计算机科学
COMPUTER SCIENCE

基于多模态双协同Gather Transformer网络的虚假信息检测方法

向旺, 王金光, 王一飞, 钱胜胜

引用本文

向旺, 王金光, 王一飞, 钱胜胜. 基于多模态双协同Gather Transformer网络的虚假信息检测方法[J]. 计算机科学, 2024, 51(12): 242-249.

XIANG Wang, WANG Jinguang, WANG Yifei, QIAN Shengsheng. Multi-modal Dual Collaborative Gather Transformer Network for Fake News Detection [J]. Computer Science, 2024, 51(12): 242-249.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000057

基于多模态双协同 Gather Transformer 网络的虚假信息检测方法

向　旺[1]　王金光[2]　王一飞[1]　钱胜胜[3]
1 郑州大学河南先进技术研究院　郑州 450000
2 合肥工业大学计算机与信息学院　合肥 230601
3 中国科学院自动化研究所多模态人工智能系统全国重点实验室　北京 100190
(1584462772@qq.com)

摘　要　社交媒体网站是人们在日常生活中分享信息、表达和交换意见的便捷平台。随着用户数量的不断增加,社交媒体网站上出现了大量的信息数据。然而,由于用户没有检查共享信息的可靠性,这些信息的真实性难以保证,从而导致大量虚假信息在社交媒体上广泛传播。然而,现有方法大多存在以下局限性:1)大多数方法通过简单提取文本与视觉特征,将其拼接后得到多模态特征来进行虚假信息判断,忽略了模态间和模态内细粒度内在联系,缺乏对关键信息的检索和筛选;2)多模态信息间缺乏指导性的特征提取,文本和视觉等特征之间缺乏交互增强,对多模态信息的理解不足。为了应对这些挑战,提出了一种新颖的基于多模态双协同 Gather Transformer 网络(Multimodal Dual-Collaborative Gather Transformer Network,MDCGTN)的虚假信息检测方法。在 MDCGTN 模型中,通过文本-视觉编码网络对文本和视觉信息的特征表示进行提取,将获得的视觉和文本特征表示输入多模态 Gather Transformer 网络进行多模态信息融合,使用 Gather 机制提取关键信息,充分捕捉和融合模态内和模态间细粒度关系。此外,设计了一个双协同机制对社交媒体帖子的多模态信息进行整合,以实现模态之间信息的交互和增强。在两个公开可用的基准数据集上进行了大量实验,结果表明,与现有的先进基准方法相比,所提方法准确率明显提升,证明了其对于虚假信息检测的优越性能。

关键词:多模态;虚假信息检测;社交媒体;Gather Transformer 网络
中图分类号　TP389

Multi-modal Dual Collaborative Gather Transformer Network for Fake News Detection

XIANG Wang[1],WANG Jinguang[2],WANG Yifei[1] and QIAN Shengsheng[3]
1 Henan Institute of Advanced Technology,Zhengzhou University,Zhengzhou 450000,China
2 School of Computer Science and Information Engineering,Hefei University of Technology,Hefei 230601,China
3 State Key Laboratory of Multimodal Artificial Intelligence Systems,Institute of Automation,Chinese Academy of Sciences,Beijing 100190,China

Abstract　Social media platforms are convenient platforms for people to share information,express opinions,and exchange ideas in their daily lives. With the increasing number of users,a large amount of data has emerged on social media websites. However,the authenticity of the shared information is difficult to be guaranteed due to users' lack of verification. This situation has led to the widespread dissemination of a large amount of fake news on social media. However,existing methods suffer from the following limitations:1)Most existing methods rely on simple text and visual feature extraction,concatenating them to obtain multimodal features for detecting fake news,while ignoring the fine-grained intrinsic connections within and between modalities,and lacking retrieval and filtering of key information. 2)There is a lack of guided feature extraction among multimodal information,with insufficient interaction and understanding between textual and visual features. To address these challenges,a novel multimodal dual-collaborative gather transformer network(MDCGTN) is proposed to overcome the limitations of existing methods. In the MDCGTN model,textual and visual features are extracted using a text-visual encoding network,and the obtained features are input into a multimodal gather transformer network for multimodal information fusion. The gathering mechanism is used to extract key information,fully capturing and fusing fine-grained relationships within and between modalities. In addition,a dual-collaborative mechanism is designed to integrate multimodal information in social media posts,enhancing interaction and understanding between modalities. Extensive experiments are conducted on two publicly available benchmark datasets. Compared to existing state-of-the-art benchmark methods,the proposed MDCGTN method achieves significant improvement in accuracy,demonstrating its

到稿日期:2023-10-10　返修日期:2024-03-09
基金项目:国家自然科学基金(62276257)
This work was supported by the National Natural Science Foundation of China(62276257).
通信作者:钱胜胜(shengsheng.qian@nlpr.ia.ac.cn)

superior performance in detecting fake news.

Keywords　Multi-modal，Fake news detection，Social media，Gather transformer network

1　引言

近年来，随着社交媒体平台的广泛普及和便利性的增加，越来越多的人加入到在线新闻的发布和使用行列当中。这一趋势的迅速发展可以归因于社交媒体的普及，以及使用便捷性的提升。由于用户群体的不断壮大，社交媒体网站已经成为各种信息的集散地。然而，用户往往会忽略对其分享内容的可靠性进行审查，导致发布信息的真实性变得不可靠。这种核实上的松懈为大量虚假信息的传播铺平了道路。虚假信息恶意地扭曲事实信息，对个人和整个社会都产生了不利影响。例如，2023 年 3 月，“四川德阳‘中江机场建设成功’”的虚假信息广泛传播于社交媒体，引起了较多的社会讨论和议论，占据了大量公众媒体资源。因此，为了避免造成不必要的损失和负面影响，我们需要开发合适的方法来帮助识别虚假新闻，以使读者能够获得真实的信息。

