，谣言检测可划分为帖子级与事件级两种类型。前者侧重于判断单个帖子的真伪，后者则基于围绕同一主题的多个相关帖子，综合评估整个事件的真实性。本文的研究范围集中在帖子级别的检测。根据时间属性，谣言检测可以分为早期检测和滞后检测。滞后检测通常依赖于谣言的传播特征，因此需要谣言在社交平台上得到较为广泛的传播后才能进行检测。相比之下，早期检测不依赖于传播特征，而是从内容特征的角度进行判断，可以最大程度地降低谣言传播带来的危害，因此更具实际应用价值。本文的研究目标是实现谣言的早期检测，给定一组帖子集合 $P=\{p_1,p_2,\cdots,p_n\}$，$p_i=(T_i,I_i)$，其中 T_i 和 I_i 分别表示第 i 个帖子的文本和图片，学习一个模型 $f:P\rightarrow C$，输入任意帖子 p_i 的文本和图片信息，输出该帖子的类别信息 $C=\{N,R,U\}$，其中 N、R 和 U 分别表示非谣言、谣言和未验证的谣言。

4 模型构建

本文提出的融合外部知识与证据的场景图注意力网络多模态谣言检测方法，命名为 SGKE，其总体架构如图 2 所示。该方法主要包括四个模块：特征提取、图文证据验证、图文语义匹配和结果分类。特征提取模块基于预训练的 BERT 和 Swin Transformer，分别提取输入帖子的文本语义特征和图像视觉特征。同时，通过错误级别分析生成 ELA 图像，并利用 Swin Transformer 提取可能被篡改的图像特征。在图文证据验证模块，首先从互联网筛选出与输入信息相关的文本和图片证据。然后，通过交叉注意力(Cross-Attention)实现证据与待验证帖子的交互对齐，从而挖掘验证特征。在图文语义匹配模块中，首先将帖子的图像和文本分别转化为视觉场景图和文本场景图。接着，通过知识蒸馏从知识图谱中获取场景图中实体的概念知识，作为场景图的补充节点，以增强模型对场景图的语义理解。随后，设计了一种场景图注意力网络(Scene Graph Attention Network，SGAT)，对视觉场景图和文本场景图进行学习，并将学习到的特征输入语义相似度感知网络，以计算图文语义匹配特征。最后，结果分类模块将不同的特征进行融合，并输入分类器进行结果分类。

4.1 特征提取

4.1.1 文本特征

预训练语言模型在提取文本语义特征方面表现出色，已成为自然语言处理(NLP)任务的重要模块。预训练语言模型 BERT 基于 Transformer 的双向

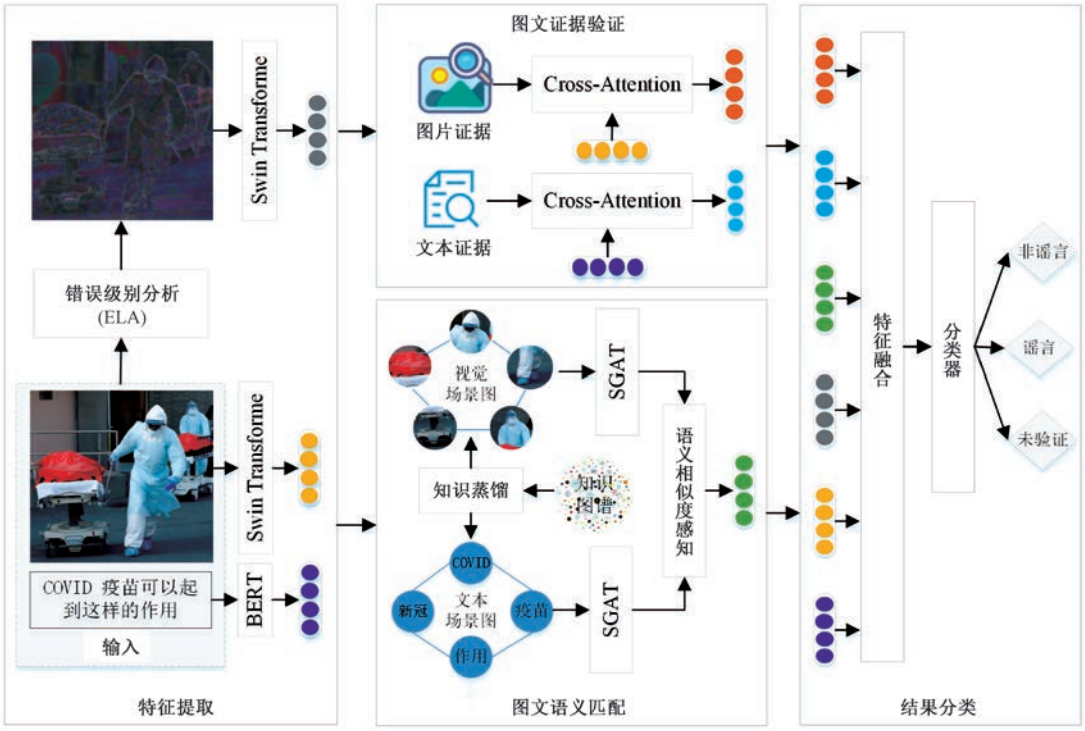


图 2　模型架构

编码器结构,利用自注意力机制显著提升了模型的学习能力和并行计算效率。与 Word2Vec 相比,BERT 在建模时充分融合上下文信息,有效缓解了词义歧义问题;此外,其分层结构能够提取多层次的语义表示,为后续任务提供了更具多样化的特征支持。为了增强对文本序列长距离依赖关系的捕捉能力,本文采用 BERT 与双向长短时记忆网络(Long Short-Term Memory, LSTM)相结合的方式提取文本的序列特征 $\overrightarrow{\boldsymbol{h}_t} \in \mathbb{R}^{l_t \times d_t}$,如公式(1)所示。其中,$T$ 为输入的文本序列,l 为文本序列长度,d_t 为文本特征维度。本文利用 $\overrightarrow{\boldsymbol{h}_t}$ 最后一个单元 $\overrightarrow{\boldsymbol{h}_{t,l}}$ 表示整个文本的语义特征 $\boldsymbol{h}_t \in \mathbb{R}^{d_t}$。

$$\overrightarrow{\boldsymbol{h}_t} = LSTM(BERT(T)) \tag{1}$$

4.1.2　图像特征

受 Transformer 在 NLP 领域取得巨大成功的影响,研究人员开始积极探索将 Transformer 引入计算机视觉(CV)领域。然而,将 Transformer 从 NLP 领域迁移到 CV 领域面临着视觉尺度变化大和图像分辨率高的两大挑战。为此,Liu 等人[59]提出了 Swin Transformer 这一网络架构,结合了层次化特征表示和滑动窗口自注意力机制。通过局部自注意力计算,Swin Transformer 显著降低了计算复杂度,能够高效处理大尺寸图像并提取多尺度特征,其架构如图 3 所示。

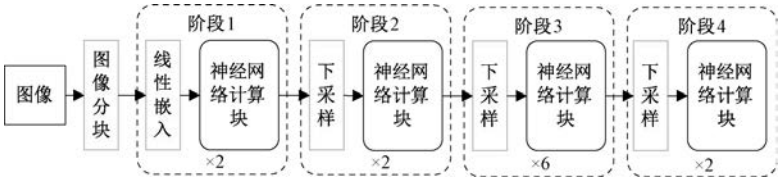


图 3　Swin Transformer 架构

本文采用在 ImageNet-1K 数据集上预训练的 Swin Transformer 来提取图像的区块特征 $\overrightarrow{\boldsymbol{h}_p} \in \mathbb{R}^{l_p \times d_p}$,如公式(2)所示。其中,$I$ 为输入的图像,l_p 为图像的区块数量,d_p 为图像的特征维度。本文采用所有区块特征的平均值作为整个图像的特征 $\boldsymbol{h}_p \in \mathbb{R}^{d_p}$。

$$\overrightarrow{\boldsymbol{h}_p} = SwinTransformer(I) \tag{2}$$

4.1.3　图像篡改特征

本文采用 ELA 算法提取图像的篡改特征,具体过程如算法 1 所示。ELA 的核心原理是利用图

像在压缩过程中的损失特性。未经篡改的图像区域在压缩后会显示出相似的损失特征，而篡改区域则可能因来源不同而表现出不同的错误水平。图 4 展示了一个基于 ELA 的篡改检测示例。获得 ELA 图像后，利用预训练的 Swin Transformer 提取 ELA 图像的特征 $\boldsymbol{h}_e \in \mathbb{R}^{d_p}$，如公式(3)所示。其中，$I_e$ 为输入的 ELA 图像，$Mean$ 表示对所有区块特征求取平均值。

$$\boldsymbol{h}_e = Mean(SwinTransformer(I_e)) \quad (3)$$

算法 1. ELA

输入：原始图像 o_img

输出：ELA 图像 e_img

$c\_img \leftarrow Compressed(o\_img)$：将原始图像 o_img 以一定的压缩率重新保存为 JPEG 格式，得到压缩后的图像 c_img；

$diff \leftarrow Difference(o\_img, c\_img)$：比较原始图像 o_img 与重新压缩后的图像 c_img 之间的差异，得到差异图像 $diff$；

$e\_img \leftarrow Enhance(diff)$：对差异图像 $diff$ 进行增强处理，并将其转化为 ELA 图像 e_img；

Return e_img：输出最终的 ELA 图像；

(a) 原始图像 (b) 压缩后的篡改图像 (c) ELA图像

图 4 基于 ELA 的篡改检测示例

4.2 图文证据验证

如图 5 所示，首先利用帖子中的图片，通过 Google 反向图像搜索 API 查找与该图像相关的新闻报道，并采用自动化过滤策略，对返回的网页进行多重验证(如比对假新闻网站列表和语言检测)，从中筛选出可靠的文本证据。接着，使用 Google 可编程搜索引擎，以帖子中的文本作为查询条件，自动检索与之相关的图片，并依据预设规则过滤掉来自虚假信息网站的图像，得到视觉证据。通过互联网获取的证据信息融合了世界知识，有助于识别图 1(e)这类造谣者精心设计的图文深层语义不匹配的谣言。

