计算机科学
COMPUTER SCIENCE

基于多模态自适应融合的短视频虚假新闻检测

朱枫, 张廷辉, 李鹏, 徐鹤

引用本文

朱枫, 张廷辉, 李鹏, 徐鹤. 基于多模态自适应融合的短视频虚假新闻检测[J]. 计算机科学, 2024, 51(11): 39-46.

ZHU Feng, ZHANG Tinghui, LI Peng, XU He. Multimodal Adaptive Fusion Based Detection of Fake News in Short Videos [J]. Computer Science, 2024, 51(11): 39-46.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700062

基于多模态自适应融合的短视频虚假新闻检测

朱　枫[1]　张廷辉[1]　李　鹏[1,2]　徐　鹤[1,2]
1 南京邮电大学计算机学院　南京 210023
2 江苏省无线传感网高技术研究重点实验室　南京 210023
(zhufeng@njupt.edu)

摘　要　随着互联网和社交媒体的迅速发展,新闻的传播途径不再局限于传统的媒体渠道。语义丰富的多模态数据成为新闻的载体,虚假新闻也随之得到了广泛的传播。由于虚假新闻的泛滥会对个人以及社会产生难以预估的影响,针对虚假新闻的检测已经成为目前的研究热点。现有的多模态虚假新闻检测方法仅针对文本和图像数据,无法充分利用短视频中的多模态信息,且忽略了不同模态间的一致性和差异性特征,难以充分发挥多种模态融合的优势。为解决该问题,提出一种基于多模态自适应融合的短视频虚假新闻检测模型。首先对短视频中多模态数据进行特征提取,采用跨模态对齐融合获取不同模态间的一致性和互补性特征;然后根据不同模态特征对最终融合结果的贡献实现自适应融合;最后利用分类器实现虚假新闻检测。在公开的短视频数据集上的实验结果表明,该模型的准确率、精确率、召回率和F1分数都高于当前的先进基线模型。

关键词:虚假新闻检测;多模态;短视频;跨模态融合;自适应融合
中图分类号　TP391

Multimodal Adaptive Fusion Based Detection of Fake News in Short Videos

ZHU Feng[1],ZHANG Tinghui[1],LI Peng[1,2] and XU He[1,2]
1 College of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
2 Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks,Nanjing 210023,China

Abstract　With the rapid development of Internet and social media,the dissemination route of news is no longer limited to traditional media channels. Semantically rich multimodal data becomes the carrier of news while fake news has been widely spread. As the proliferation of false news will have an unpredictable impact on individuals and society,the detection of false news has become a current research hotspot. Existing multimodal false news detection methods only focus on text and image data,which not only fail to fully utilize the multimodal information in short videos but also ignore the consistency and difference features between different modalities. As a result,it is difficult for them to give full play to the advantages of multimodal fusion. To solve this problem,a fake news detection model for short videos based on multimodal adaptive fusion is proposed. This model extracts features from multimodal data in short videos,uses cross-modal alignment fusion to obtain the consistency and complementarity features among different modalities,and then achieves adaptive fusion based on the contribution of different modal features to the final fusion result. Finally,a classifier is used to achieve fake news detection. The results of the experiment conducted on a publicly available short video dataset demonstrate that the accuracy,precision,recall,and F1-score of the proposed model are higher than those of the state-of-the-art models.

Keywords　Fake news detection,Multimodal,Short video,Cross-modal fusion,Adaptive fusion

1　引言

如今,互联网已经成为生活中不可或缺的一部分,报纸和电视等获取新闻信息的传统渠道已经无法满足日益增长的需求,基于网络的社交媒体平台已经成为人们日常生活中获取信息必不可少的途径[1]。互联网是社交媒体网站的载体,随着网络技术的进步,信息的创建、共享和传播都在以极快的速度进行。因此,社会新闻等信息能够快速而自由地传播,社会公众可以随时随地不受阻碍地获取到公开的海量信息。截至2024年,在中国有超过70%的人通过网络社交媒体获取新闻等社会公共信息,微博、抖音和快手等短视频社交媒体的受欢迎程度正在呈指数级增长。据报道,2024年5月微博的

到稿日期:2024-07-10　返修日期:2024-08-29
基金项目:国家自然科学基金(61902196,62102196);江苏省科技支撑计划项目(BE2019740);江苏省六大人才高峰高层次人才项目(RJFW-111)
This work was supported by the National Natural Science Foundation of China(61902196,62102196),Scientific and Technological Support Project of Jiangsu Province(BE2019740) and Six Talent Peaks Project of Jiangsu Province(RJFW-111).
通信作者:李鹏(lipeng@njupt.edu.cn)

月活跃用户达5亿，抖音全球月活跃用户已经接近10亿，并且随着时间的推移，各个网络社交平台的用户数量还在持续增长。

由于信息变得触手可及和完备审查机制的匮乏，网络社交媒体上各种信息的质量远远低于传统媒体方式，整个网络环境中充斥着虚假新闻，信息的可信度成为了我们面临的新问题。虚假新闻指恶意用户在网络媒体上发布的不准确和虚假的信息，它通常以虚假或带有偏见的故事误导人们，以操控社会公众情绪，影响他人的思想和行为，进而达到谋取私利的目的。虚假新闻的泛滥严重影响社会舆论，甚至会扰乱正常秩序、引发恐慌，对个人乃至整个社会都会产生巨大的负面影响[2]。例如，2010年智利地震发生后，一些虚假新闻在Twitter上传播，加剧了当地民众的恐慌和混乱[3]。这些谣言不仅引发了广泛的误解，还导致救援工作复杂化，并严重干扰了社会秩序。此外，假新闻的传播也对经济有重大影响。例如，美国联合通讯社被黑客攻击后发布了一则关于爆炸的假新闻，造成了股市的剧烈震荡，导致价值1 300亿美元的股票瞬间蒸发，这巨大的经济损失引起了人们对假新闻的激烈讨论[4]。