近年来，虚假信息检测一直是研究关注的焦点。早期的方法主要依赖于用户报告和专家确认，然而这些方法非常耗时且劳动密集。因此，研究人员开始探索检测自动化的方法来解决该问题，这些方法可分为两种类型：1）基于手工制作特征的方法[1-4]，这些方法通过从帖子内容和用户社交背景中提取特征，使用支持向量机（SVM）[1,4]、决策树分类器[2-3]等机器学习算法实现虚假信息检测。然而，手工制作的特征可能无法完全捕捉虚假信息中的复杂内容，存在一定局限性。2）基于深度学习的方法，这类方法利用神经网络来捕捉深层次的特征，例如，Ma 等[5]利用循环神经网络（RNNs）从帖子中提取潜在的特征表示，而 Yu 等[6]则利用卷积神经网络（CNNs）进行特征提取，并辨别它们在虚假信息内容中的高阶相互作用。近年来，社交媒体内容构成已经从单纯的文本扩展到包括图片和视频等多媒体格式[7-10]，这些格式为识别虚假信息提供了补充信息。此外，用户评论也包含了确定帖子内容真实性的关键线索。然而，上述大部分方法仅集中在文本内容中，忽视了具有多模态信息的社交媒体帖子。为解决该问题，基于深度神经网络的多模态方法被提出。虽然这些方法在虚假信息检测任务中取得了不错的效果，但大部分方法仍没有充分考虑多模态信息间和信息内的细粒度内在联系，缺乏对关键信息的检索和筛选[11-13]。

此外，社交媒体帖子对应的用户评论同样可以提供有价值的线索和补充信息，进而提高帖子内容真实性的检测性能[9,14-15]。例如，Lin 等[16]提出了一种基于无向交互图的创新的声明导向分层图注意力网络，通过考虑包括声明和相关响应评论在内的全面社交上下文信息，增强了对响应帖子特征表示的学习。Shang 等[15]设计了 DGExplain（Duo-Generative Explainable Misinformation Detection）模型，它能巧妙利用用户评论定位和阐明相关新闻中的错误信息。然而，这些方法往往缺乏对用户评论的筛选和利用，不同模态信息之间缺乏相互指导和交互，导致对多模态信息理解不足。

为了构建一种更有效的虚假信息检测方法，需要先解决以下问题：

1）如何探索和捕捉多模态信息间和信息内的细粒度关系，并对关键信息进行有效检索和提取？

2）如何提高帖子中不同模态的指导性特征提取以及多模态信息间彼此交互增强的能力，进而提高对多模态信息的理解？

针对上述局限，本文提出了一种新颖的基于多模态双协同 Gather Transformer 网络（MDCGTN）的虚假信息检测方法，该网络模型充分融合了文本和视觉信息。针对问题 1，引入了 Gather Transformer 网络，通过索引和选择检测信息中的关键内容，提高模型对多模态信息关键特征的提取能力。针对问题 2，引入双协同机制，增强文本与视觉、评论信息间的彼此指导关系，提高特征间交互增强能力，增强对多模态信息的理解。在得到社交媒体帖子的统一表示之后，我们利用具有相应激活函数的全连接层来对帖子的真实性进行分类。

总而言之，本文工作有以下几个贡献点：1）通过添加一个 Gather 机制来改进传统 Transformer 网络结构，充分探索和捕捉多模态信息间和信息内的细粒度内在联系，有效筛选多模态信息中的关键部分；2）提出了一个双协同机制，以增加多模态信息间的彼此指导和协同，进而提高多模态信息间的交互增强能力，提高对多模态信息的理解能力；3）在 MMCoVaR[17] 和 ReCOVery[18] 两个公共基准数据集上的广泛实验结果表明，所提模型比目前最先进的虚假信息检测方法表现更好。

2　相关工作

2.1　基于多模态的虚假信息检测

一些学者利用机器学习技术，通过视觉特征提取器和文本特征提取器提取图像和文本的特征，再将它们拼接起来，用于虚假信息检测任务。例如，2019 年提出的 Spotfake 方法利用 VGG19 和 BERT 分别提取视觉信息和文本信息，将它们拼接起来输入分类器中进行虚假信息分类；Spotfake+[19] 方法则采用 VGG 和 XLNET 提取多模态特征；MSRD[20] 方法考虑了帖子的图片中包含的文本信息，使用 LSTM 建模文本信息以及图像中的文本信息，并使用 VGG 建模视觉信息，将多模态信息进行拼接，得到最终的多模态特征表示。由于直接将多模态信息拼接的方法过于简单，无法充分利用多模态信息，因此，一些学者设计了辅助任务，以帮助模型更好地理解多模态信息。例如，EANN[21] 采用 VGG 提取视觉特征和 Text-CNN[22] 提取文本特征，将它们拼接起来得到帖子的特征表示，并利用事件鉴别器将拼接的多模态信息作为输入，输出事件类别用于辅助判断。

Khattar 等[14]提出的 MVAE 同样采用 VGG 和双向 LSTM 分别提取图像特征和文本特征，并设计了信息重构的辅助任务，以提高模型利用多模态信息的效率。还有一些学者认为，社交媒体帖子的图片内容和文本内容是否相符，是判断该帖子是否为虚假信息的重要指标。基于该假设，学者们

提出了另一种多模态检测方法，该方法可以检测帖子的图文相符性，进而判断帖子的真实性。具体方法是将社交媒体帖子的图片信息和文本信息分别编码，并通过计算它们的相似度来判断它们是否匹配。如果相似度很高，则说明该帖子的文本信息和视觉信息匹配，即为真实信息；如果相似度很低，则说明该帖子的文本信息和视觉信息不匹配，即为虚假信息。Zhou 等[23]基于以上方式提出了一种多模态检测方法，该方法利用 image2text[24]模型将图片信息转化为文本信息，并通过全连接层将文本信息和图像信息映射到同一向量空间中。最终，该方法通过比较两者之间的相似度来判断信息的真实性。Xue 等[25]则提出了另一种方法，该方法分别使用 BERT 和 ResNet 来提取文本特征和图像特征，并计算两者之间的相似度，最终也是通过判断相似度进而判断帖子的真实性。