图 5 证据检索

4.2.1 证据检索

首先，使用输入帖子的图片作为查询条件，通过 Google 反向图片搜索引擎检索文本证据。搜索引擎返回一系列与查询图片相似的图像及包含这些图像的网页 URL，初步保留前 20 条记录。随后，根据维基百科列出的假新闻网站列表，过滤掉出现在该列表中的 URL，并通过 fastText 库进行语言识别，丢弃网页标题非英语和中文的页面。接着，利用网页爬虫收集过滤后的网页的标题和图片描述信息(即<img>和<figcaption>标签的文本属性)，并去除冗余的文本片段，最终形成文本证据集 $TE = \{te_1, te_2, \cdots, te_k\}$，其中 te_i 表示第 i 条文本证据，k 为证据的总数。然后，利用 BERT 提取证据的语义特征表示 $\overrightarrow{\boldsymbol{h}_{te}} \in \mathbb{R}^{k \times d_t}$。

同样，使用输入帖子的文本作为查询条件，通过 Google 可编程搜索引擎来检索相关的图片，并过滤掉来自虚假信息网站的图像，得到图片证据集合 $VE = \{ve_1, ve_2, \cdots, ve_k\}$，其中 ve_i 为第 i 个图片证据，k 为证据的总数。然后，使用 Swin Transformer 提取图片证据特征 $\overrightarrow{\boldsymbol{h}_{ve}} \in \mathbb{R}^{k \times d_p}$。

4.2.2 证据对齐

本文利用交叉注意力机制(Cross-Attention)实现证据和待检测帖子内容的对齐。交叉注意力机制能够实现两个不同输入序列间的信息交互，有效地将来自不同来源的上下文进行对齐，从而更好地捕

捉两个输入之间的相关性,其计算流程如图 6 所示。

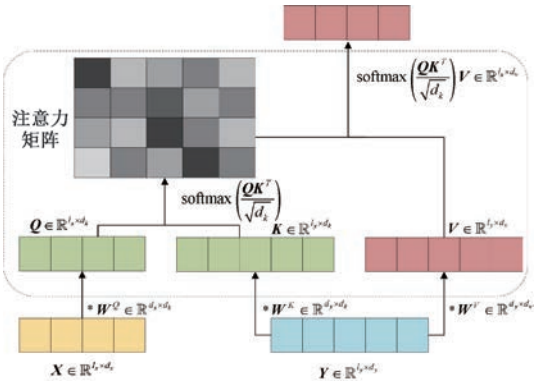


图 6　交叉注意力计算流程

多头交叉注意力是对 Cross-Attention 的扩展,通过并行计算多个注意力头,模型能够从多个角度关注输入信息,从而增强其表达能力。其计算如公式(4)～(6)所示。

$$MulCoA(\boldsymbol{X},\boldsymbol{Y})=Concat(head_1,\cdots,head_h)\boldsymbol{W}^O \tag{4}$$

$$head_i=cross-attention(\boldsymbol{X}\boldsymbol{W}_i^Q,\boldsymbol{Y}\boldsymbol{W}_i^K,\boldsymbol{Y}\boldsymbol{W}_i^V) \tag{5}$$

$$cross-attention(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)=softmax(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^T}{\sqrt{d_k}})\boldsymbol{V}_i \tag{6}$$

其中,$\boldsymbol{X}\in\mathbb{R}^{l_x\times d_x}$ 和 $\boldsymbol{Y}\in\mathbb{R}^{l_y\times d_y}$ 表示两个不同的输入序列(X 作为查询序列,Y 作为键-值序列),l_x 和 l_y 表示对应的序列长度,d_x 和 d_y 表示对应的特征维度。$\boldsymbol{Q}_i=\boldsymbol{X}\boldsymbol{W}_i^Q\in\mathbb{R}^{l_x\times d_k}$ 表示查询(Query),$\boldsymbol{K}_i=\boldsymbol{Y}\boldsymbol{W}_i^K\in\mathbb{R}^{l_y\times d_k}$ 表示键(Key),$\boldsymbol{V}_i=\boldsymbol{Y}\boldsymbol{W}_i^V\in\mathbb{R}^{l_y\times d_v}$ 表示值(Value)。$\boldsymbol{W}_i^Q\in\mathbb{R}^{d_x\times d_k}$,$\boldsymbol{W}_i^K\in\mathbb{R}^{d_y\times d_k}$ 和 $\boldsymbol{W}_i^V\in\mathbb{R}^{d_y\times d_v}$ 表示第 i 个头的参数矩阵。

本文将帖子的文本的序列特征 $\overrightarrow{\boldsymbol{h}_t}\in\mathbb{R}^{l_t\times d_t}$ 和文本证据特征 $\overrightarrow{\boldsymbol{h}_{te}}\in\mathbb{R}^{k\times d_t}$ 分别作为多头交叉注意力的输入序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$。同样,将帖子的图像特征 $\overrightarrow{\boldsymbol{h}_p}\in\mathbb{R}^{l_p\times d_p}$ 和图片证据特征 $\overrightarrow{\boldsymbol{h}_{ve}}\in\mathbb{R}^{k\times d_p}$ 作为另一个多头交叉注意力的输入序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$。最终,对输出序列特征进行平均,得到文本证据对齐特征 $\boldsymbol{h}_{te}\in\mathbb{R}^{d_t}$ 和图片证据对齐特征 $\boldsymbol{h}_{ve}\in\mathbb{R}^{d_p}$,如公式(7)和(8)所示。

$$\boldsymbol{h}_{te}=Mean(MulCoA(\overrightarrow{\boldsymbol{h}_t},\overrightarrow{\boldsymbol{h}_{te}})) \tag{7}$$

$$\boldsymbol{h}_{ve}=Mean(MulCoA(\overrightarrow{\boldsymbol{h}_p},\overrightarrow{\boldsymbol{h}_{ve}})) \tag{8}$$

4.3　图文语义匹配

4.3.1　场景图构建

由于从预训练语言模型和视觉模型中提取的文本和图像语义特征分布在不同的特征空间,存在语义鸿沟,直接在这些异构空间中计算语义相似度的难度较大。因此,本文将图像和文本转化为视觉场景图和文本场景图,从而在同构的场景图空间中计算它们的语义相似度。

(1)视觉场景图生成

视觉场景图是一种将图像结构化表示的方法,能够清晰地表达图像中的对象及其之间的关系,从而实现对图像的高层次理解和更细粒度的推理,这使其成为当前的研究热点。本文采用一种基于 Transformer 的轻量级场景图生成方法 EGTR[60],来生成视觉场景图。EGTR 的总体架构如图 7 所示。

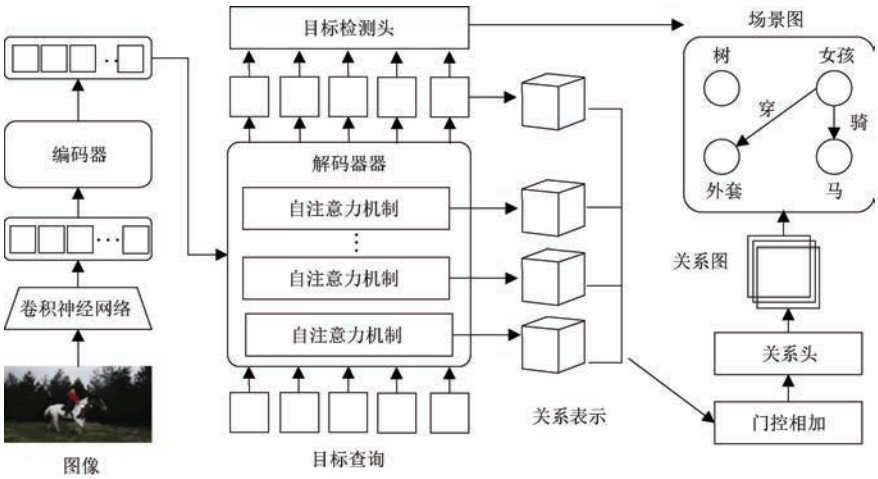


图 7　场景图生成方法 EGTR 的总体架构

然而,由于数据集标签呈现出高度偏向的长尾分布,生成的场景图往往存在偏见问题。例如,原本信息丰富的谓语(如“人在海滩上行走”、“坐在海滩上”或“躺在海滩上”)可能会被简单归纳为“人在海

滩上”。为了解决这一问题，本文在 EGTR 模型的基础上引入了一种基于反事实推理的去偏场景图生成方法（如图 8 所示），其核心思想源自因果推理，通过模拟干预揭示物体对象视觉特征对关系预测的因果影响。

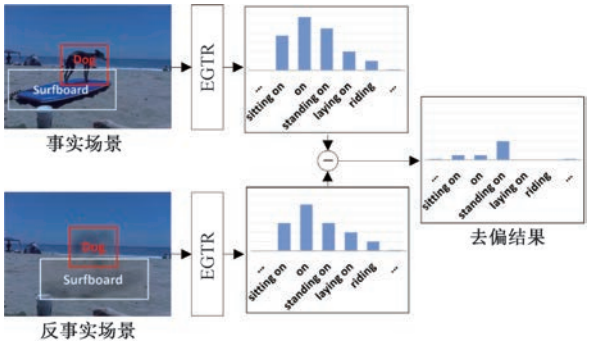


图 8 基于反事实推理的去偏场景图生成方法

具体来说，从因果图的角度对场景图生成过程进行建模[61]，如图 9 所示。其中，每个节点代表一个关键变量，其中 I 表示整幅输入图像的特征、O 表示物体对象的视觉特征、Z 表示物体对象的标签信息、Ra 表示物体对象之间的谓语关系，有向边表示节点间存在因果关系。例如，$I \rightarrow Ra$、$O \rightarrow Ra$ 和 $Z \rightarrow Ra$ 表示谓语关系 Ra 是由 I、O 和 Z 三者共同作用产生的结果。其中，视觉特征 O 是影响关系预测的内因（主要因素），而 I 和 Z 则是外因[62]。