如何解决虚假新闻在网络社交媒体上的广泛传播是一个具有挑战性的问题。目前已经有许多研究人员对虚假新闻的识别方法展开研究，并取得了一系列有价值的成果。最初，虚假新闻检测主要依靠机器学习方法进行分析，Xu等[5]研究基于Jaccard相似度度量方法来探索虚假新闻和真实新闻之间的文档相似程度，实现对真假新闻的区分。Li等[6]构建了关于糖尿病和中医的数据集，提出了虚假信息的5个特征，并采用朴素贝叶斯模型对虚假内容进行检测。随着人工智能技术的进步，深度学习算法在各种研究中被广泛应用，涌现出了许多新的虚假新闻检测方法。针对文本单模态数据，Liao等[7]提出一种假新闻检测多任务学习模型，通过引入新闻的主题和语境信息，提升对虚假新闻文本的检测性能。除了文本模态以外，也有研究表明图像模态可以用作虚假新闻检测[8]。随着网络社交媒体的发展，多模态信息成为人们获取新闻的主体，因此针对虚假新闻多模态内容的检测变得更为重要。Raj等[9]提出一种多模态耦合卷积神经网络架构，利用卷积神经网络(CNN)对文本和视觉特征进行分析，有效地对在线新闻进行分类。Amri等[10]提出一个可解释的基于多模态内容的模型，通过学习文本和视觉特征的联系进行虚假新闻检测。此外，有研究[11]采用双向长短期记忆网络(Bidirectional Long Short-Term Memory，BiLSTM)提取文本特征，并使用残差网络(Residual Network，ResNet)融合视觉特征来增强语义，实现了对微博虚假信息的检测。

以上研究在虚假新闻检测领域取得了不错的成果，但仍然存在诸多挑战。首先，大多数早期的研究都集中在从单模态数据(文本)中提取特征信息。一些近期的研究开始关注多模态数据，并提出相关虚假新闻检测的方法。但这些方法仅针对文本和图像信息，不能很好地适用于短视频这一主要的多模态信息载体，如图1所示。因此需要研究面向短视频的多模态虚假新闻检测方法。

其次，现有的虚假新闻检测方法主要基于多模态的特征级融合及决策级融合，无法充分利用跨模态数据的一致性和互补性信息。一致性和互补性信息指不同模态数据表达的相似性信息和相互补充的信息，这些信息能够增强模型的整体性和全面性。

图1 新闻载体对比

Fig. 1 News carrier comparison

此外，目前一些关于虚假新闻分类的研究将不同模态的所有信息都放在同样重要的位置，缺少对重要模态数据的关注。因此，最大化不同模态之间的相关性融合，对于增强短视频虚假新闻检测效果至关重要。

通过对上述挑战的分析，并受已有工作的启发，本文提出了一种基于多模态自适应融合的短视频虚假新闻检测模型。该模型主要针对短视频，对包括视频标题、文本描述、图像、声音、用户评论和用户信息等在内的多模态数据进行特征提取和自适应融合，最终实现对虚假新闻的检测。本文的主要贡献总结如下：

1)针对多模态短视频中的异构数据，采用基于注意力机制的跨模态融合方法，以文本为中心充分结合不同模态间的一致性和互补性特征，获得丰富的视频语义信息；

2)聚焦关键模态信息，采用差异化融合的方式，通过自适应融合机制重点提取各个模态的关键特征，进一步提升了虚假新闻检测的效果；

3)在公开的FakeSV短视频数据集上进行实验，结果表明所提模型在多个评价指标上均高于当前先进的基线模型。

本文第2章对相关工作进行阐述；第3章介绍了所提模型；第4章进行实验并对实验结果进行分析；最后总结全文并展望未来。

2 相关工作

2.1 单模态检测方法

单模态虚假新闻检测主要倾向于从新闻文本或图像中提取有价值的信息，通过机器学习或深度学习的方法获取丰富的语义特征，进而提升检测效果。针对递归神经网络(Recursive Neural Network，RNN)存在缺陷的问题，Yu等[12]提出一种基于CNN的虚假信息识别方法，有效地提取分散在输入语句中的关键特征，提高了检测任务的准确率。同样针对单模态文本数据，Mohtarami等[13]提出了一个端到端的记忆网络模型。该模型结合了CNN和RNN的优点，并在推理层引入相似矩阵，准确地提取了输入文本片段的特征信息。为了提升谣言检测的鲁棒性，Ma等[14]受到生成对抗网络的启发，将生成器用于产生不确定的噪声，使输入数据复杂化，从而加强判别器学习更强特征表示的能力。这些方法仅仅关注

了新闻的文本内容，而忽略了其他重要的文本信息。Dou 等[15]研究了利用社交用户偏好进行虚假新闻检测的问题，证明了用户的历史社交活动提供了关于新闻偏好的丰富信息，且可以提高虚假新闻的检测能力。Kausar 等[16]则提出了一种混合假新闻检测模型，通过结合机器学习和深度学习，有效地保留了上下文的特征信息。此外，针对多领域假新闻，Yu 等[17]提出了一种历史新闻环境感知框架，在考虑上下文因素的基础上设计了感知识别模块，并结合域融合实现了对多领域虚假新闻检测的显著改进。Song 等[18]考虑到信息传播的动态性，提出了一种基于动态传播图的虚假新闻检测方法，通过捕获静态网络中缺失的动态传播信息实现对虚假新闻的分类。然而，新闻不仅包含文本信息，还有大量的图像在社交媒体上传播。Qi 等[19]提出了一种多域视觉神经网络框架，从新闻图像中提取频率域和像素域的特征进行融合，提升了虚假新闻检测效果。以上方法有效地实现了单模态虚假新闻检测，但对多模态虚假新闻的识别能力不佳。