然而，上述方法仅简单将文本与图像特征进行提取和拼接，忽略了模态间和模态内的复杂关系。多模态信息间缺乏彼此指导和交互，导致模型对模态信息理解不足。

2.2 Attention 机制

Attention 机制已经被广泛应用于各种任务中，如图像字幕生成[26]、机器翻译[27]和推荐系统[28-29]等。Attention 机制最初由 Bahdanau 等[27]引入，被应用于机器翻译任务中，能在预测目标词时关注句子中相关部分。随后，Transformer[30]模型被设计出来，它利用 Attention 机制作为 LSTM 的替代品来解决 Sequence-to-Sequence 的问题，并在机器翻译任务中取得了前所未有的效果。Baevski 等[31]提出了自适应单词输入表示，使模型能够为常见单词分配更多空间，同时限制不常见单词的空间。Dai 等[32]提出了 Transformer-XL 模型，它具有相对位置编码和缓存内容，增强了模型扩展上下文信息的建模能力。在此基础上，Rae 等[33]将 Transformer-XL 内存段扩展到更为复杂的压缩内存中，进一步扩展上下文长度，并在 WikiText-103 基础测试中达到了较好的效果。尽管以上方法都表明更长的上下文信息更有利于语言建模任务，但利用密集点数据生成增强上下文信息表示仍未得到充分探索。同时，各种实验表明，经过大型文本语料库训练的大规模神经语言模型有其显著优势。OpenAI 的 GPT[34]和 BERT[35]模型分别采用自回归语言建模任务和掩码语言建模任务进行训练。最近，一些学者开始将 Attention 机制应用于虚假信息检测任务中。例如，Chen 等[36]建议使用深度注意力模型，基于循环神经网络(RNN)来有选择地学习序列帖子的时间隐藏表示，从而帮助识别虚假信息；此外，还引入了一种复杂的分层多模态注意力网络[15]，采用两个 Transformer 单元来联合建模多模态上下文数据，用于虚假信息检测。

受 Attention 机制成功应用的启发，本文引入了一种多模态 Gather Transformer 网络，以合并多模态特征，全面捕捉模态之间的细粒度关系，同时筛选和提取模态中的关键信息。

3 MDCGTN 模型方法

3.1 任务描述

虚假信息检测任务可以被定义为一个二元分类问题，其主要关注社交媒体上的帖子是否为虚假信息。给定社交媒体上由文本信息和相应图片组成的多模式帖子 P，该模型将输出 $Y=\{0,1\}$来表示帖子的标签，其中 $Y=0$ 和 $Y=1$ 分别表示帖子是真实信息和虚假信息。

3.2 整体框架

图 1 展示了本文模型的总体框架。我们在模型中引入了多模态双协同 Gather Transformer 网络，通过充分融合和协同文本、图像与评论间的信息，并有效筛选模态信息中的关键信息来提高虚假信息检测任务的性能。

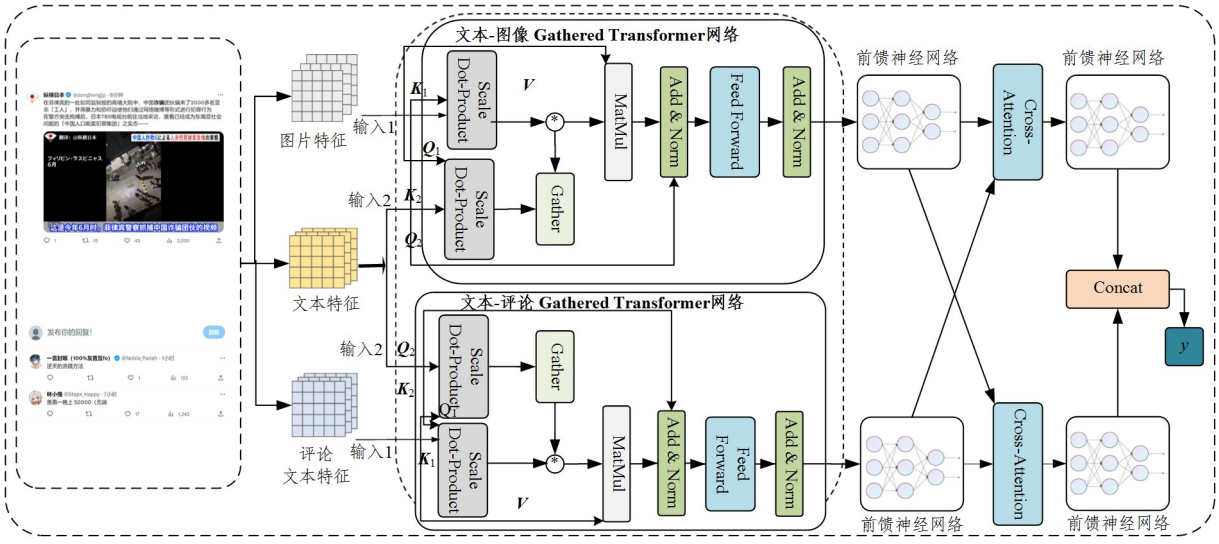

图 1 MDCGTN 模型

Fig. 1 MDCGTN model

整个模型由以下组件构成：

1)文本-图像编码网络：对于给定帖子的文本和图像信息，分别使用 BERT 和 ResNet 提取文本内容和视觉内容的特征。

2)多模态 Gather Transformer 网络：由于不同模态信息间具有细腻的内在联系，因此提出了一种双协同 Gather Transformer 网络，用于充分融合和协同多模态细粒度特征。