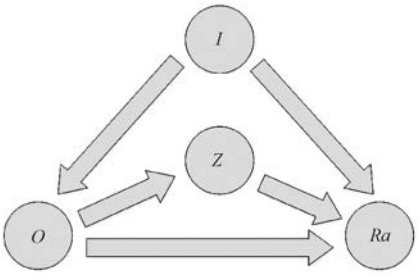


图 9 因果图

为了准确识别视觉特征 O 在关系推理中的核心因果贡献，我们实施因果干预，对原始（事实）场景中的每个对象进行预处理，保留对象的语义标签 Z 和上下文特征 I，同时将视觉特征屏蔽，从而构造出一个反事实场景。该反事实场景实际上相当于在因果图中对视觉特征变量进行了干预，使其不再发挥作用。接下来，分别将事实场景和反事实场景输入预训练的 EGTR 模型，通过多层 Transformer 捕捉对象间的交互关系，分别预测出两种场景下的关系分布 P_{fact} 和 P_{counter}。随后，设计一个差值计算模块，对两种预测结果分布进行差值计算 $D = P_{fact} - P_{\text{counter}}$。这一差值信号反映了视觉特征在关系预测中的因果作用，揭示了在干预下预测结果的变化，并可作为校正因子对原始预测结果进行调整，从而消除长尾数据分布所带来的偏向影响。经过去偏的 EGTR 处理后，生成的视觉场景图表示为 $G^v = \{V^v, E^v\}$，其中 $V^v = \{v_1, v_2, \cdots, v_n\}$ 表示由对象组成的节点集，$E^v = \{e_1, e_2, \cdots, e_m\}$ 表示对象之间关系的边集。每个对象 $v_i = \{c_i, a_i, f_i\}$ 由一个对象类别标签 c_i、属性标签 a_i 和对象的框坐标 f_i 组成，而每个关系 $e_j = (s_j, r_j, o_j)$ 则表示为一个三元组，其中 s_j 和 o_j 分别表示首尾对象实体，r_j 表示关系类别标签。

（2）文本场景图生成

文本场景图解析是将自然语言文本转换为结构化的场景图表示，以捕捉文本中提到的对象、属性及其相互关系，从而使模型更好地理解文本表达的语义。该解析过程通常涉及对文本的语法和语义分析，通过识别名词、动词及其他相关成分来构建场景图的节点和边。本文采用在 FACTUAL 数据集上微调的 Flan-T5 模型进行文本场景图解析[63]。Flan-T5 是 Google 发布的一款具有强大泛化能力的大语言模型，能够高效地完成下游任务的微调。FACTUAL 数据集对关系、对象和属性有严格且一致的定义，有助于提升场景图解析模型的质量和准确性。经过微调的 Flan-T5 处理后，得到文本场景图 $G^t = \{V^t, E^t\}$，其中节点集合 V^t 和边集合 E^t 与视觉场景图 G^v 的含义一致，但文本场景图中的对象不包含框坐标。图 10 展示了一个文本场景图解析的示例。

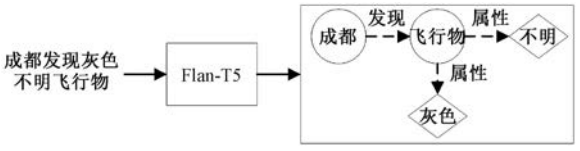


图 10 文本场景图解析示例

4.3.2 知识蒸馏

知识蒸馏的目的是从外部大型知识图谱中提取与场景图实体相关的概念知识，并将这些知识作为补充节点融入场景图，从而提升模型对场景图语义的理解能力。与传统的模型压缩技术不同，本文中的知识蒸馏主要用于丰富场景图的语义信息。其具体过程如图 11 所示。首先，利用实体链接技术（例如 EDEL[64] 和 Rel-Norm[65]），将文本中模棱两可的实体提及关联到知识图谱中对应的实体。该技术首

先识别出文本中的潜在实体,并借助消歧算法[66]消除同名或相似名称之间的歧义,确保每个实体均能被精确映射到知识图谱中的唯一实体。接着,采用概念化技术[67],从知识图谱中抽取出与已识别实体相关的概念知识。这些概念通常包括实体的定义、所属类别、主要属性以及与其他实体之间的关系等信息,为场景图提供更为丰富的语义背景。例如,对于文本"电联表示 5G 不会传播 COVID-19",首先通过实体链接技术获得集合 $ES=\{5G,电联,COVID\text{-}19\}$。然后,以 isA 关系为例,对集合中的每个实体进行概念化处理,分别得到实体概念 $EC_{5G}=\{移动网络,通讯网络\}$、$EC_{电联}=\{机构组织,国际电信联盟\}$和 $EC_{COVID\text{-}19}=\{新冠病毒,病毒\}$。随后,在场景图中,为每个实体添加相应的补充概念节点,并建立实体与概念之间的关系边,以增强场景图的语义表达。

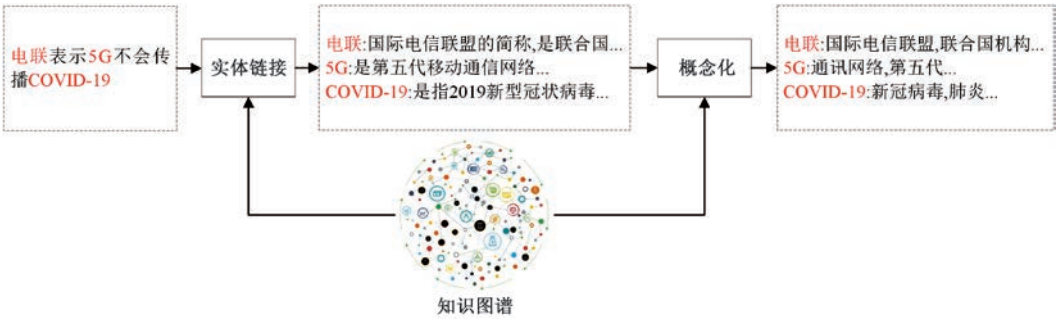


图 11　知识蒸馏

4.3.3　场景图注意力网络

为了实现对场景图的有效学习,本文设计了一种场景图注意力网络(Scene Graph Attention Network, SGAT)。SGAT 是在图注意力网络(Graph Attention Network, GAT)的基础上进行改进的。GAT 通过引入注意力机制增强节点之间的交互,其核心思想是自适应地为不同邻居节点分配权重,以突出重要节点并抑制不重要节点,从而实现灵活的特征聚合。然而,GAT 在计算注意力权重时未考虑节点之间的关系特征,可能导致在处理包含复杂关系的场景图时存在不足。因此,本文提出了融合关系特征的 SGAT,在计算注意力权重和特征聚合时引入关系特征,并增加残差结构,以减轻过平滑问题,其计算如公式(9)。

$$SGAT=\begin{cases}\boldsymbol{h}'_{v_i}=\Big\|_{h=1}^{H}\Big(\sum\limits_{j\in N(v_i)}\alpha_{j,i}^{h}(\boldsymbol{W}_n^h\cdot\boldsymbol{h}_{v_j}+\boldsymbol{W}_e^h\cdot\boldsymbol{h}_{e_{j,i}})+\boldsymbol{W}_s^h\cdot\boldsymbol{h}_{v_i}\Big)\\ e^h=\mathrm{LeakyReLU}(\boldsymbol{a}^T[\boldsymbol{W}_n^h\cdot\boldsymbol{h}_{v_i}\,\|\,\boldsymbol{W}_n^h\cdot\boldsymbol{h}_{v_j}\,\|\,\boldsymbol{W}_e^h\cdot\boldsymbol{h}_{e_{j,i}}])\\ \alpha_{j,i}^h=\mathrm{softmax}(e^h)\end{cases}\tag{9}$$

其中,$\boldsymbol{h}'_{v_i}$ 表示节点 v_i 聚合更新后的特征,$\|$ 表示拼接操作,H 表示多头注意力的个数,$\alpha_{j,i}^h$ 表示在第 h 个注意力类型中邻接节点 v_j 对节点 v_i 的注意力权重,$N(v_i)$ 表示节点 v_i 的邻接节点集合,$\boldsymbol{W}_n^h$、$\boldsymbol{W}_e^h$ 和 $\boldsymbol{W}_s^h$ 表示第 h 个头的可学习参数矩阵,$\boldsymbol{h}_{v_i}$、$\boldsymbol{h}_{v_j}$ 和 $\boldsymbol{h}_{e_{j,i}}$ 分别表示节点 v_i、v_j 和边 $e_{j,i}$ 的特征表示。

通过多层 SGAT 对知识增强的视觉场景图 G^v 和文本场景图 G^t 进行学习,得到图的特征表示 $\boldsymbol{h}_g^v\in\mathbb{R}^{n_v\times d}$ 和 $\boldsymbol{h}_g^t\in\mathbb{R}^{n_t\times d}$,如公式(10)和(11)所示,其中节点的初始特 $\boldsymbol{H}_n^v$ 和 $\boldsymbol{H}_n^t$,边的初始特征 $\boldsymbol{H}_e^v$ 和 $\boldsymbol{H}_e^t$ 为预训练语言模型提取的语义向量,n_v 和 n_t 分别表示视觉和文本场景图的节点数量。然后,设计一种基于交叉注意力的双向协同注意力机制(Co-Attention),即将视觉场景图特征 $\boldsymbol{h}_g^v$ 和文本场景图特征 $\boldsymbol{h}_g^t$ 交替作为交叉注意力的查询(Query)输入,另一个特征作为键(Key)和值(Value)输入,如图 12 所示。Co-Attention 的核心思想是通过注意力机制,使一种模态的特征能够基于另一种模态的特征进行动态加权,从而实现模态间的信息融合与对齐。最后,对协同注意力的输出序列进行平均池化,并输入全连接神经网络,得到语义一致性特征 $\boldsymbol{h}_g\in\mathbb{R}^d$,如公式(12)和(13)所示。其中,$\sigma$ 表示激活函数,$\boldsymbol{W}_g$ 表示可以学习的参数矩阵。图文语义匹配的完整过程如算法 2 所示。

$$\boldsymbol{h}_g^v=SGAT(G_v,\boldsymbol{H}_n^v,\boldsymbol{H}_e^v)\tag{10}$$

$$\boldsymbol{h}_g^t=SGAT(G_t,\boldsymbol{H}_n^t,\boldsymbol{H}_e^t)\tag{11}$$

$$\ddot{\boldsymbol{h}}_g=co\text{-}attention(\boldsymbol{h}_g^v,\boldsymbol{h}_g^t)\tag{12}$$

$$\boldsymbol{h}_g=\sigma(\boldsymbol{W}_g(mean(\ddot{\boldsymbol{h}}_g)))\tag{13}$$

算法 2. 图文语义匹配

输入:帖子的文本 T 和图片 I

输出:文本 T 和图片 I 的语义匹配特征 $\boldsymbol{h}_g$

$G^v\leftarrow EGTR(I)$:利用去偏的 EGTR 算法把图片 I 转化为

视觉场景图 $G^v=\{V^v,E^v\}$；

$G^t \leftarrow Flan-T5(T)$：利用微调的 Flan-T5 模型把文本 T 转化为文本场景图 $G^t=\{V^t,E^t\}$；

$KnowledgeDistillation(KG)$：从知识图谱中获取实体的语义概念，并融合到场景图中；

$\boldsymbol{H}_n^v,\boldsymbol{H}_n^t \leftarrow BERT([V^v,V^t])$ $\boldsymbol{H}_e^v,\boldsymbol{H}_e^t \leftarrow BERT([E^v,E^t])$：利用 BERT 获取场景图节点和边的初始特征表示；