2.2　多模态检测方法

随着技术的发展，信息以多种形式存在，网络社交媒体中多模态数据成为信息传播的主流。为了克服单模态方法的局限性，越来越多的研究人员开始关注多模态虚假新闻检测。Liang 等[20]采用多层 CNN 实现文本和视觉特征的混合融合，有效提高了虚假信息检测的准确率。同样针对文本和图像数据，Ye 等[21]结合社交网络图实现多模态虚假信息检测，在检测效率和准确率方面均有提高。为解决数据的时间敏感性和大量性给虚假新闻检测带来的问题，Qu 等[22]提出一种基于量子多模态融合的假新闻检测模型。该模型将文本和视觉特征通过量子卷积神经网络进行传递，并通过融合特征获取检测结果。然而，社交媒体中的多模态数据不仅限于文本和图像，其他形式的数据信息同样不可忽视。Rezayi 等[23]提出一种利用文本、主题标签、转发数和收藏数等结构特征的假新闻分类方法，采用后期融合的方式将各个特征相互连接并完成预测，进一步提升了假新闻分类性能。Gôlo 等[24]则提出一种从文本、主题和语言等组合数据中学习特征表示的方法，然后使用单分类完成虚假新闻检测。此外，大多数关于虚假新闻检测的探索只是针对某个特定领域进行研究，如果新闻信息来自不同的领域（如政治、娱乐和体育），那么许多方法的性能会显著下降。针对这一问题，Silva 等[25]提出一种用于跨域新闻的多模态虚假新闻检测框架，它将特定领域知识和跨域知识联合保存，并采用无监督学习技术训练模型，最终实现对来自不同领域的假新闻的识别。类似地，Mosallanezhad 等[26]提出基于强化学习的领域自适应假新闻检测模型，将辅助信息纳入领域特征学习，解决了现有方式的局限性问题。不同于传统的多模态数据，多模态短视频中所蕴含的信息更加丰富，吸引了更多研究者的关注。Hou 等[27]专注于研究 YouTube 视频中错误信息的检测，探索使用文本、声音和用户参与特征来开发分类模型，并以此识别错误信息。还有研究者使用迁移学习预训练模型来检测虚假信息视频，取得了较高的效率和准确性[28]。Choi 等[29]则提出了一种基于对抗性学习和主题建模的假新闻视频检测模型，该模型通过构建一个对抗神经网络，从视频、标题/描述和评论 3 种类型的数据中提取有效特征，在检测主题变化的同时判断视频的真假。除此之外，为了从短视频中聚合不同模态的异构信息，Shang 等[30]提出了一个多模态错误信息检测框架。该框架利用字幕从虚假内容中准确捕获关键信息，并有效地学习由视频和声音共同传达的融合特征，在检测 TikTok 虚假短视频实验中取得了显著成果。

2.3　虚假新闻数据集

近年来，关于虚假新闻检测的研究日益增多，与之相关的数据集也在不断丰富。本节对公开的虚假新闻数据集进行介绍，并将相关信息汇总，如表 1 所列。其中，单模态虚假新闻数据集只包括文本内容，Wang 等[31]公开的 LIAR 数据集包含 2007－2017 年间美国不同政党政客发表的约 12800 条声明，旨在用于事实核查和假新闻检测研究。Shahi 等[32]对 Twitter 上虚假信息的传播进行了探索性研究，发现部分假新闻传播速度更快，他们收集并发布的数据集弥补了当前研究领域的空白。类似地，Kochkina 等[33]于 2023 年发布了 Twitter 谣言数据集，与其他现有的数据集不同，它能够更好地评估模型的时间稳定性，并为提升模型的泛化能力提出建议。

不同于单模态数据集，多模态虚假新闻数据集包括文本、图像、视频、评论和元数据等多样化的信息。其中，Twitter 数据集[34]和 Weibo[35]数据集来源于网络社交平台的大量帖子，Fakeddit 数据集[36]包含数万条多种类型的假新闻条目，NewsBag 数据集[37]则是一个用于检测假新闻的多模态数据集。为了更好地检测虚假新闻，Gao 等[38]从新浪微博和推特等网络社交平台收集了文本和图片数据，构建了包含多个领域的中英文多模态虚假新闻数据集（Multi-modal Fake News Dataset，MFND）。此外，Zhou 等收集了 2020 年 1 月至 5 月期间的 2029 条新闻，构建为 ReCOVery 数据集[39]，其中包括文本、图像、新闻来源和发布时间等信息。Qi 等[40]从抖音和快手收集了与虚假新闻有关的短视频，构建了最大的中文假新闻短视频数据集（Fake News Short Video dataset，FakeSV），该数据集不仅包含新闻文本和视频等内容，还包括用户评论、发布者简介和元数据等信息。与其他现有的数据集相比，FakeSV 包含不同领域的丰富数据，为评估虚假新闻检测提供了充分的实验数据。

表 1　虚假新闻数据集

Table 1　Fake news datasets

数据集	发布时间	数据格式	标签数量	语言
Wang 等[31]	2017	文本	6	英文
Shahi 等[32]	2021	文本	2	英文
Kochkina 等[33]	2023	文本	4	英文
Boididou 等[34]	2018	文本、图片	2	英文
Jin 等[35]	2017	文本、图片	2	中文
Nakamura 等[36]	2019	文本、图片	2,3,6	英文
Jindal 等[37]	2020	文本、图片	2	英文
Gao 等[38]	2023	文本、图片	2	中文，英文
Zhou 等[39]	2023	文本、图片和元数据	2	英文
Qi 等[40]	2023	文本、图片、视频和元数据	2	中文

3 模型介绍

3.1 模型概述

针对现有的虚假新闻检测存在的问题，本文提出了基于多模态自适应融合的短视频虚假新闻检测模型，其结构如图 2 所示。将该模型分为 4 个主要部分，分别为：多模态特征提取、跨模态特征融合、多模态自适应融合以及多模态虚假新闻分类器。在多模态特征提取部分，首先将短视频分解为多个不同模态数据，包括声音、文本描述、图像（视频帧和视频片段）、用户评论、视频标题及用户信息；其次分别提取对应的单模态特征，分别为 $\boldsymbol{X}_a$，$\boldsymbol{X}_t$，$\boldsymbol{X}_{v1}$，$\boldsymbol{X}_{v2}$，$\boldsymbol{X}_c$，$\boldsymbol{X}_u$，$\boldsymbol{X}_u$；再次，以文本为中心对不同模态数据进行跨模态对齐融合，充分结合不同模态间的一致性和互补性特征，并进一步处理元数据，得到多模态融合结果 $\boldsymbol{X}_{ta}$，$\boldsymbol{X}_{tv}$，$\boldsymbol{X}_{cc}$，$\boldsymbol{X}_{hu}$；然后在多模态自适应融合部分，根据不同模态特征对最终融合结果的贡献度来实现自适应融合；最后通过虚假新闻分类器完成对短视频虚假新闻的检测。下面将对本文提出的模型进行详细介绍。