3)双协同机制：由前馈神经网络和 Cross-attention 网络

构成，用于提高文本、视觉和用户评论间的相互指导关系以及多模态信息间的互动增强能力。

4）虚假信息检测器：旨在将社交媒体帖子分为真实和虚假。它采用全连接层，并配有相应的激活函数，通过生成一个预测概率，最终确定帖子内容的真实性。

3.3　文本-图像编码网络

正如任务描述中所述，本文模型输入为多模态新闻 $P=\{T,V,C\}$。其中，T 表示文本内容，V 表示视觉内容，C 表示对应的用户评论。为了更好地获得社交媒体帖子细粒度的图像和文本特征表示，对于文本信息，使用预训练模型 BERT 来进行文本特征提取；对于图像信息，使用 ResNet50 来获取图像特征表示。

3.3.1　文本编码网络

为了精确地获取文本语义和上下文信息，我们采用 BERT 作为文本处理模型的核心模块。BERT 已被证明在许多领域是有效的，如问题回答、翻译、阅读理解和文本分类等任务。

给定文本内容 T，将 T 建模为单词序列 $T=\{w_1,w_2,\cdots,w_m\}$（m 表示文本中的单词数），将处理后的文本特征表示为 $P_T=\{P_{t1},P_{t2},\cdots,P_{tm}\}$。其中，$P_{ti}$ 表示第 i 个单词 w_i 的特征。词的表示方法 P_{ti} 是由预先训练的 BERT 模型计算的：

$$P_T=\{P_{t1},P_{t2},\cdots,P_{tm}\}=BERT(W) \tag{1}$$

其中，$P_{ti}\in\mathbb{R}^{d_t}$ 是 BERT 中对应标记的输出层隐藏状态，d_t 为词嵌入的维度。

3.3.2　视觉编码网络

对给定视觉内容 V，使用自下而上的注意力预训练模型 ResNet50 来提取区域特征。输出是一组区域特征 $P_V=\{P_{v1},P_{v2},\cdots,P_{vn}\}$（$n$ 表示图像中的区域数量），其中每个 P_{vj} 被定义为第 j 个区域的平均池化卷积特征。在训练期间，预训练模型参数保持固定。同时，在处理给定的视觉内容 V 的过程中，视觉特征提取器中倒数第二层池化层的操作可以表示为：

$$P_V=\{P_{v1},P_{v2},\cdots,P_{vn}\}=ResNet50(V) \tag{2}$$

其中，$P_{v_j}\in\mathbb{R}^{d_v}$，$d_v$ 表示图像嵌入层的维度。此外，通过另外计入 2D 卷积层，将嵌入维度 d_v 调整为 d_t，以满足任务需求。

3.3.3　评论编码网络

对于给定评论内容 C，我们同样使用 BERT 提取评论文本特征。给定内容评论 $C=\{c_1,\cdots,c_o\}$（o 表示评论数量），转换后得到评论特征，用 $P_{C_h}=\{P_{c_1},\cdots,P_{c_o}\}$ 表示。其中，每个 P_{C_h} 对应第 h 个评论 C_h 的特征，评论特征 P_{C_h} 通过训练的 BERT 计算得到：

$$P_{C_h}=BERT(c_h) \tag{3}$$

其中，$P_{C_h}\in\mathbb{R}^{dt}$ 是 $BERT$ 中隐藏层的池化输出，d_t 表示单词嵌入的维度。

3.4　多模态 Gather Transformer 网络

为了有效整合社交媒体帖子中文本、视觉和评论特征，本文设计了多模态双协同 Gather Transformer 网络来构建多模态上下文信息，并充分捕捉和融合高阶互补信息。如图 1 所示，双协同 Gather Transformer 网络由文本-图像 Gather Transformer、文本-评论 Gather Transformer 两个模块构成。

在 Transformer 模型中，Attention 机制是不可或缺的，它构成了 Sequence-to-Sequence 处理中编码器-解码器架构的核心，有利于建立输入与输出之间的长程依赖关系。对于长度为 L 的输入序列 $X\in\mathbb{R}^{L*D}$，单头 self-attention 机制的表达式如下：

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{4}$$

其中，softmax(·)和 $\boldsymbol{V}$ 的乘积表示由查询值 $\boldsymbol{Q}$ 和键值 $\boldsymbol{K}$ 之间的交互得到的注意力分数对输入的加权组合。分母中缩放因子 $\sqrt{d}$ 能有效抑制点积幅度增长，进而稳定度量标准。

为了检索和捕捉模态信息中的关键部分，充分理解和提取模态之间细粒度相互关系，本文通过添加一个多模态 Gather 网络来增强传统 Transformer 网络的多头注意力架构。

3.4.1　文本-图像 Gather Transformer 网络

为了充分捕捉和检索文本与图像模态信息间细粒度内在联系，我们设计了文本-图像 Gather Transformer 网络。

将经过预训练后得到的视觉特征 P_V 和文本特征 P_T 分别作为 input1 和 input2（如图 1 所示）输入到文本-图像 Gather Transformer 网络当中。在自注意力机制中，对于多模态输入 $\boldsymbol{P}_V=\{P_{v1},P_{v2},\cdots,P_{vn}\}\in\mathbb{R}^{d_v}$ 和 $\boldsymbol{P}_T=\{P_{t1},P_{t2},\cdots,P_{tm}\}\in\mathbb{R}^{d_t}$，计算相似度矩阵 $\boldsymbol{S}$：

$$\boldsymbol{S}=\mathrm{Softmax}\left(\frac{\boldsymbol{P}_T\boldsymbol{P}_V^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{5}$$

其中，$S_{i,j}$ 表示 P_T 中第 i 个单词与 P_V 中第 j 个区域之间的相似度得分。将相似度矩阵 $\boldsymbol{S}$ 展开为元素个数为 N（N 为 n 与 m 的乘积）的线性序列 s：

$$s=(s_0,s_1,\cdots,s_{N-1}) \tag{6}$$

$$s_g=gather(s) \tag{7}$$

通过引入 Gather 机制，筛选线性序列 s 中元素权重最高的 K 个元素，得到新的线性序列 s_g。最后，将序列 s_g 还原为经 Gather 机制处理后的相似度矩阵 $\boldsymbol{S}_g$。为方便理解，图 2 给出了 Gather 机制的具体实现流程。