$\boldsymbol{h}_g^v \leftarrow SGAT(G^v,\boldsymbol{H}_n^v,\boldsymbol{H}_e^v)$ $\boldsymbol{h}_g^t \leftarrow SGAT(G^t,\boldsymbol{H}_n^t,\boldsymbol{H}_e^t)$：利用 SGAT 对场景图 G^v 和 G^t 进行学习，得到特征 $\boldsymbol{h}_g^v$ 和 $\boldsymbol{h}_g^t$；

$\ddot{\boldsymbol{h}}_g \leftarrow co\text{-}attention(\boldsymbol{h}_g^v,\boldsymbol{h}_g^t)$：利用双向协同注意力学习视觉场景图和文本场景图的细粒度语义交互，得到特征 $\ddot{\boldsymbol{h}}_g$；

$\boldsymbol{h}_g \leftarrow FCNN(mean(\ddot{\boldsymbol{h}}_g))$：对 $\ddot{\boldsymbol{h}}_g$ 进行平均池化，并输入全连接神经网络，得到图文语义匹配特征 $\boldsymbol{h}_g$；

Return $\boldsymbol{h}_g$；

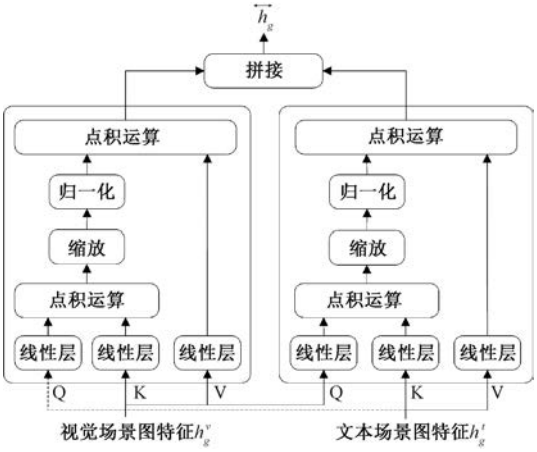


图 12 协同注意力的计算流程

4.4 结果分类

在获得文本语义特征 $\boldsymbol{h}_t$、图像视觉特征 $\boldsymbol{h}_p$、图像篡改特征 $\boldsymbol{h}_e$、图文语义匹配特征 $\boldsymbol{h}_g$、文本证据对齐特征 $\boldsymbol{h}_{te}$ 和图片证据对齐特征 $\boldsymbol{h}_{ve}$ 后，将它们进行拼接得到最终的融合特征 $\boldsymbol{F}$。接着，将融合特征 $\boldsymbol{F}$ 输入多层感知机(MLP)，最后通过 softmax 层输出分类结果，如公式(14)所示。

$$\hat{y}=softmax(\boldsymbol{W}*MLP(\boldsymbol{h}_t\,||\,\boldsymbol{h}_p\,||\,\boldsymbol{h}_e\,||\,\boldsymbol{h}_g\,||\,\boldsymbol{h}_{te}\,||\,\boldsymbol{h}_{ve})+\boldsymbol{b}) \tag{14}$$

其中，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别是线性层的参数和偏置项。本文基于最小化交叉熵损失函数对模型进行训练，其计算如公式(15)所示。

$$L=-\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}y_{i,c}\log(\hat{y}_{i,c}) \tag{15}$$

其中，M 表示样本数，C 表示类别数，$y_{i,c}$ 表示样本 i 在类别 c 上的实际标签(若样本属于类别 c 则为 1，否则为 0)，$\hat{y}_{i,c}$ 为模型预测样本 i 属于类别 c 的概率。

4.5 时间复杂度分析

整体模型的时间复杂度主要由各个模块的计算开销决定。在特征提取模块中，BERT 和 Swin Transformer 均基于 Transformer 架构，其主要计算瓶颈在于自注意力机制，其时间复杂度通常为 $O(L^2\cdot d+L\cdot d^2)$，其中 L 为输入序列长度，d 为特征维度。在图文证据验证模块中，主要计算开销来自多头交叉注意力网络。假设两个输入序列平均长度为 L，特征维度为 d，并将 d 分为 H 个头(每个头的维度为 $d_k=d/H$)，则生成查询、键和值的线性变换时间复杂度为 $O(L\cdot d_k^2)$，而查询与键的点积计算以及注意力矩阵与值的点积计算的时间复杂度均为 $O(L^2\cdot d_k)$，从而多头交叉注意力模块的总体时间复杂度大致为 $O(H(L^2\cdot d_k+L\cdot d_k^2))$。

在图文语义匹配模块中，SGAT 的时间复杂度主要依赖于图中节点数 $|V|$ 和边数 $|E|$ 以及特征维度 d。首先，对节点和边进行线性变换的时间复杂度为 $O((|V|+|E|)\cdot d^2)$，随后计算每条边的注意力权重的时间复杂度为 $O(|E|\cdot d)$，因此单层 SGAT 的总体时间复杂度为 $O((|V|+|E|)\cdot d^2+|E|\cdot d)$。若采用多头注意力(头数为 H)并堆叠 L 层，则 SGAT 模块的整体时间复杂度约为 $O(H\cdot L((|V|+|E|)\cdot d^2+|E|\cdot d))$。此外，Co-Attention 模块由两个多头交叉注意力网络构成，其时间复杂度为 $O(2H(L^2\cdot d_k+L\cdot d_k^2))$。在结果分类模块中，2 层感知机的时间复杂度为 $O(d^2)$。考虑到实际文本长度、图像 patch 数量以及场景图节点数通常有限，因此整体模型的时间复杂度处于可控范围内。

5 实验分析

在本节中，将详细介绍实验部分的相关内容，以验证本文提出的 SGKE 方法的性能。首先，将描述所使用的两个实验数据集：Twitter 和 Weibo。其次，将介绍实验的具体设置和评估指标。然后，将展示实验结果并进行深入分析，包括超参数实验、与基线方法的对比实验、消融实验和误差分析。最后，通过案例分析，直观地展示 SGKE 方法的优势和局限性。

5.1 实验数据

实验采用 Hu 等人[68]最新公布的多模态谣言数据集 MR^2。该数据集包含英文和中文两个谣言数据集。英文数据集由 Twitter 上的帖子构建，而

中文数据集则来源于 Weibo 上的帖子。数据集中的帖子被分类为谣言、非谣言和未验证的谣言(根据现有信息无法验证真假)。每个帖子包含一段文本和对应的一张图片。两个数据集的统计信息如表 1 所示。

表 1　实验数据集的统计信息

统计	Weibo	Twitter
样本总数量	7699	6102
谣言样本数量	1775	1288
非谣言样本数量	2425	1999
未验证的谣言数量	3499	2815
文本平均长度	66	67

同时,图 13 展示了这两个数据集的主题分布词云。词云显示,Weibo 数据集主要涉及公共健康和社会新闻,而 Twitter 数据集则关注政治、公共健康和社会新闻等话题。这表明 Weibo 和 Twitter 上传播的谣言在领域上存在一定的重叠。

图 13　Weibo 和 Twitter 数据集的主题对比

5.2　实验设置与评价指标

我们将数据集按 8∶1∶1 的比例划分为训练集、验证集和测试集。在特征提取模块,采用多语言预训练的语言模型 bert-base-multilingual-uncased 提取文本特征,文本的最大长度设置为 40。图像特征则使用在 ImageNet-1K 数据集上预训练的微型 Swin Transformer 提取。在 ELA 算法中,JPEG 压缩的质量级别设为 90%。在图文语义匹配模块中,采用去偏的轻量级场景图生成方法 EGTR 构建视觉场景图,同时利用在 FACTUAL 数据集上微调的 Flan-T5 模型生成文本场景图。在知识蒸馏过程中,采用 DBpedia 知识图谱,并通过 Spotlight 实现实体链接和概念化。SGAT 的层数和多头注意力个数均设置为 2。在图文证据验证模块,文本证据和图片证据的最大数量为 5。用于证据对齐的 Cross-Attention 的多头个数设为 8,所有特征的输出维度统一为 768。在结果分类模块中,使用了一个两层的感知机。在模型训练过程中,采用 AdamW 优化器,初始学习率设为 5e-5,批量大小为 32,训练轮次为 10。学习率的动态调整采用 ExponentialLR 策略,在训练初期保持较大的学习率以促进快速收敛,而在训练后期减小学习率,以实现更优的收敛效果。为减轻过拟合问题,训练过程中引入了早停策略,耐心值设置为 3 个轮次。实验方法基于 PyTorch 深度学习框架实现,具体实验代码可见 GitHub。模型及其训练的主要超参数详见表 2。

表 2　超参数设置

参数类型	参数名称	参数值
模型参数	特征维度 d	768
	文本最大长度	40
	SGAT 层数	2
	SGAT 多头注意力个数	2
	图文证据最大数量	5
	Cross-Attention 多头个数	8
	感知机层数	2
	LSTM 层数	2
训练参数	初始学习率	5e-5
	Batch size	32
	最大训练轮次	10
	早停等待轮次	3
	优化器 Optimizer	AdamW
	学习率动态调整策略	ExponentialLR

实验采用宏准确率(M_Acc)、每个类别的精确率(Pre)、召回率(Rec)和 F1 值作为方法性能评价指标,其计算如公式(16)～(20)所示。

$$\mathrm{M\_Acc}=\frac{1}{C}\sum_{c=1}^{C}Acc_c \tag{16}$$

$$\mathrm{Acc}_c=\frac{TP_c}{N_c} \tag{17}$$

$$\mathrm{Pre}_c=\frac{TP_c}{TP_c+FP_c} \tag{18}$$

$$\mathrm{Rec}_c=\frac{TP_c}{TP_c+FN_c} \tag{19}$$

$$\mathrm{F1}_c=\frac{2\times Pre_c\times Rec_c}{Pre_c+Rec_c} \tag{20}$$

其中,c 表示特定类别,C 为类别总数,N_c 为类别 c 的总样本数,TP_c、TN_c、FP_c 和 FN_c 分别表示类别 c 的真阳性、真阴性、假阳性和假阴性的数量。