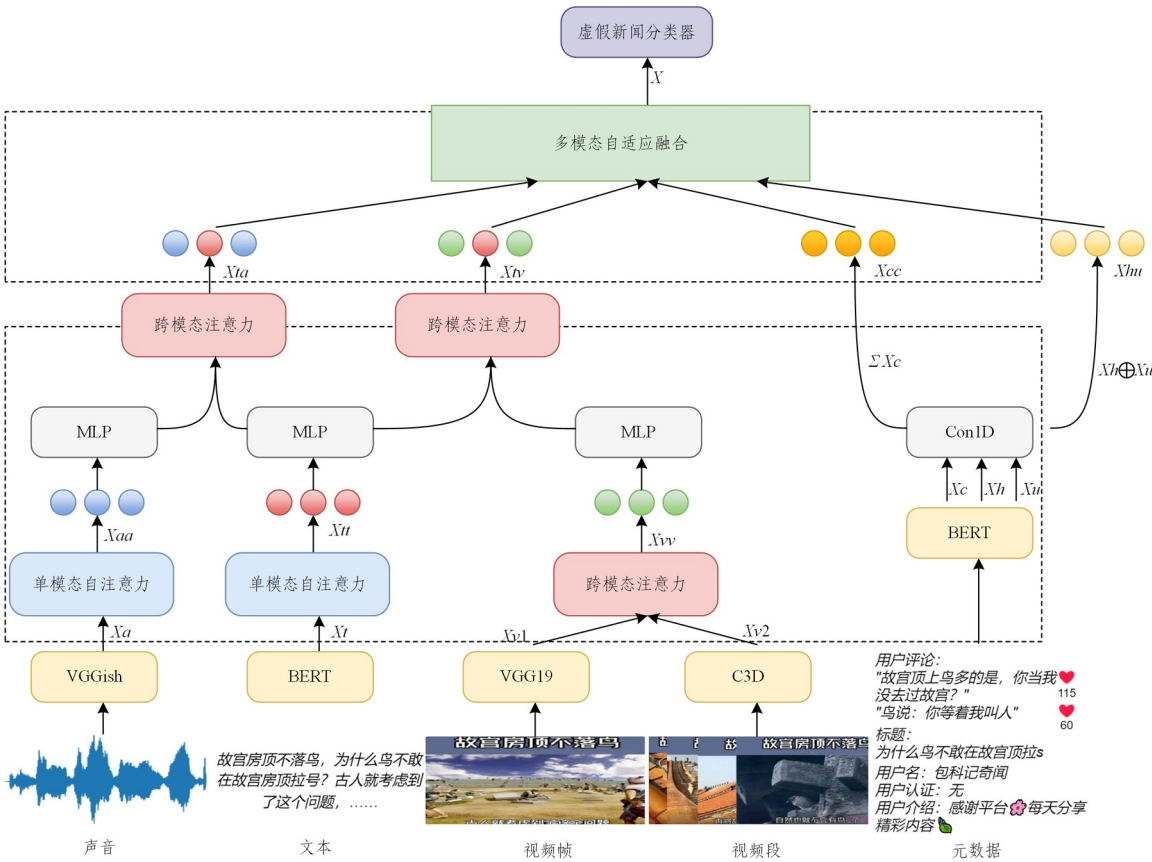

图 2 模型结构

Fig. 2 Model structure

3.2 多模态特征提取

短视频新闻中包含多种模态数据，本节采用不同方式提取 FakeSV 短视频数据集中各个单模态特征。对于声音数据来说，它不仅能够传达语义信息，还能通过环境和背景声音等提供额外的上下文信息，有助于理解说话者的情绪和意图。我们首先将声音从视频中提取出来，然后使用 VGGish 模型将音频预处理为梅尔频谱图，并通过卷积层提取特征 $\boldsymbol{X}_a$。文本数据则包括文本描述、用户评论、视频标题和用户信息，反映了新闻短视频中丰富的语义和个性化的偏好信息。为了保证提取特征的一致性，我们均采用预训练的 BERT 模型来捕捉新闻语义内容和用户评论情感观点，提取模态特征 $\boldsymbol{X}_t$，$\boldsymbol{X}_c$，$\boldsymbol{X}_h$，$\boldsymbol{X}_u$。对于图像数据，选择提取视频帧和视频片段中的视觉特征，以便更好地描述静态和动态信息。具体来说，将视频帧作为输入，经过预训练的 VGG19 模型中的卷积层来提取特征 $\boldsymbol{X}_{v1}$；然后将每个时间步前后的 16 个视频帧作为一个视频片段，并采用 C3D 模型中的 3D 卷积神经网络捕捉动态特征 $\boldsymbol{X}_{v2}$。

3.3 跨模态特征融合

如何有效实现短视频中不同模态之间的融合一直是虚假新闻检测领域的研究热点，但目前的大多数方法仅仅是针对文本和图像模态，无法完全覆盖新闻短视频这一主要信息载体。针对这一问题，本节基于注意力机制提出一种面向短视频的跨模态特征融合方法。考虑到目前针对文本模态的研究较为成熟，且文本信息能够表达新闻语义内容，在提取模态内部特征的同时，以文本为中心充分结合模态间的一致性和互补性特征，获得丰富的新闻多模态信息。

注意力机制能够根据输入序列的不同位置提取对应的特征[41]，我们首先对提取的声音特征 $\boldsymbol{X}_a$ 和文本特征 $\boldsymbol{X}_t$ 使用单模态自注意力机制进一步提取高级特征，计算过程如下：

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

其中，$\boldsymbol{Q}$，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别为查询、键和值，$\boldsymbol{d}$ 代表输入特征的维度。结合多头注意力：

$$\boldsymbol{head}_i = Attention(\boldsymbol{QW}_i^Q, \boldsymbol{KW}_i^K, \boldsymbol{VW}_i^V) \tag{2}$$