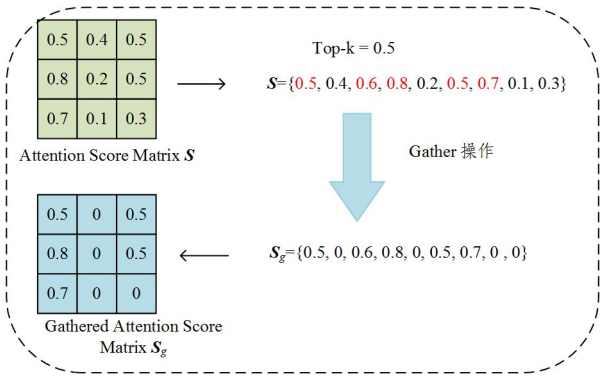

图 2　Gather 机制处理流程

Fig. 2　Process of Gather mechanism

基于此，我们将文本-图像 Gather Transformer 网络中修改后的单头多模态自注意力机制定义为：

$$GatheredAtt(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{G})=\mathrm{Softmax}\left(\frac{\mathbb{G}(\boldsymbol{QK}^{\mathrm{T}})}{\sqrt{d}}\right)\boldsymbol{V} \tag{8}$$

其中，$\boldsymbol{Q}=\boldsymbol{W}_Q^1\boldsymbol{P}_T$，$\boldsymbol{K}=\boldsymbol{W}_K^1\boldsymbol{P}_V$，$\boldsymbol{V}=\boldsymbol{W}_V^1\boldsymbol{P}_V$；$\boldsymbol{W}_Q^1$，$\boldsymbol{W}_K^1$，$\boldsymbol{W}_V^1$ 分别表示将输入投影到查询、键和值的不同线性变换；符号 $\mathbb{G}(\cdot)$ 表示 Gather 机制的处理操作。

3.4.2 文本-评论 Gather Transformer 网络

为了充分捕捉和检索文本与评论模态信息间细粒度内在联系，我们设计了文本-评论 Gather Transformer 网络。

将经过预训练后得到的评论特征 P_{C_h} 和文本特征 P_T 分别作为 input1 和 input2（如图 1 所示）输入到文本-评论 Gather Transformer 网络当中。在自注意力机制中，对于多模态输入 $\boldsymbol{P}_{C_h}=\{P_{c_1},\cdots,P_{c_o}\}\in\mathbb{R}^{d_t}$ 和 $\boldsymbol{P}_T=\{P_{t1},P_{t2},\cdots,P_{tm}\}\in\mathbb{R}^{d_t}$，可计算相似度矩阵 $\boldsymbol{S}$：

$$\boldsymbol{S}=\mathrm{Softmax}\left(\frac{\boldsymbol{P}_T\boldsymbol{P}_C^{\mathrm{T}}}{\sqrt{d_t}}\right) \tag{9}$$

其中，$S_{q,p}$ 表示 P_T 中第 q 个单词与 P_C 中第 p 个评论之间的相似度得分。接下来，Gather 机制的筛选、检索流程与 3.4.1 节相似。

基于此，将文本-评论 Gather Transformer 网络中改进后的单头多模态自注意力机制定义为：

$$GatheredAtt(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{G})=\mathrm{Softmax}\left(\frac{\mathbb{G}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})}{\sqrt{d}}\right)\boldsymbol{V} \tag{10}$$

其中，$\boldsymbol{Q}=\boldsymbol{W}_Q^2\boldsymbol{P}_T$，$\boldsymbol{K}=\boldsymbol{W}_K^2\boldsymbol{P}_C$，$\boldsymbol{V}=\boldsymbol{W}_V^2\boldsymbol{P}_C$；$\boldsymbol{W}_Q^2$，$\boldsymbol{W}_K^2$，$\boldsymbol{W}_V^2$ 分别表示将输入投影到查询、键和值的不同线性变换；符号 $\mathbb{G}(\cdot)$ 表示 Gather 机制的处理操作。

3.5 双协同机制

设文本-图像 Gather Transformer 的输出为 $\boldsymbol{P}_{TV}$，文本-评论 Gather Transformer 的输出为 $\boldsymbol{P}_{TC}$。为了探索文本、图像和评论之间的内在关系，我们通过两个 Cross-Attention 网络进一步对 $\boldsymbol{P}_{TV}$ 和 $\boldsymbol{P}_{TC}$ 进行编码，使其充分协同和交互。

如图 1 所示，对于上方的 Cross-Attention，将 $\boldsymbol{P}_{TV}$ 作为输入（即 $\boldsymbol{Q}$ 值），$\boldsymbol{P}_{TC}$ 作为输入（即 $\boldsymbol{K}$ 值和 $\boldsymbol{V}$ 值），对应的 Attention 操作如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{P}_{TV}\boldsymbol{P}_{TC}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{P}_{TC} \tag{11}$$

与此对应，下方的 Cross-Attention 操作，将 $\boldsymbol{P}_{TC}$ 作为输入（即 $\boldsymbol{Q}$ 值），$\boldsymbol{P}_{TV}$ 作为输入（即 $\boldsymbol{K}$ 值和 $\boldsymbol{V}$ 值）。

3.6 虚假信息检测器

在双协同机制中，Cross-Attention 模块输出两个多模态表示 F_{ca1} 和 F_{ca2} 后，将其连接起来得到最终特征表示。

$$F_f=Concat(F_{ca1}+F_{ca2}) \tag{12}$$

其中，虚假信息检测器以最终多模态表示 F_f 作为输入，它包含相应激活函数的全连接构成。

$$\hat{y}=\mathrm{softmax}(W*F_f+b) \tag{13}$$

其中，softmax(·)表示激活函数，$\hat{y}$ 表示该帖子的预测概率，F_f 表示该帖子的最终多模态表示。为了增强模型检测效果，我们采用交叉熵损失函数 $Loss$ 来训练本模型。

$$Loss=-\sum_{n=1}^{N}y_n\log(\hat{y}_n) \tag{14}$$

其中，N 表示社交媒体帖子的数量，$\hat{y}_n$ 表示第 n 个帖子的预测概率，y_n 表示第 n 个帖子的真实标签。

4 实验

4.1 实验准备

4.1.1 数据集

本文在两个公开可用的数据集 ReCOVery[18] 和 MMCoVaR[17] 上将 MDCGTN 模型与其他基准模型进行有效性验证。表 1 列出了实验中的两个数据集所包含的虚假信息、真实信息、图像和评论数量。其中，ReCOVery 包含从 2020 年 1 月到 2020 年 5 月涵盖多模态信息的 COVID-19 新闻帖子。MMCoVaR 包含特定 COVID-19 疫苗相关的多模态社交媒体帖子，帖子编辑时间为 2020 年 2 月至 2021 年 5 月。根据文献[15]的方法，我们将两个数据集按 8∶2 的比例划分为训练集和测试集。