5.3　实验结果与分析

5.3.1　超参数实验

图文证据的最大数量和 SGAT 网络的层数是两个关键的超参数。为了确定最大证据数量,我们将检索到的最大图文证据数量从 1 逐渐增加到 10,其他参数保持与表 2 一致,并在 Weibo 和 Twitter 数据集上观察宏准确率的变化,如图 14(a)所示。

结果表明，当图文证据数量为 5 时，方法的宏准确率最高，而过多或过少的证据均会影响性能。证据数量不足时，无法为待检测的帖子提供充足的解释性信息；而证据数量过多则可能引入不相关的噪声，进而影响方法的整体性能。为了确定 SGAT 网络的层数，我们分别将其层数设置为 1、2 和 3，其他参数同样保持与表 2 一致，比较结果如图 14(b)所示。当层数为 2 时，方法的性能最佳。这是因为当 SGAT 网络的层数过少时，卷积的感受野范围受限，导致学到的节点特征不够充分。而层数过多时，感受野的范围变得过大，不同节点间的覆盖区域重叠较多，容易出现过平滑现象。

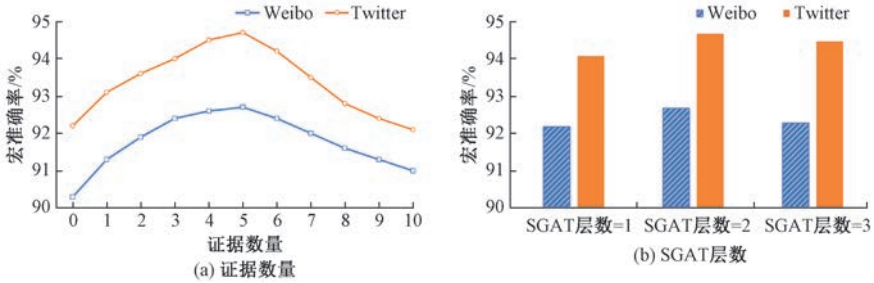


图 14 超参数实验

5.3.2 对比实验

为了验证本文方法的有效性，我们选择以下 12 个谣言检测方法作为对比基线：

(1) LSTM_Word2vec[18] (2020)：结合 LSTM 和 Word2vec 挖掘文本的谣言语义特征。

(2) TDRD[13](2020)：在文本特征中融合主题信息。

(3) MCAN[41] (2021)：通过交叉注意力融合文本和视觉特征。

(4) HMCAN[49] (2021)：通过分层多模态上下文注意力网络挖掘图文信息。

(5) CCN[56] (2022)：基于外部证据与待检验帖子内容一致性检查的方法。

(6) DC-CNN[14] (2023)：通过双通道卷积神经网络增强上下文语义的学习能力。

(7) DEDA[19] (2023)：通过注意力机制融合文本情感特征。

(8) MTTV[37] (2023)：通过多模态 Transformer 融合图像和文本的语义信息。

(9) HESN[36] (2023)：通过文本实体和视觉实体增强图像语义和知识语义。

(10) ITS[6] (2024)：基于文本信息和视觉信息相似度的方法。

(11) HGA-MMRD[43] (2024)：利用层次图注意力网络捕捉模态内和跨模态的语义交互。

(12) MMCAN[42] (2024)：基于知识蒸馏的多模态匹配感知方法。

此外，实验还对比了 ChatGPT 3.5 和 ChatGPT 4o 这两种零样本学习方法。对于 ChatGPT 3.5，输入帖子的文本，并使用提示(Prompt)“请判断该信息是谣言、非谣言，还是未验证的信息”。对于 ChatGPT 4o，输入帖子的文本及其相关图片，并给出两种提示“请根据所提供的文本和图片判断信息是否属于谣言、非谣言或未验证的信息，并且不得搜索互联网资料”以及“请根据所提供的文本和图片判断信息是否属于谣言、非谣言或未验证的信息，可以搜索互联网资料作为补充证据”。我们将 Weibo 和 Twitter 两个数据集上五次实验的平均结果作为对比，具体结果如表 3 和表 4 所示。

ChatGPT 3.5 的宏准确率为 36.6%和 36.9%，仅略好于随机猜测。ChatGPT 4o 的宏准确率也仅为 50.3%和 51.7%，在增加了证据搜索的情况下，性能有所提升，但仍然无法与其他监督式学习方法相匹配。这表明，需要设计更复杂的提示或上下文学习策略，才能激发大模型在谣言识别这一特定领域问题上的能力。在单模态谣言检测方法中，TDRD 方法在两个数据集上的宏准确率比 LSTM_Word2vec 方法有所提高，这证明了融合文本主题信息对谣言判断具有积极作用。DC-CNN 方法通过双通道池化层替代传统卷积神经网络的池化层，解决了池化层容易丢失局部和全局特征相关性的问题，从而增强了上下文语义和全局依赖关系的学习，因此其性能相较于 LSTM_Word2vec 和 TDRD 方法有所提升。在所有单模态谣言检测方法中，DEDA 获得了最高的宏准确率，凸显了情感特征在提高谣言检测性能方面的重要性。

在多模态谣言检测方法中，所有多模态方法的性能均优于单模态方法，因为某些谣言仅依赖文本

表 3　Weibo 数据集上实验对比结　　(单位:%)

方法	类型	M_Acc	非谣言			谣言			未验证谣言		
			Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
ChatGPT 3.5	零样本	36.6	35.4	48.6	41.0	22.7	34.5	27.4	25.0	26.7	25.8
ChatGPT 4o		50.3	55.2	57.3	56.2	45.4	51.2	48.1	38.8	42.5	40.6
ChatGPT 4o+搜索		53.0	60.8	59.7	60.2	48.5	54.1	51.1	40.4	45.3	42.7
LSTM_Word2vec	单模态	86.5	90.7	91.1	90.9	82.1	82.1	82.1	87.7	86.3	87.0
TDRD		87.4	91.5	91.9	91.7	82.3	82.9	82.6	88.4	87.5	87.9
DC-CNN		87.6	92.0	92.1	92.1	82.9	83.7	83.3	89.5	87.0	88.3
DEDA		88.4	92.6	92.8	92.7	83.6	84.5	84.1	90.2	87.7	88.9
MCAN		89.3	93.2	93.3	93.3	84.4	85.0	84.7	90.5	89.8	90.1
HMCAN		89.3	93.2	93.4	93.3	84.5	85.1	84.8	90.6	89.5	90.1
MTTV		89.4	93.6	93.7	93.7	84.7	84.6	84.7	91.0	89.7	90.4
ITS		90.6	92.1	96.3	94.2	83.9	86.0	84.9	93.9	89.6	91.7
HGA-MMRD	多模态	90.7	94.2	95.2	94.7	84.3	85.8	85.0	93.3	91.0	92.2
MMCAN		90.7	93.0	96.4	94.6	86.2	84.2	85.2	93.6	91.1	92.3
CCN		90.9	93.1	96.2	94.6	85.3	85.9	85.6	93.0	90.7	91.8
HESN		91.1	93.5	96.2	94.8	85.9	86.2	86.0	92.7	90.8	91.7
SGKE(本文)		**92.7**	**94.8**	**96.7**	**95.7**	**87.2**	**90.1**	**88.6**	**94.8**	**91.4**	**93.0**
△SOTA		↑1.6	↑0.6	↑0.3	↑0.9	↑1.0	↑3.9	↑2.6	↑0.9	↑0.3	↑0.7

注:加粗表示最高值,下划线表示第二高值,△SOTA 表示与最先进方法的对比,↑表示提升。

表 4　Twitter 数据集上实验对比结果　　(单位:%)

方法	类型	M_Acc	非谣言			谣言			未验证谣言		
			Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
ChatGPT 3.5	零样本	36.9	27.1	50.0	35.1	45.0	22.5	30.0	40.6	38.2	39.4
ChatGPT 4o		51.7	57.3	58.2	57.7	55.1	50.6	52.8	44.5	46.4	45.4
ChatGPT 4o+搜索		54.7	62.4	60.7	61.5	57.6	53.5	55.5	46.1	49.8	47.9
LSTM_Word2vec	单模态	87.8	89.5	92.5	91.0	83.8	78.6	81.1	91.4	92.2	91.8
TDRD		88.5	90.9	93.3	92.1	84.7	79.1	81.8	92.0	93.2	92.6
DC-CNN		88.8	91.6	93.8	92.7	85.6	79.4	82.4	92.4	93.4	92.9
DEDA		89.3	92.1	94.5	93.3	86.4	79.4	82.7	93.2	94.0	93.6
MCAN		89.6	93.3	94.2	93.7	86.4	81.0	83.6	93.5	93.8	93.7
HMCAN		89.7	93.2	94.3	93.7	86.8	80.9	83.7	93.7	93.9	93.8
MTTV		90.0	93.8	94.8	94.3	87.3	80.8	83.9	94.0	94.3	94.2
ITS		90.7	94.2	95.5	94.8	87.6	80.9	84.1	94.2	95.6	94.9
HGA-MMRD	多模态	91.4	94.7	97.3	95.9	88.9	80.7	84.6	94.6	96.3	95.5
MMCAN		91.9	94.1	96.5	95.3	91.0	82.8	86.7	95.0	96.2	95.6
CCN		92.1	94.9	96.1	95.5	91.1	84.0	87.4	94.7	96.2	95.4
HESN		92.5	95.1	97.2	96.1	90.7	84.3	87.4	94.8	96.0	95.4
SGKE(本文)		**94.7**	**97.0**	**98.2**	**97.6**	**91.5**	**89.5**	**90.4**	**96.2**	**96.4**	**96.3**
△SOTA		↑2.2	↑1.9	↑0.9	↑1.5	↑0.4	↑5.2	↑3.0	↑1.2	↑0.1	↑0.7

注:加粗表示最高值,下划线表示第二高值,△SOTA 表示与最先进方法的对比,↑表示提升。

特征无法准确识别。MCAN 和 HMCAN 这两个方法均利用预训练的语言和视觉模型从文本和图片中提取特征,然后通过分层和注意力机制实现不同模态特征的交互,因此它们的总体性能相差不大。MTTV 方法在全局视觉特征的基础上,利用 Faster R-CNN 提取了图片的局部特征,因而其性能略优于 MCAN 和 HMCAN。相比之下,ITS 方法的性能有所提升,证明了挖掘图文语义相似度对多模态谣言检测的积极作用。HGA-MMRD 通过层次图注意力网络捕捉模态内和模态间的不同语义交互,增强了模型对图文细粒度语义的理解能力,从而提高了谣言检测的性能。MMCAN 方法通过图像-文本匹配感知捕捉图文的对齐情况,并通过两个协同注意力网络进行相互知识蒸馏,从而实现更好的多模态融合,性能也有所提升。CCN 方法通过融合文本和视觉两种模态的证据,学习待检验帖子和证据之间的一致性,性能也得到了一定程度的提升。HESN 在两个数据集上分别获得了 91.1%和 92.5%的宏准确率,证明了文本实体和视觉实体能够有效增强图像语义和知识语义。本文提出的 SGKE 方法在所有指