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V};\boldsymbol{\theta}) = Concat(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h)\boldsymbol{W}_h^O \tag{3}$$

其中，h 为多头注意力的头数，i 代表第几个头，$\boldsymbol{\theta}=\{\boldsymbol{W}^Q,\boldsymbol{W}^K,\boldsymbol{W}^V,\boldsymbol{W}^O\}$，$\boldsymbol{W}_i^Q \in R^{d^Q\times d^Q}$，$\boldsymbol{W}_i^K \in \boldsymbol{R}^{d^K\times d^K}$，$\boldsymbol{W}_i^V \in \boldsymbol{R}^{d^V\times d^V}$ 分别为多头注意力中不同模态的第 i 个查询、键和值的权重变化矩阵，$\boldsymbol{W}_h^O \in \boldsymbol{R}^{d\times d}$ 是连接权重矩阵，$\boldsymbol{d}^Q=\boldsymbol{d}^K=\boldsymbol{d}^V=\dfrac{d}{h}$ 为每个头的特征维度。

最后基于式(1)－式(3)可以提取声音和文本模态的高级特征，如式(4)所示：

$$\boldsymbol{X}_{mm} = MultiHead(\boldsymbol{X}_m,\boldsymbol{X}_m,\boldsymbol{X}_m;\boldsymbol{\theta}_m),\boldsymbol{m}\in\{\boldsymbol{a},\boldsymbol{t}\} \tag{4}$$

对于视觉特征 $\boldsymbol{X}_{v1}$ 和 $\boldsymbol{X}_{v2}$，则采用跨模态注意力机制来进行对齐融合，如式(5)所示：

$$\boldsymbol{X}_{vv} = \mathrm{Softmax}\left(\frac{\boldsymbol{X}_{v2}\boldsymbol{W}_{vv}^Q\boldsymbol{W}_{vv}^{K^T}\boldsymbol{X}_{v1}^T}{\sqrt{\boldsymbol{d}}}\right)\boldsymbol{X}_{v1}\boldsymbol{W}_{vv}^V \tag{5}$$

其中，$\boldsymbol{W}_{vv}^Q \in \boldsymbol{R}^{d^{v2}\times d}$，$\boldsymbol{W}_{vv}^K \in \boldsymbol{R}^{d^{v1}\times d}$，$\boldsymbol{W}_{vv}^V \in \boldsymbol{R}^{d^{v1}\times d}$ 为线性变化权重矩阵。

为了提取多模态间更加复杂的特征，先将 $\boldsymbol{X}_m$ 经过 MLP 层，其中 $\boldsymbol{n}\in\{\boldsymbol{a},\boldsymbol{t},\boldsymbol{v}\}$，然后以文本为中心采用跨模态注意力机制来融合不同模态，如式(6)和式(7)所示：

$$\boldsymbol{X}_{ta} = \mathrm{Softmax}\left(\frac{\boldsymbol{X}_{aa}\boldsymbol{W}_{ta}^Q\boldsymbol{W}_{ta}^{K^T}\boldsymbol{X}_{tt}^T}{\sqrt{\boldsymbol{d}}}\right)\boldsymbol{X}_{tt}\boldsymbol{W}_{ta}^V \tag{6}$$

$$\boldsymbol{X}_{tv} = \mathrm{Softmax}\left(\frac{\boldsymbol{X}_{vv}\boldsymbol{W}_{tv}^Q\boldsymbol{W}_{tv}^{K^T}\boldsymbol{X}_{tt}^T}{\sqrt{\boldsymbol{d}}}\right)\boldsymbol{X}_{tt}W_{tv}^V \tag{7}$$

其中，$\boldsymbol{W}_{ta}^Q \in \boldsymbol{R}^{d^{aa}\times d}$，$\boldsymbol{W}_{ta}^K \in \boldsymbol{R}^{d^{tt}\times d}$，$\boldsymbol{W}_{ta}^V \in \boldsymbol{R}^{d^{tt}\times d}$，$\boldsymbol{W}_{tv}^Q \in \boldsymbol{R}^{d^{vv}\times d}$，$\boldsymbol{W}_{tv}^K \in \boldsymbol{R}^{d^{tt}\times d}$，$\boldsymbol{W}_{tv}^V \in \boldsymbol{R}^{d^{tt}\times d}$ 为线性变化权重矩阵。

考虑到用户信息与每个视频标题的关联，把标题特征 $\boldsymbol{X}_h$ 和用户信息特征 $\boldsymbol{X}_u$ 整合作为新的融合特征 $\boldsymbol{X}_{hu}$，如式(8)所示：

$$\boldsymbol{X}_{hu} = Concat(\boldsymbol{X}_h,\boldsymbol{X}_u) \tag{8}$$

对于用户评论特征 $\boldsymbol{X}_c$，则使用点赞数来衡量不同评论的重要性，并采用平滑的指数归一化方式来突出高点赞的评论特征，如式(9)所示：

$$\boldsymbol{X}_{cc} = \sum_i^n \frac{\boldsymbol{e}^{s_i+\epsilon}}{\sum_i^n \boldsymbol{e}^{s_i+\epsilon}}\boldsymbol{X}_c \tag{9}$$

其中，$\boldsymbol{n}$ 为评论总数，$\boldsymbol{s}_i$ 为第 $\boldsymbol{i}$ 条评论点赞数，ϵ 为平滑常数。

3.4　多模态自适应融合

目前，多数虚假新闻检测方法将多种不同模态特征放在同样重要的位置进行融合，但由于不同模态数据所包含的信息量不同，这种方式不仅无法关注到重要特征，而且存在引入无关噪声的局限性[42]。此外，也有研究表明不同模态特征对最终融合结果的贡献度不同，对不同模态分配不同权重能够有效提高模型的效果。