表 1 ReCOVery 和 MMCoVaR 数据集

Table 1 ReCOVery and MMCoVaR datasets

News Articles	ReCOVery	MMCoVaR
# of Fake News	1364	1635
# of Real News	665	958
# of Images	1675	22357
# of Comments	**140820**	**24183**

4.1.2 评价指标

虚假信息检测任务属于二分类，一般使用准确率（Accuracy）作为主要评估指标，但在样本数据不均衡的情况下，其可靠性可能会降低。因此，在实验过程中，我们增加了精度（Precision）、召回率（Recall）和加权 F1 分数（Weighted F1-score）作为补充评估指标，以解决数据不平衡导致的问题。

4.1.3 实验设置

对于社交媒体帖子多模态嵌入，我们使用预训练得到的 ResNet 来提取视觉特征，使用预训练的 BERT 模型来提取帖子文本和评论中的文本特征。其中，图像嵌入维度为 2048，文本嵌入维度为 768。为了适应本模型，引入了一个 2D 卷积层，将视觉区域特征维度从 2048 转换为 768。本文模型使用自适应矩估计（Adam）优化器进行训练，一共训练 200 个 epoch，学习率为 0.001，小批量设置为 64。

4.2 比较方法

选取了 11 种最先进的模型进行比较：

1）HSA[37]：一种利用媒体中的用户评论和层次化社交网络结构进行虚假信息检测的方法。

2）ExFaux[38]：一种基于图的虚假图像解释框架，可以为检测结果提供内容解释。

3）dEFEND[39]：一种可解释的虚假信息检测方法。它利用新闻文本与用户评论之间的相关性，既可以对虚假新闻进行分类，又可以确定用户评论对分类结果提供解释的方式。

4）BTIC[40]：采用基于 BERT 的多模态框架，用于不可靠新闻检测，使用对比学习策略来利用可疑文章中的文本和视觉信息。

5）SAFE[23]：一种以相似性为中心的多模态方法，用于虚假信息检测。它从新闻素材中提取文本和视觉元素，并探索它们之间的相互关系以获得最终表示。

6）EANN[21]：包括多模态特征提取器、虚假新闻检测器

和后处理鉴别器。

7）SpotFake[41]：使用预训练的 BERT 提取文本特征，同时使用在 ImageNet 数据集上预训练的 VGG-19 提取图像特征，进而辅助判断帖子的真实性。

8）MVAE[14]：通过将多模态特征引入到双模态变分自编码器（VAE）中，获得多模态表示。再经过二元分类器，对双模态 VAE 产生的多模态潜在表示进行分类。

9）FMFN[10]：一种细粒度多模态融合网络，它利用缩放点积注意力机制来融合文本中单词的词嵌入，以及表示图像不同特征的多个特征向量。

10）MMTN[42]：一种多模态掩码 Transformer 网络，它利用掩码 Transformer 网络联合建模多模态信息的模态间和模态内关系，并屏蔽无关上下文信息，实现虚假信息检测任务。

11）DGExplain[15]：一种生成性方法，通过分析视觉和文本信息的模态间联系，进而识别与 COVID-19 相关的多模态帖子的真实性。

4.3　实验结果分析

4.3.1　定量分析

实验结果如表 2 和表 3 所列，通过观察表中的数据可以得出以下结论：

1）基于特征融合的方法（如 MVAE，SpotFake，EANN 和 ExFaux）在两个数据集中表现一般，主要原因在于其忽视了社交媒体帖子中视觉和文本元素之间的复杂关系。它们直接从混合的多模态内容中推断待检测内容的可信度，而不是考虑这种相互关联。

2）DGExplain 性能超过了其他基线方法，原因在于其能精确生成跨模态特征，结合生成的和原始的多模态特征之间的一致性评估，进而辅助检测虚假信息。同时，DGExplain 对内容-评论图的结合有效促进了生成的、原始的多模态特征和用户评论的整合，进而有助于虚假信息的检测。

3）所提出的 MDCGTN 在两个数据集上的性能超过了所有基准方法。结果表明，所提出的基于多模态双协同 Gather Transformer 网络的虚假信息检测方法能够充分捕捉模态间和模态内的细粒度特征关系，检索和收集模态信息中的关键部分，最终达到不错的检测效果。

表 2　不同方法在 ReCOVery 数据集上的实验结果

Table 2　Detection results of different methods on ReCOVery dataset

（%）

Methods	*Acc*	*P*	*R*	*F*1
HSA	77.90	73.70	73.60	73.60
ExFaux	76.30	71.90	69.50	70.40
dEFEND	85.60	82.60	81.30	82.30
BTIC	76.30	71.90	69.50	70.40
SAFE	83.10	80.30	78.90	79.50
EANN	84.70	81.60	83.40	82.40
SpotFake	68.10	63.70	65.00	64.10
MVAE	82.50	81.30	75.50	77.40
FMFN	87.40	85.30	84.50	84.90
MMTN	88.20	88.30	82.90	85.00
DGExplain	89.70	89.00	86.10	87.30
MDCGTN(Ours)	**90.15**	**89.24**	**88.01**	**88.53**