标上均优于最先进的基线方法，在宏准确率上分别提高了 1.6%和 2.2%，在谣言类别的 F1 值上提高了 2.6%和 3.0%，证明了本文方法的有效性。

5.3.3 消融实验

为了验证不同模块对 SGKE 方法的贡献，我们设计了以下 12 个消融实验：(1)w/o 文本证据：去除文本证据；(2)w/o 图片证据：去除图片证据；(3)w/o 图文证据：同时去除文本和图片证据；(4)w/o 辟谣证据：过滤掉来源于常见辟谣平台和辟谣公众号的证据；(5)w/o ELA：去除图片篡改特征；(6)w/o 知识蒸馏：不对场景图做知识增强；(7)w/o 图文语义匹配：去除图文语义匹配模块；(8)w/o 去偏的场景图：利用传统的 EGTR 模型生成场景图；(9)o/h 文本和图像：仅利用文本和图像特征；(10)o/h 文本：仅利用文本特征；(11)o/h 图像：仅利用图像特征；(12)SGAT→GAT：将 SGAT 替换为传统的 GAT。在 Weibo 和 Twitter 两个数据集上的对比结果分别如表 5 和表 6 所示。

表 5 Weibo 数据集上的消融实验对比结果 (单位：%)

方法	M_Acc	非谣言			谣言			未验证谣言		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
SGKE(本文)	**92.7**	**94.8**	**96.7**	**95.7**	**87.2**	**90.1**	**88.6**	**94.8**	**91.4**	**93.0**
w/o 文本证据	90.8	95.3	94.4	94.8	84.8	87.0	85.8	91.7	91.0	91.4
w/o 图片证据	91.6	95.4	94.4	94.8	82.7	90.4	86.3	93.5	90.0	91.7
w/o 图文证据	90.3	94.9	94.0	94.4	88.2	84.3	86.2	90.0	92.6	91.3
w/o 辟谣证据	92.5	94.6	96.5	95.5	87.0	89.9	88.4	94.7	91.2	92.9
w/o ELA	92.1	95.3	94.3	94.8	87.4	89.3	88.3	92.8	92.7	92.7
w/o 知识蒸馏	91.9	94.6	96.1	95.4	87.4	86.7	87.0	93.5	92.8	93.1
w/o 图文语义匹配	88.3	86.9	89.4	88.1	87.2	86.3	86.8	90.2	89.1	89.7
w/o 去偏的场景图	92.4	95.9	95.2	95.5	84.9	90.4	87.5	94.3	91.7	92.9
o/h 文本和图像	89.1	93.5	92.9	93.1	83.8	84.2	83.9	90.0	90.2	90.1
o/h 文本	87.4	94.4	93.1	93.7	76.9	80.9	78.8	89.8	88.3	89.0
o/h 图像	83.2	86.1	86.3	86.1	76.2	80.2	78.1	85.9	83.0	84.4
SGAT→GAT	92.0	94.8	95.0	94.9	83.0	90.9	86.8	94.7	90.2	92.4

表 6 Twitter 数据集上的消融实验对比结果 (单位：%)

方法	M_Acc	非谣言			谣言			未验证谣言		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
SGKE(本文)	**94.7**	**97.0**	**98.2**	**97.6**	**91.5**	**89.5**	**90.4**	**96.2**	**96.4**	**96.3**
w/o 文本证据	93.8	95.7	97.7	96.7	89.4	87.8	88.6	96.5	95.8	96.1
w/o 图片证据	93.2	95.4	97.7	96.5	89.7	85.8	87.7	96.0	96.2	96.1
w/o 图文证据	92.2	94.3	97.2	95.7	91.3	83.2	87.0	94.5	96.1	95.3
w/o 辟谣证据	94.6	96.9	98.1	97.5	91.2	89.3	90.2	96.1	96.3	96.2
w/o ELA	91.7	94.3	96.8	95.5	87.3	82.6	84.8	95.0	95.6	95.3
w/o 知识蒸馏	94.2	96.3	96.9	96.6	88.8	90.1	89.4	96.8	95.6	96.1
w/o 图文语义匹配	91.9	93.7	98.0	95.8	89.4	82.1	85.6	95.0	95.5	95.2
o/h 文本和图像	91.4	95.9	95.7	95.8	87.5	81.7	84.4	94.1	96.8	95.5
w/o 去偏的场景图	94.4	96.3	97.1	96.7	90.3	90.0	90.1	96.6	96.0	96.3
o/h 文本	89.4	89.9	96.4	93.0	90.4	75.4	82.2	93.4	96.4	94.9
o/h 图像	77.7	84.1	85.9	84.9	72.9	60.3	65.9	82.2	86.9	84.4
SGAT→GAT	93.3	96.1	94.8	95.4	85.6	90.2	87.8	96.9	94.9	95.9

从实验结果中可以看出，SGKE 方法的每个模块都发挥着独特的作用，去除任何一个模块都会影响方法的性能。具体而言，去除文本证据、图文证据或图文双证据时，在 Weibo 数据集上的宏准确率分别下降了 1.9%、1.1%和 2.4%；而在 Twitter 数据集上则分别下降了 0.9%、1.4%和 2.5%。互联网上检索的文本和图片证据蕴含了丰富的世界知识，能够从不同角度帮助模型验证信息的真实性，因此去除这些证据会降低方法的性能。此外，我们发现去除证据模块后，对谣言类别的识别影响更为显著。这可能是因为，相较于非谣言和未验证谣言，判断谣言类别更依赖于证据的支持。在 Weibo 数据集上，去除图文证据的方法在谣言类别的 F1 值略高于去除文本证据的方法，可能是由于随机性引起的样本分布和初始化参数波动所致。由于辟谣证据占比很少，所以去除后对方法的性能影响较小。去除图片

篡改特征后,方法在 Weibo 数据集上的宏准确率和谣言类别的 F1 值分别下降了 0.6%和 0.3%,而在 Twitter 数据集上则分别下降了 3.0%和 5.6%。这可能是由于 Twitter 数据集中包含了更多与图片篡改相关的谣言。去除知识蒸馏后,方法的性能也有下降,因为从知识图谱中获取的知识能够有效增强模型对场景图的语义理解。去除图文匹配模块后,方法在两个数据集上的宏准确率分别下降了 4.4%和 2.8%,这表明某些谣言中存在明显的图文语义不匹配,因此图文语义匹配特征对于多模态谣言检测至关重要。基于反事实推理的去偏方法能够生成语义更丰富、表达更精准的场景图。因此,在去除该模块后,Weibo 数据集上的宏准确率、非谣言类别和谣言类别 F1 分别下降了 0.3%、0.2%和 1.1%,而在 Twitter 数据集上分别下降了 0.3%、0.9%和 0.3%。

如果仅保留帖子的文本和图像特征,方法的性能将显著下降,因为仅依靠不同模态信息的互补尚不足以有效应对多样化的谣言信息。如果退化为基于文本的单模态方法,Weibo 和 Twitter 数据集上的宏准确率分别仅为 87.4%和 89.4%;而退化为基于图像的单模态方法,准确率更是下降至 83.2%和 77.7%。这表明某些多模态谣言仅凭文本或图像信息无法准确判断,需要不同模态信息的交互与互补。基于图像的单模态方法性能低于基于文本的单模态方法,原因在于文本提供的谣言特征(如语义、情感和风格)比图像更加明显。将本文设计的 SGAT 算法替换为传统的 GAT 后,方法在 Weibo 数据集上的宏准确率和谣言类别的 F1 值分别下降了 0.7%和 1.8%;在 Twitter 数据集上分别下降了 1.4%和 2.6%。这证明了图神经网络计算注意力权重和特征聚合时,引入场景图关系特征的重要性。通过这些消融实验,进一步验证了本文提出的 SGKE 方法的有效性。

5.3.4　误差分析

为了对预测结果的误差进行详细分析,我们对五次随机实验的测试结果混淆矩阵进行了可视化,如图 15 所示。其中,纵坐标表示真实类别,横坐标表示预测类别。本文方法在 Weibo 数据集上对非谣言、谣言和未验证谣言的识别率分别达到了 96.7%、89.6%和 91.4%;在 Twitter 数据集上则分别为 98.2%、89.5%和 94.4%。两个数据集上的识别率均呈现出非谣言>未验证谣言>谣言的顺序,这一结果是可以理解的,因为对于人类而言,判断这三种类别的难度也是依次递增的。通过误差分析发现,SGKE 方法容易将谣言误判为未验证的谣言,反之亦然。这也是容易理解的,因为这两类谣言之间的区分度不高,从狭义的定义来看,未验证的谣言实际上也属于谣言。因此,这种更细粒度的划分影响了本文方法在谣言类别上的识别性能。

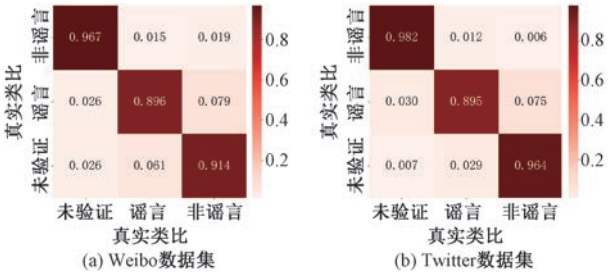


图 15　测试结果的混淆矩阵

5.3.5　二分类实验

为了进一步验证本文方法在二分类谣言检测(即谣言与非谣言)上的效果,我们使用了 Pheme 和 Weibo[69](记为 Weibo-2)两个传统数据集。Pheme 数据集来源于 Twitter,包含关于五个热点新闻的谣言和非谣言;Weibo-2 数据集则包括来自新华社和微博平台的非谣言和谣言。在数据预处理过程中,我们确保每条文本对应一张图像,并剔除了重复度较高的帖子。最终,Pheme 数据集包含 2001 个样本(其中 1416 条为非谣言,585 条为谣言),Weibo-2 数据集包含 7959 个样本(其中 3640 条为非谣言,4319 条为谣言)。在实验中,我们选取了对比实验中的多模态谣言检测方法作为基线方法。实验对比结果如表 7 所示。