因此我们采用多模态自适应融合的方式，结合目标权重以聚焦关键模态特征。具体来说，首先将多模态特征作为输入，经过线性变换和非线性函数的处理，最后通过 Softmax 函数计算不同模态的权重结果，如式(10)所示。这些权重值反映了各模态在最终决策中的重要性。

$$\boldsymbol{\omega}_f = \mathrm{Softmax}(\tanh(\boldsymbol{\omega}_f\boldsymbol{X}_f+\boldsymbol{b}_f)),\boldsymbol{f}\in\{\boldsymbol{ta},\boldsymbol{tv},\boldsymbol{cc},\boldsymbol{hu}\} \tag{10}$$

其中，$\boldsymbol{\omega}_f$ 为不同模态的权重，$\boldsymbol{W}_f$ 和 $\boldsymbol{b}_f$ 为对应参数。然后根据权重计算多模态融合结果 $\boldsymbol{X}$，如式(11)所示：

$$\boldsymbol{X} = \sum \boldsymbol{\omega}_f\boldsymbol{X}_f \tag{11}$$

3.5　虚假新闻分类器

在完成多模态自适应融合后，将融合结果 X 输入一个带有 Softmax 激活函数的全连接网络中得到虚假新闻检测分数，并将其转换为虚假新闻分类结果，如式(12)和式(13)所示：

$$\hat{\boldsymbol{Y}} = \mathrm{Softmax}(\boldsymbol{W}_y\boldsymbol{X}+\boldsymbol{b}_y) \tag{12}$$

$$\boldsymbol{L} = \mathrm{argmax}(\hat{\boldsymbol{Y}}),\boldsymbol{L}\in\{0,1\} \tag{13}$$

其中，$\boldsymbol{W}_y$ 和 $\boldsymbol{b}_y$ 为可学习的权重和偏置项。对于短视频虚假新闻检测的损失 $\boldsymbol{Loss}$ 则采用二分类交叉熵损失函数来计算，如式(14)所示：

$$\boldsymbol{Loss} = \sum_i^n (\boldsymbol{Y}_i \log \hat{\boldsymbol{Y}}_i + (1-\boldsymbol{Y}_i)\log(1-\hat{\boldsymbol{Y}}_i)) \tag{14}$$

其中，n 为每批次中样本数量大小，$\boldsymbol{Y}_i$ 表示真实标签，$\hat{\boldsymbol{Y}}_i$ 表示预测值。

4　实验与分析

4.1　数据集

在公开的 FakeSV 数据集[40]上进行实验，并将本文所提出的模型与其他基线模型进行对比。该数据集由从我国的两个短视频平台(抖音和快手)中抓取的 2019－2022 年的相关视频组成，并采用人工注释的方法进行标注，是当前最大的中文假新闻短视频数据集，其中包含了由 738 个事件衍生的 1 827 个假新闻短视频、1 827 个真新闻短视频和 1 884 个辟谣的短视频。不同于传统数据集仅包含文本和图像，该数据集还包含声音、用户评论、发布者简介和元数据等信息，为各种与虚假新闻检测相关的研究提供了支持。

4.2　实验设置

实验环境为一台操作系统为 Ubuntu 20.04 的服务器，其处理器为 Intel(R) Xeon(R) Platinum 8368 CPU 和 NVIDIA A800 GPU。根据已有的经验，相关的实验参数设置如下：学习率设置为 0.000 1，批量大小为 128，训练轮次设置为 50，采用自适应矩估计优化器来实现模型损失函数最小化的目标。

4.3　评价指标

实验采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 分数 4 个指标对模型进行评估。为了更好地计算评价指标，引入混淆矩阵中的真阳性、真阴性、假阳性和假阴性指标，如表 2 所列。

表 2　混淆矩阵描述

Table 2　Description of confusion matrix

分类	实际为正	实际为负
预测为正	TP	FP
预测为负	FN	TN

准确率、精确率、召回率和 F1 分数的计算方法如式(15)－式(18)所示：

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{15}$$

$$Precision=\frac{TP}{TP+FP} \tag{16}$$

$$Recall=\frac{TP}{TP+FN} \tag{17}$$

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{18}$$

4.4 基线模型

为了从单模态和多模态方面验证本文模型的有效性，我们选择了 6 种在公开数据集上表现出优秀性能的虚假信息检测模型进行对比实验，具体如下：

1) VGGish[43]：针对声音单模态，通过提取视频中的声音特征，并采用全连接层实现对虚假新闻的分类。

2) Text-CNN[44]：针对文本单模态，通过提取文本描述特征实现假新闻分类。

3) VGG19[45]+Attention：针对图像单模态，抽取视频帧特征，结合注意力机制实现假新闻检测。

4) FANVM[29]：一种基于对抗性学习和主题建模的多模态虚假新闻检测模型。该模型通过视频的标题、文本描述和评论评估新闻的主题分布，结合对抗神经网络识别虚假新闻。

5) TikTec[30]：一个多模态错误信息检测框架，它从视频中提取音频，从视频帧中提取字幕，然后通过共同注意力机制有效地提取关键信息。

6) SV-FEND[40]：一个面向短视频的多模态虚假新闻检测模型。该模型对短视频中多种模态特征进行建模，并通过模态融合实现虚假新闻分类。

4.5 对比实验

将本文所提出的虚假新闻检测模型与相关基线模型在 FakeSV 数据集上进行对比实验，结果如表 3 所列。首先与 3 个单模态基线模型进行对比分析，可以看到，针对文本单模态的 Text-CNN 模型在准确率、精确率、召回率和 F1 分数 4 个评价指标上均高于另外两个单模态模型，说明视频中的文本描述在虚假新闻检测中的重要性高于声音和图像模态。

与 3 个单模态模型的实验结果相比，本文模型在 4 个评价指标上均取得了更好的结果，且在假新闻检测的准确率和 F1 分数上比最好的单模态模型还要高出 7.4%以上，在精确率和召回率上则高出 8.4%以上。这些结果说明多模态数据所包含的信息多于单模态数据，有助于虚假新闻检测，验证了本文模型的合理性和有效性。

表 3 对比实验结果

Table 3 Results of comparison experiments

(%)