表 3　不同方法在 MMCoVaR 数据集上的实验结果

Table 3　Detection results of different methods on MMCoVaR dataset

（%）

Methods	*Acc*	*P*	*R*	*F*1
HSA	80.30	78.20	78.50	78.40
ExFaux	76.90	78.40	69.40	70.70
dEFEND	85.60	84.70	83.10	83.80
BTIC	82.90	82.30	79.10	80.30
SAFE	78.80	77.30	74.90	75.70
EANN	83.30	81.90	81.00	81.40
SpotFake	69.90	67.00	62.00	62.30
MVAE	81.50	80.50	83.40	80.80
FMFN	87.30	87.10	85.50	86.20
MMTN	88.40	87.70	87.60	87.70
DGExplain	89.50	89.60	87.10	88.10
MDCGTN(Ours)	**90.40**	**90.20**	**89.13**	**89.59**

4.3.2　消融实验

为了验证所提模型的有效性，我们比较了 MDCGTN 的几种变体：

1）MDCGTN ¬ V：去掉视觉信息，仅保留文本信息；

2）MDCGTN ¬ C：去掉评论信息，仅使用帖子文本信息和视觉信息；

3）MDCGTN ¬ G：去掉 Gather 机制，仅使用传统的 Transformer 模块；

4）MDCGTN ¬ D：去掉双协同机制，并利用自注意力网络替换交叉注意力网络。

表 4 和表 5 列出了模型的不同变体在两个数据集上的实验结果。

表 4　模型不同变体在 ReCOVery 数据集上的实验结果

Table 4　Performance comparison of different variants of MDCGTN on ReCOVery dataset

（%）

Methods	*Acc*	*P*	*R*	*F*1
MDCGTN ¬ V	85.47	85.17	81.50	82.50
MDCGTN ¬ C	87.83	87.14	84.42	85.50
MDCGTN ¬ G	88.22	87.59	85.09	86.13
MDCGTN ¬ D	87.98	87.04	85.32	86.28
MDCGTN	**90.15**	**89.24**	**88.01**	**88.53**

表 5　模型不同变体在 MMCoVaR 数据集上的实验结果

Table 5　Performance comparison of different variants of MDCGTN on MMCoVaR dataset

（%）

Methods	*Acc*	*P*	*R*	*F*1
MDCGTN ¬ V	82.35	84.79	78.02	79.54
MDCGTN ¬ C	87.67	87.89	85.95	86.47
MDCGTN ¬ G	87.57	87.48	85.94	86.47
MDCGTN ¬ D	89.15	89.36	87.39	88.12
MDCGTN	**90.40**	**90.20**	**89.13**	**89.59**

根据表 4 和表 5 中的结果，可以得出以下结论：

1）为了评估视觉信息的效果，我们对比了 MDCGTN 和 MDCGTN ¬ V 在两个数据集上的性能。可观察到，所提出的 MDCGTN 优于 MDCGTN ¬ V，这表明视觉信息能为本模型提供有价值的补充信息。

2）通过对比 MDCGTN 和 MDCGTN ¬ C 可以得到 MD-

CGTN 的性能优于 MDCGTN ¬C 的结论，进而表明用户评论能为本模型提供有效的补充信息。

3)为了评估 Gather Transformer 网络的效果，我们对比了 MDCGTN 和 MDCGTN ¬G 的性能。可观察到，所提出的 MDCGTN 优于 MDCGTN ¬G，这表明 Gather Transformer 网络能为本模型提供有价值的补充信息。

4)通过对比 MDCGTN 和 MDCGTN ¬D 可观察到，所提出的 MDCGTN 优于 MDCGTN ¬D，这表明了双协同机制的有效性。

4.3.3 参数分析

图 3 展示了参数 top K 值的变化对模型性能的影响。在 ReCOVery 数据集上，本文模型准确率在 top $K=0.1$ 时到达最佳，在此之后的 top K 值的准确性呈下降趋势，不能再达到同样水平。在 MMCoVaR 数据集上，我们注意到模型的准确率同样在 top $K=0.1$ 时到达最高。因此，可以得出当 top $K=0.1$ 时，本模型能够达到最佳性能。

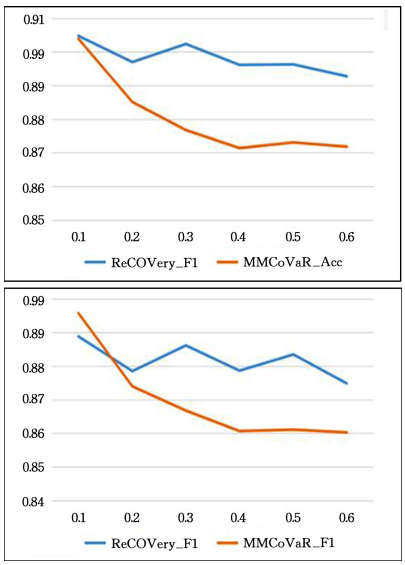

图 3 Top K 值对两个数据集中虚假信息检测的准确率和 F1 分数的影响

Fig. 3 Influence of Top K on accuracy and F1 scores of fake news detection on two datasets

结束语 本文设计了一个多模态双协同 Gather Transformer 模型来对社交媒体虚假信息进行检测。现有的方法大多缺乏对多模态信息中关键部分的提取，未能有效实现多模态信息间的相互指导，多模态信息间缺乏彼此交互性。为了应对上述挑战，本文提出了多模态双协同 Gather Transformer网络来建模多模态信息的模态间和模态内关系，并对关键信息进行筛选。该方法包含 3 个主要组件：1)利用 ResNet 学习视觉表示，并利用 BERT 学习文本表示；2)使用 Gather Transformer 网络更好地探索和捕捉模态间和模态内的细粒度关系，有效检索和筛选模态信息中的关键部分；3)引入一种双协同机制，提高多模态信息间的指导性以及文本与视觉特征间的交互增强能力。在两个基准数据集上的实验表明，本文提出的方法更加有效。下一步将讨论如何更好地利用用户评论信息，以及如何更好地实现多模态信息融合。

参 考 文 献

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

[2] KWON S, CHA M, JUNG K, et al. Prominent features of rumor propagation in online social media[C]// 2013 IEEE 13th International Conference on Data Mining. IEEE, 2013: 1103-1108.