表 7　在二分类数据集上的实验对比

数据集	方法	准确率	精确率	召回率	F1 值
Pheme	MCAN	88.7	85.0	84.6	84.8
	HMCAN	88.8	85.2	84.7	84.9
	MTTV	89.0	85.2	85.1	85.0
	ITS	89.5	86.4	85.2	85.8
	HGA-MMRD	89.9	87.0	85.2	86.0
	MMCAN	90.1	87.6	86.4	87.0
	CCN	90.3	87.5	88.6	88.0
	HESN	90.8	88.6	89.6	89.0
	SGKE(本文)	**91.6**	**90.3**	**90.0**	**90.2**
Weibo-2	MCAN	90.4	90.7	90.3	90.4
	HMCAN	90.5	90.1	90.6	90.2
	MTTV	91.2	90.9	91.1	91.0
	ITS	91.7	91.5	91.6	91.5
	HGA-MMRD	92.6	92.4	92.7	92.5
	MMCAN	92.8	92.6	92.8	92.7
	CCN	93.0	92.9	92.9	92.9
	HESN	93.3	92.9	93.3	93.0
	SGKE(本文)	**94.3**	**94.2**	**94.3**	**94.2**

实验结果表明,在二分类数据集上,本文方法的性能仍然优于所有对比基线方法。在 Pheme 数据集中,本文方法相较于最优基线方法,在准确率、精确率、召回率和 F1 值上分别提高了 0.8%、1.7%、0.4%和 1.2%;而在 Weibo-2 数据集中,分别提高了 1.0%、1.3%、1.0%和 1.2%。这些结果证明了本文方法在二分类谣言检测任务中的有效性。

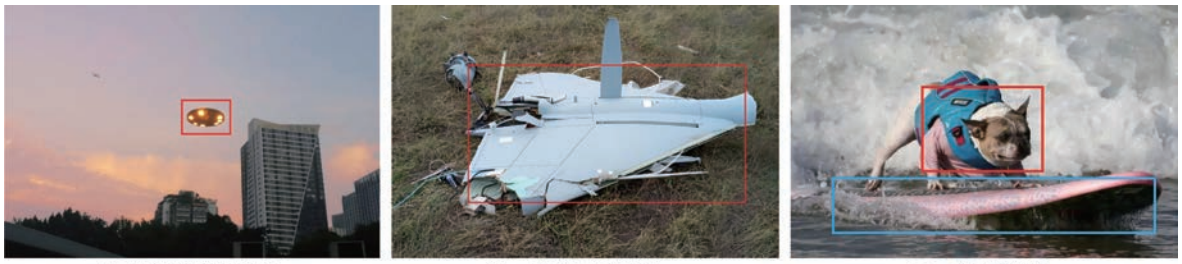
5.4 案例分析

图 16 展示了三个案例,清晰地揭示了本文提出的 SGKE 方法的优势与不足。在图 16(a)中,这则“比尔·盖茨近日被美国联邦法警逮捕”的谣言模仿了华盛顿邮报真实新闻的报道风格,因此基于文本的单模态方法误判其为“非谣言”。然而,通过分析图片内容,可以发现其过于夸张且有篡改的痕迹,基于多模态的基线方法和本文提出的方法都准确地识别出该信息为“谣言”。在图 16(b)中,这则“竞选团队工作人员因选民欺诈被捕”谣言的表达风格类似于真实新闻,且帖子中的文本与图片表达的语义相似,因此基于文本的单模态方法和基于图文语义相似度的多模态方法均将其误判为“非谣言”。然而,基于图片证据中的“FALSE”注释,以及文本证据中的“虚假”和“毫无根据”等线索,本文提出的 SGKE 方法正确预测了该帖子为“谣言”。在图 16(c)中,这则未验证的谣言的文本似于真实新闻的表达风格,图文语义也一致,且检索到的文本证据和图片证据与帖子内容相互验证,因此所有方法均错误地将其预测为“非谣言”。然而,所有方法都忽略了帖子中提到的实际伤亡情况尚未确认这一事实,因此该帖子应标记为“未验证的谣言”。这也是本文方法容易混淆“谣言”和“未验证的谣言”的主要原因。

(a) 单模态方法误判的谣言 (b) 仅本文方法判断正确的谣言 (c) 所有方法都误判的未验证谣言

图 16 案例分析

图 17 展示了三个图文细粒度语义匹配的示例,其中,图像中框选的物体对象与文本中相应颜色标注的实体在协同注意力矩阵中具有较高的权重值。图像中的关键物体对象与文本中的相应实体实现了良好的匹配,验证了本文提出的基于场景图注意力网络的图文语义匹配算法能够有效地实现图文之间的细粒度语义的对齐。



(a) 天空出现不明飞碟 (b) 一架无人机被摧毁 (c) 一条狗在冲浪

图 17 图文细粒度语义匹配示例

6 总　结

针对现有多模态方法在识别图文深层语义不匹配谣言方面的困难,本文提出了一种融合外部知识和证据的多模态谣言检测方法。通过知识蒸馏获取背景知识,增强模型对图文语义的理解;同时,从互联网中筛选文本和图片证据,以提高检测的准确性和可解释性。此外,针对图文异构特征空间中的语义鸿沟问题,设计了一种基于场景图注意力网络的图文细粒度语义对齐方法。通过在 Weibo 和 Twitter 这两个真实社交网络谣言数据集上进行大量实验与分析,结果表明本文提出的方法优于目前最先进的基线方法,证明了其有效性。

尽管本文的方法取得了良好的效果,但也面临一些挑战。首先,使用搜索引擎虽然能够提供相关证据,但也可能引入噪声信息,从而误导预测结果。其次,并非所有证据都是可信的,而且证据之间可能存在相互矛盾,这对方法的证据对齐提出了挑战。此外,除了文本和图片之外,还可以在其他模态中找到证据,例如音频和视频。最后,随着深度伪造技术的发展,篡改不仅出现在图片中,也逐渐出现在视频中。因此,如何挖掘更多不同模态的证据和知识并有效融合,将是未来多模态谣言检测的重要研究方向。

参　考　文　献

[1] Liu Ya-Hui, Jin Xiao-Long, Shen Hua-Wei, et al. A survey rumor identification over social media. Chinese Journal of Computers, 2018,41(7): 1536-1558 (in Chinese)
(刘雅辉,靳小龙,沈华伟等. 社交媒体中的谣言识别研究综述. 计算机学报,2018,41(7):1536-1558)

[2] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171-4186

[3] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2020:1877-1901

[4] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[5] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Virtual, Austria, 2021: 1-21

[6] Zhang X C, Dadkhah S, Weismann A G, et al. Multimodal fake news analysis based on image-text similarity. IEEE Transactions on Computational Social Systems, 2024,11(1): 959-972

[7] Nadeem M I, Ahmed K, Zheng Z Y, et al. SSM: Stylometric and semantic similarity oriented multimodal fake news detection. Journal of King Saud University-Computer and Information Sciences, 2023,35(5):101559

[8] Sun M Z, Zhang X, Ma J Q, et al. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. IEEE Transactions on Knowledge and Data Engineering, 2023,35(12):12736-12749

[9] Jiang Y, Yu X M, Wang Y M, et al. Similarity-aware multi-modal prompt learning for fake news detection. Information Sciences, 2023,647:119446

[10] Gao Yu-Jun, Liang Gang, Jiang Fang-Ting, et al. Social network rumor detection: A survey. Acta Electonica Sinica, 2020,48(7):1421-1435 (in Chinese)
(高玉君,梁刚,蒋方婷等. 社会网络谣言检测综述. 电子学报,2020,48(7):1421-1435)

[11] 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)

[12] Zhang F, Liu J, Xie J, et al. ESCNet: Entity-enhanced and stance checking network for multi-modal fact-checking//Proceedings of the ACM on Web Conference 2024. Singapore, 2024: 2429-2440

[13] Xu F, Sheng V S, Wang M W. Near real-time topic-driven rumor detection in source microblogs. Knowledge-Based Systems, 2020, 207:106391

[14] Ma K, Tang C H, Zhang W J, et al. DC-CNN: Dual-channel convolutional neural networks with attention-pooling for fake news detection. Applied Intelligence, 2023,53(7):8354-8369

[15] Zhang H, Li Z L, Liu S Y, et al. Do sentence-level sentiment interactions matter? sentiment mixed heterogeneous network for fake news detection. IEEE Transactions on Computational Social Systems, 2024,11(4):5090-5100

[16] Chakraborty A, Khatri I, Choudhry A, et al. An emotion-guided approach to domain adaptive fake news detection using adversarial learning//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 16178-16179

[17] Huang X J, Ma T H, Jia L, et al. An effective multimodal representation and fusion method for multimodal intent rec-

ognition. Neurocomputing, 2023,548:126373

[18] Alkhodair S A, Ding S H H, Fung B C M, et al. Detecting breaking news rumors of emerging topics in social media. Information Processing & Management, 2020,57(2):102018

[19] Luvembe A M, Li W M, Li S H, et al. Dual emotion based fake news detection: A deep attention-weight update approach. Information Processing & Management, 2023,60(4):103354

[20] Su Xing, Yu Ke, Wu Xiao-Fei. Gated interactive fusion network for rumor detection. Journal of Beijing University of Posts and Telecommunications, 2023,46(4):97-102 (in Chinese)
(苏兴,禹可,吴晓非. 基于层次门控交互融合网络的谣言检测方法. 北京邮电大学学报,2023,46(4):97-102)

[21] Wang You-Wei, Feng Li-Zhou, Wang Wei-Qi, et al. Weibo rumor detection based on heterogeneous graph of event-word-feature. Journal of Chinese Information Processing, 2023,37(9):161-174 (in Chinese)
(王友卫,凤丽洲,王炜琦等. 基于事件-词语-特征异质图的微博谣言检测新方法. 中文信息学报,2023,37(9):161-174)

[22] Xu Fan, Li Ming-Hao, Huang Qi, et al. Knowledge graph-driven graph neural network-based model for rumor detection. Scientia Sinica (Informationis), 2023,53(4):663-681 (in Chinese)
(徐凡,李明昊,黄琪等. 知识图谱驱动的图卷积神经网络谣言检测模型. 中国科学:信息科学,2023,53(4):663-681)