模态	模型	准确率	精确率	召回率	F1
单模态	VGGish	66.95	67.26	66.81	67.02
	Text-CNN	75.38	74.52	74.36	74.40
	VGG19+Attention	69.47	69.56	69.32	69.43
多模态	FANVM	75.85	75.57	75.32	75.41
	TikTec	75.46	75.34	75.19	75.24
	SV-FEND	80.68	79.79	79.68	79.73
	Ours	**82.78**	**83.06**	**82.76**	**82.93**

在多模态模型的对比实验中，本文提出的模型在假新闻检测的准确率、精确率、召回率和 F1 分数这 4 个评价值指标上同样取得了最优结果。其中，本文模型在准确率和 F1 分数上高出最优的模型 2.1%以上，在精确率和召回率上则高出 3.2%以上。值得注意的是，本文模型与 SV-FEND 模型使用了相同的多模态，但由于本文跨模态对齐融合与多模态自适应融合部分发挥了作用，因此取得了更优的结果。

相比之下，FANVM 和 TikTec 模型使用了更少的模态，虚假新闻检测性能也更差。这些结果说明不同模态数据对多模态虚假新闻检测具有不同的作用，也验证了本文模型的合理性和有效性。

4.6 消融实验

为了进一步验证本文模型在不同模态上的效果，本文基于同样的实验条件进行消融实验，结果如表 4 所列。通过移除原始模型中一些模态来进行消融实验。其中“w/o A”代表移除声音，“w/o T”代表移除文本描述，“w/o V”代表移除图像，“w/o C”代表移除评论和元数据，“w/o U”则代表移除视频标题和用户信息，“ALL”为原始模型。从实验结果可以看出，与原始模型相比，在移除文本描述或移除视频标题和用户信息后，模型的准确率、精确率、召回率和 F1 分数 4 个指标明显降低，其中“w/o U”的准确率降低最多，为 4.41%，“w/o U”和“w/o T”的精确率、召回率和 F1 分数也下降 4%左右。当实验模型为“w/o A”“w/o V”和“w/o C”时，4 个实验指标同样有所下降，但下降程度小于前两种消融实验模型，约为 2%左右。

表 4 消融实验结果

Table 4 Results of ablation experiments

(%)

模型	准确率	精确率	召回率	F1
w/o A	80.93	80.90	80.22	80.56
w/o T	78.66	78.63	78.57	78.60
w/o V	80.67	80.71	79.62	80.16
w/o C	80.29	80.54	79.98	80.27
w/o U	78.37	78.72	78.53	78.61
ALL	**82.78**	**83.06**	**82.76**	**82.93**

这些结果说明不同模态数据对虚假新闻检测具有不同作用，文本描述、视频标题和用户信息对假新闻检测影响较大，声音、图像、评论和元数据的影响则相对较小，验证了多模态自适应融合的合理性。此外，在消融实验中，即使只使用部分模态，本文模型的实验结果依然优于多数基线模型，验证了跨模态对齐融合的有效性。

结束语 本文围绕虚假新闻检测这一核心任务，提出了一种基于多模态自适应融合的短视频虚假新闻检测模型。通过跨模态对齐融合和多模态自适应融合有效地实现特征提取，并由虚假新闻分类器得出最终检测结果。本文模型在对比实验中取得了优于当前先进基线模型的结果，验证了模型的有效性；在消融实验中移除不同模态，验证了模型的合理性。随着互联网及社交媒体的快速发展，多模态数据的内容和形式日新月异，本文提出的方法尽管取得了一定的成果，但仍有改进的空间。在未来的研究中，我们将针对模型缺乏可解释性和鲁棒性的问题，不断优化和改进，进一步提升模型的准确性和适用性，为打击虚假新闻传播提供支持。

参 考 文 献

[1] OLAN F, JAYAWICKRAMA U, ARAKPOGUN E O, et al.

Fake News on Social Media:The Impact on Society[J]. Information Systems Frontiers,2024,26(2):443-458.

[2] TUFCHI S,YADAV A,AHMED T. A Comprehensive Survey of Multimodal Fake News Detection Techniques:Advances,Challenges,and Opportunities[J]. International Journal of Multimedia Information Retrieval,2023,12(2):28.

[3] ZHANG X,GHORBANI A A. An Overview of Online Fake News:Characterization,Detection,and Discussion[J]. Information Processing & Management,2020,57(2):102025.

[4] VOSOUGHI S,ROY D,ARAL S. The Spread of True and False News Online[J]. Science,2018,359(6380):1146-1151.

[5] XU K,WANG F,WANG H,et al. Detecting Fake News over Online Social Media via Domain Reputations and Content Understanding[J]. Tsinghua Science and Technology,2020,25(1):20-27.

[6] LI X,LI S,LI J,et al. Detection of Fake-video Uploaders on Social Media Using Naive Bayesian Model with Social Cues[J]. Scientific Reports,2021,11(1):16068.

[7] LIAO Q,CHAI H,HAN H,et al. An Integrated Multi-task Model for Fake News Detection[J]. IEEE Transactions on Knowledge and Data Engineering,2022,34(11):5154-5165.

[8] CAO J,QI P,SHENG Q,et al. Exploring the Role of Visual Content in Fake News Detection[J]. arXiv:2003.05096,2020.

[9] RAJ C,MEEL P. Convnet Frameworks for Multi-modal Fake News Detection[J]. Applied Intelligence,2021,51(11):8132-8148.

[10] AMRI S,SALLAMI D,AÏMEUR E. Exmulf:An Explainable Multimodal Content-based Fake News Detection System[C]// 14th International Symposium on Foundations and Practice of Security. 2021:177-187.

[11] WANG H,GONG L,ZHOU Z,et al. Detecting Mis/Dis-information from Social Media with Semantic Enhancement[J]. Data Analysis and Knowledge Discovery,2023,7(2):48-60.

[12] YU F,LIU Q,WU S,et al. A Convolutional Approach for Misinformation Identification[C]// 26th International Joint Conference on Artificial Intelligence. 2017:3901-3907.

[13] MOHTARAMI M,BALY R,GLASS J,et al. Automatic Stance Detection Using End-to-End Memory Networks[C]// 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2018:767-776.

[14] MA J,GAO W,WONG K. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning[C]// 2019 The World Wide Web Conference. 2019:3049-3055.