[3] LIU X, NOURBAKHSH A, LI Q, et al. Real-time rumor debunking on twitter[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 1867-1870.

[4] MA J, GAO W, WEI Z, et al. Detect rumors using time series of social context information on microblogging websites[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 1751-1754.

[5] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016: 3818-3824.

[6] YU F, LIU Q, WU S, et al. A Convolutional Approach for Misinformation Identification[C]// IJCAI. 2017: 3901-3907.

[7] WANG J, WANG Y C, HUANG M J. False information in social networks: Definition, detection and control[J]. Computer Science, 2021, 48: 263-277.

[8] HAO X, MING L. Deepfake Video Detection Based on 3D Convolutional Neural Networks[J]. Computer Science, 2021, 48(7): 86-92.

[9] PROCTER R, CRUMP J, KARSTEDT S, et al. Reading the riots: What were the police doing on Twitter? [J]. Policing and Society, 2013, 23(4): 413-436.

[10] WANG J, MAO H, LI H. FMFN: Fine-grained multimodal fusion networks for fake news detection[J]. Applied Sciences, 2022, 12(3): 1093.

[11] QIAN S S, ZHANG T Z, XU C S. Survey of Multimedia Social Events Analysis[J]. Computer Science, 2021, 48(3): 97-112.

[12] WU X K, ZHAO T F. Application of natural language processing in social communication: A review and future perspectives [J]. Computer Science, 2020, 47(6): 184-193.

[13] HAN Z M, ZHENG C Y, DUAN D G, et al. Associated Users Mining Algorithm Based on Multi-information Fusion Representation Learning[J]. Computer Science, 2019, 46(4): 77-82.

[14] KHATTAR D, GOUD J S, GUPTA M, et al. Mvae: Multimodal variational autoencoder for fake news detection[C]// The World Wide Web Conference. 2019: 2915-2921.

[15] SHANG L, KOU Z, ZHANG Y, et al. A duo-generative approach to explainable multimodal covid-19 misinformation detection[C]// Proceedings of the ACM Web Conference. 2022: 3623-3631.

[16] LIN H, MA J, CHENG M, et al. Rumor detection on twitter with claim-guided hierarchical graph attention networks[J]. arXiv: 2110.04522, 2021.

[17] CHEN M, CHU X, SUBBALAKSHMI K P. MMCoVaR: multimodal COVID-19 vaccine focused data repository for fake news

detection and a baseline architecture for classification[C]// Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2021:31-38.

[18] ZHOU X, MULAY A, FERRARA E, et al. Recovery: A multimodal repository for covid-19 news credibility research[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020:3205-3212.

[19] SINGHAL S, KABRA A, SHARMA M, et al. Spotfake+: A multimodal framework for fake news detection via transfer learning(student abstract)[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:13915-13916.

[20] LIU J, FENG K, PAN J Z, et al. MSRD: Multimodal Web Rumor Detection Method[J/OL]. https://www.researchgate.net/publication/347001640.

[21] WANG Y, MA F, JIN Z, et al. Eann: Event adversarial neural networks for multi-modal fake news detection[C]// Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. 2018:849-857.

[22] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:7370-7377.

[23] ZHOU X, WU J, ZAFARANI R. Safe: similarity-aware multimodal fake news detection[C]// Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference. 2020:354-367.

[24] WEI Z X, LIANG J M. Design of Image Retrieval System Based on Speech Recognition[J]. Applied Mechanics and Materials, 2012, 220:2371-2374.

[25] XUE J, WANG Y, TIAN Y, et al. Detecting fake news by exploring the consistency of multimodal data[J]. Information Processing & Management, 2021, 58(5):102610.

[26] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// International Conference on Machine Learning. PMLR, 2015:2048-2057.

[27] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2014.

[28] CHEN J, ZHANG H, HE X, et al. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017:335-344.

[29] WANG S, HU L, CAO L, et al. Attention-based transactional context embedding for next-item recommendation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[31] BAEVSKI A, AULI M. Adaptive input representations for neural language modeling[J]. arXiv: 1809.10853, 2018.

[32] DAI Z, YANG Z, YANG Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv: 1901.02860, 2019.

[33] RAE J W, POTAPENKO A, JAYAKUMAR S M, et al. Compressive transformers for long-range sequence modelling[J]. arXiv: 1911.05507, 2019.

[34] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J/OL]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[35] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[36] CHEN T, LI X, YIN H, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]// Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing, 2018:40-52.

[37] GUO H, CAO J, ZHANG Y, et al. Rumor detection with hierarchical social attention network[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018:943-951.

[38] KOU Z, ZHANG D Y, SHANG L, et al. Exfaux: A weakly supervised approach to explainable fauxtography detection[C]// 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020:631-636.

[39] SHU K, CUI L, WANG S, et al. defend: Explainable fake news detection[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019:395-405.

[40] ZHANG W, GUI L, HE Y. Supervised contrastive learning for multimodal unreliable news detection in covid-19 pandemic[C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021:3637-3641.

[41] SINGHAL S, SHAH R R, CHAKRABORTY T, et al. Spotfake: A multi-modal framework for fake news detection[C]// 2019 IEEE fifth International Conference on Multimedia Big Data (BigMM). IEEE, 2019:39-47.

[42] WANG J, QIAN S, HU J, et al. Positive Unlabeled Fake News Detection Via Multi-Modal Masked Transformer Network[J]. IEEE Transactions on Multimedia, 2023, 26:234-244.

XIANG Wang, born in 1996, postgraduate. His main research interests include natural language processing and multimedia computing analysis.

QIAN Shengsheng, born in 1991, Ph.D, associate professor. His main research interests include data mining and multimedia content analysis.

（责任编辑：何杨）