[23] Alsaif H F, Aldossari H D. Review of stance detection for rumor verification in social media. Engineering Applications of Artificial Intelligence, 2023, 119: 105801

[24] Zhang Q, Yang Y Y, Shi C Y, et al. Rumor detection with hierarchical representation on bipartite ad hoc event trees. IEEE Transactions on Neural Networks and Learning Systems, 2024,35(10):14112-14124

[25] Huang X J, Ma T H, Jin W W, et al. Multiview spatio-temporal learning with dual dynamic graph convolutional networks for rumor detection. IEEE Transactions on Computational Social Systems, 2025(early access)

[26] Truică C O, Apostol E S, Karras P. DANES: Deep neural network ensemble architecture for social and textual context-aware fake news detection. Knowledge-Based Systems, 2024,294:111715

[27] Wang J G, Qian S S, Hu J, et al. Comment-context dual collaborative masked transformer network for fake news detection. IEEE Transactions on Multimedia, 2024,26:5170-5180

[28] Yang Yan-Jie, Wang Li, Wang Yu-Hang. Rumor detection based on source information and gating graph neural network. Journal of Computer Research and Development, 2021,58(7):1412-1424 (in Chinese)
(杨延杰,王莉,王宇航. 融合源信息和门控图神经网络的谣言检测研究. 计算机研究与发展,2021,58(7):1412-1424)

[29] Yang Fan, Li Shao-Mei. Incorporating user features for Weibo rumor detection via graph attention network. Journal of Chinese Information Processing, 2024,38(8):140-146 (in Chinese)
(杨帆,李邵梅. 融合用户特征的图注意力微博谣言检测模型. 中文信息学报,2024,38(8):140-146)

[30] Pi De-Chang, Wu Zhi-Yuan, Cao Jian-Jun. Early rumor detection method based on knowledge graph representation learning. Acta Electronica Sinica, 2023,51(2):385-395 (in Chinese)
(皮德常,吴致远,曹建军. 基于知识图谱表示学习的谣言早期检测方法. 电子学报,2023,51(2):385-395)

[31] Yang C, Zhang P, Qiao W B, et al. Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 5705-5717

[32] Ke Jing, Xie Zhe-Yong, Xu Tong, et al. An implicit semantic enhanced fine-grained fake news detection method based on large language models. Journal of Computer Research and Development, 2024,61(5): 1250-1260 (in Chinese)
(柯婧,谢哲勇,徐童等. 基于大语言模型隐含语义增强的细粒度虚假新闻检测方法. 计算机研究与发展,2024,61(5):1250-1260)

[33] Zhang X, Gao W. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method//Proceedings of the 13th International Joint Conference on Natural Language. Nusa Dua, Indonesia, 2023: 996-1011

[34] Caramancion K M. News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in news fact-checking. arXiv Preprint arXiv: 2306.17176, 2023

[35] Li J, Bin Y, Peng L, et al. Focusing on relevant responses for multi-modal rumor detection. IEEE Transactions on Knowledge and Data Engineering, 2024,36(11): 6225-6236

[36] Zhang Q, Liu J, Zhang F, et al. Hierarchical semantic enhancement network for multimodal fake news detection//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 3424-3433

[37] Wang B, Feng Y, Xiong X C, et al. Multi-modal transformer using two-level visual features for fake news detection. Applied Intelligence, 2023,53(9):10429-10443

[38] Wu L W, Long Y Z, Gao C, et al. MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection. Information Fusion, 2023,100:101944

[39] Huang Xue-Jian, Ma Ting-Huai, Wang Gen-Sheng. Multimodal learning method based on intra-and inter-sample cooperative representation and adaptive fusion. Journal of Computer Research and Development, 2024,61(5):1310-1324 (in Chinese)
(黄学坚,马廷淮,王根生. 基于样本内外协同表示和自适应融合的多模态学习方法. 计算机研究与发展,2024,61(5):1310-1324)

[40] Zheng J Q, Zhang X, Guo S C, et al. MFAN: Multi-modal

feature-enhanced attention networks for rumor detection//Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria, 2022: 2413-2419

[41] Wu Y, Zhan P W, Zhang Y J, et al. Multimodal fusion with co-attention networks for fake news detection//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2021: 2560-2569

[42] Hu L M, Zhao Z W, Qi W J, et al. Multimodal matching-aware co-attention networks with mutual knowledge distillation for fake news detection. Information Sciences, 2024, 664:120310

[43] Xu F, Zeng L, Huang Q, et al. Hierarchical graph attention networks for multi-modal rumor detection on social media. Neurocomputing, 2024,569:127112

[44] 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)

[45] Xiong S F, Zhang G P, Batra V, et al. TRIMOON: Two-round inconsistency-based multi-modal fusion network for fake news detection. Information Fusion, 2023,93:150-158

[46] Wu L W, Liu P S, Zhao Y Q, et al. Human cognition-based consistency inference networks for multi-modal fake news detection. IEEE Transactions on Knowledge and Data Engineering, 2024,36(1): 211-225

[47] Wang J, Yang Y, Liu K Y, et al. Instance-guided multi-modal fake news detection with dynamic intra-and inter-modality fusion//Proceedings of the 26th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Chengdu, China, 2022: 510-521

[48] Liu Jin-Shuo, Feng Kuo, Jeff Z. Pan, et al. MSRD: Multi-modal web rumor detection method. Journal of Computer Research and Development, 2020,57(11):2328-2336 (in Chinese)
(刘金硕,冯阔,Jeff Z. Pan,等. MSRD:多模态网络谣言检测方法. 计算机研究与发展,2020,57(11):2328-2336)

[49] Qian S S, Wang J G, 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. Online, 2021: 153-162

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

[51] Meel P, Vishwakarma D K. HAN, image captioning, and forensics ensemble multimodal fake news detection. Information Sciences, 2021,567:23-41

[52] Zhang H W, Qian S S, Fang Q, et al. Multi-modal meta multi-task learning for social media rumor detection. IEEE Transactions on Multimedia, 2022,24:1449-1459

[53] Zhou H H, Ma T H, Rong H, et al. MDMN: Multi-task and domain adaptation based multi-modal network for early rumor detection. Expert Systems with Applications, 2022, 195:116517

[54] Zhang F, Liu J, Zhang Q, et al. ECENet: Explainable and context-enhanced network for muti-modal fact verification//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 1231-1240

[55] Huang X J, Ma T H, Rong H, et al. Dual evidence enhancement and text-image similarity awareness for multimodal rumor detection. Engineering Applications of Artificial Intelligence, 2025, 153:110845

[56] 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. New Orleans, USA, 2022: 14940-14949

[57] Qi P, Yan Z, Hsu W, et al. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 13052-13062

[58] Dong Y, He D, Wang X, et al. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 8354-8362

[59] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 10012-10022

[60] Im J, Nam J Y, Park N, et al. EGTR: Extracting graph from transformer for scene graph generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 24229-24238

[61] Tang K H, Niu Y L, Huang J Q, et al. Unbiased scene graph generation from biased training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3716-3725

[62] Sun S Z, Zhi S F, Liao Q, et al. Unbiased scene graph generation via two-stage causal modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12562-12580

[63] Li Z, Chai Y Y, Zhuo T Y, et al. Factual: A benchmark for faithful and consistent textual scene graph parsing//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 6377-6390

[64] Kolitsas N, Ganea O E, Hofmann T. End-to-end neural entity linking//Proceedings of the 22nd Conference on Computational Natural Language Learning. Brussels, Belgium, 2018: 519-529

[65] Le P, Titov I. Improving entity linking by modeling latent relations between mentions//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 1595-1604

[66] Yang J, Li Y, Gao C, et al. Entity disambiguation with context awareness in user-generated short texts. Expert Systems with Applications, 2020, 160: 113652

[67] Zhang Z, Zhuang F, Qu M, et al. Knowledge graph embedding with hierarchical relation structure//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3198-3207

[68] Hu X M, Guo Z J, Chen J Z, et al. MR2: A benchmark for multimodal 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

[69] 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. Mountain View, USA, 2017: 795-816

HUANG Xue-Jian, Ph. D. candidate, lecturer. His main research areas include rumor detection, multi-modal learning, and natural language processing.

MA Ting-Huai, Ph. D., professor. His main research areas include privacy protection in social networks, big data mining, and text sentiment analysis.

RONG Huan, Ph. D., associate professor. His main research areas include social media mining, social network content security, and knowledge engineering.

WANG Gen-Sheng, Ph. D., professor. His main research areas include data mining and social networks.

LIAO Guo-Qiong, Ph. D., professor. His main research areas include databases, data mining, and social networks.

LIU De-Xi, Ph. D., professor. His main research areas include social media processing, natural language processing, and computational psychology.

Background

The widespread proliferation of social networks has greatly facilitated the dissemination and sharing of information, but it has also created fertile ground for the spread of rumors. Compared to purely textual rumors, multimodal rumors that combine text and images are more deceptive and transmissible due to their strong visual impact and the "seeing is believing" cognitive bias. This makes multimodal rumor detection an urgent issue in the field of network security.

Current mainstream approaches mainly focus on learning features from each modality and employing various attention mechanisms to fuse them for detection. However, they still face two major challenges: (1) difficulty in effectively capturing fine-grained semantic correlations between images and text; and (2) limited ability to detect well-crafted rumors that exhibit deep semantic inconsistency across modalities.

To address these challenges, this paper proposes a multimodal rumor detection method based on a Scene Graph Attention Network (SGAT) that integrates external knowledge and evidence. Specifically, the proposed approach constructs a scene graph attention mechanism to explicitly model the semantic relationships between visual objects and textual terms, enabling fine-grained alignment between visual and linguistic features. In addition, knowledge distillation is introduced to incorporate background knowledge from external knowledge bases, thereby enhancing the model's ability to understand implicit semantics. Furthermore, by incorporating textual and visual evidence, the model's performance in identifying deeply inconsistent rumors is significantly improved, both in terms of accuracy and interpretability.

Experimental results on two real-world social media datasets, Weibo and Twitter, demonstrate that the proposed method outperforms existing state-of-the-art baseline models across multiple evaluation metrics, validating its effectiveness and advancement.

This work was supported in part by the National Natural Science Foundation of China (Nos. 62372243, 62272206), the Natural Science Foundation of Jiangxi Province (No. 20242BAB20074), and the Humanities and Social Sciences Research Project of Jiangxi Provincial Universities (No. JC24219).