[15] DOU Y,SHU K,XIA C,et al. User Preference-aware Fake News Detection[C]// 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021:2051-2055.

[16] KAUSAR N,ALIKHAN A,SATTAR M. Towards Better Representation Learning Using Hybrid Deep Learning Model for Fake News Detection[J]. Social Network Analysis and Mining,2022,12(1):165.

[17] YU W,GE J,YANG Z,et al. Multi-domain Fake News Detection for History News Environment Perception[C]// 17th IEEE Conference on Industrial Electronics and Applications. 2022:428-433.

[18] SONG C,TENG Y,ZHU Y,et al. Dynamic Graph Neural Network for Fake News detection[J]. Neurocomputing,2022,505:362-374.

[19] QI P,CAO J,YANG T,et al. Exploiting Multi-domain Visual Information for Fake News Detection[C]// 2019 IEEE International Conference on Data Mining. 2019:518-527.

[20] LIANG Y,TOHTI T,HAMDULLA A. Multi-modal False Information Detection via Multi-layer CNN-based Feature Fusion and Multi-classifier Hybrid Prediction[J]. Computer Engineering and Science,2023,45(6):1087-1096.

[21] YE Z,LUO S,YU J. Multimodal Misinformation Detection Model with Social Network Graph[J]. Application Research of Computers,2024,41(7):1-8.

[22] QU Z,MENG Y,MUHAMMAD G,et al. QMFND:A Quantum Multimodal Fusion-based Fake News Detection Model for Social Media[J]. Information Fusion,2024,104:102172.

[23] REZAYI S,SOLEYMANI S,ARABNIA H R,et al. Socially Aware Multimodal Deep Neural Networks for Fake News Classification[C]// 4th International Conference on Multimedia Information Processing and Retrieval. 2021:253-259.

[24] GÔLO M P S,DE SOUZA M C,ROSSI R G,et al. One-class Learning for Fake News Detection through Multimodal Variational Autoencoders[J]. Engineering Applications of Artificial Intelligence,2023,122:106088.

[25] SILVA A,LUO L,KARUNASEKERA S,et al. Embracing Domain Differences in Fake News:Cross-domain Fake News Detection Using Multi-modal Data[C]// 35th AAAI Conference on Artificial Intelligence. 2021:557-565.

[26] MOSALLANEZHAD A,KARAMI M,SHU K,et al. Domain Adaptive Fake News Detection via Reinforcement Learning[C]// 31st ACM World Wide Web Conference. 2022:3632-3640.

[27] HOU R,PÉREZ-ROSAS V,LOEB S,et al. Towards Automatic Detection of Misinformation in Online Medical Videos[C]// 21st International Conference on Multimodal Interaction. 2019:235-243.

[28] SERRANO J C M,PAPAKYRIAKOPOULOS O,HEGELICH S. NLP-based Feature Extraction for The Detection of Covid-19 Misinformation Videos on YouTube[C]// 1st Workshop on NLP for COVID-19 at the 58th Annual Meeting of the Association for Computational Linguistics. 2020:1-7.

[29] CHOI H,KO Y. Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection[C]// 30th ACM International Conference on Information & Knowledge Management. 2021:2950-2954.

[30] SHANG L,KOU Z,ZHANG Y,et al. A Multimodal Misinformation Detector for Covid-19 Short Videos on TikTok[C]//

2021 IEEE International Conference on Big Data. 2021: 899-908.
[31] WANG W Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection[C]// 55th Annual Meeting of the Association for Computational Linguistics. 2017: 422-426.
[32] SHAHI G K, DIRKSON A, MAJCHRZAK T A. An Exploratory Study of Covid-19 Misinformation on Twitter[J]. Online Social Networks and Media, 2021, 22: 100104.
[33] KOCHKINA E, HOSSAIN T, LOGAN R L, et al. Evaluating the Generalisability of Neural Rumour Verification Models[J]. Information Processing & Management, 2023, 60(1): 103116.
[34] BOIDIDOU C, PAPADOPOULOS S, ZAMPOGLOU M, et al. Detection and Visualization of Misleading Content on Twitter [J]. International Journal of Multimedia Information Retrieval, 2018, 7(1): 71-86.
[35] JIN Z, CAO J, GUO H, et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs[C]// 25th ACM International Conference on Multimedia. 2017: 795-816.
[36] NAKAMURA K, LEVY S, WANG W Y. Fakeddit: A New Multimodal Benchmark Dataset for Fine-Grained Fake News Detection[C]// 12th Language Resources and Evaluation Conference. 2020: 6149-6157.
[37] JINDAL S, SOOD R, SINGH R, et al. Newsbag: A Multimodal Benchmark Dataset for Fake News Detection[C]// The AAAI-20 Workshop on Artificial Intelligence Safety. 2020: 138-145.
[38] GAO G, FANG Y, HAN Y, et al. Construction of Multi-modal Social Media Dataset for Fake News Detection[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 144-154.
[39] ZHOU X, MULAY A, FERRARA E, et al. Recovery: A Multimodal Repository for Covid-19 News Credibility Research[C]// 29th ACM International Conference on Information & Knowledge Management. 2020: 3205-3212.
[40] QI P, BU Y, CAO J, et al. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms[C]// 37th AAAI Conference on Artificial Intelligence. 2023: 14444-14452.
[41] 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.
[42] BAI C, CHEN H, KUMAR S, et al. M2p2: Multimodal Persuasion Prediction Using Adaptive Fusion[J]. IEEE Transactions on Multimedia, 2021, 25: 942-952.
[43] HERSHEY S, CHAUDHURI S, ELLIS D P W, et al. CNN Architectures for Large-Scale Audio Classification[C]// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. 2017: 131-135.
[44] LUO W. Research and Implementation of Text Topic Classification Based on Text CNN[C]// 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications. 2022: 1152-1155.
[45] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-scale Image Recognition[C]// 3rd International Conference on Learning Representations. 2015: 1-14.

ZHU Feng, born in 1987, Ph.D, assistant professor, master supervisor. His main research interests include cyberspace security, Internet of Things security, and operating system security.

LI Peng, born in 1979, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 48573M). His main research interests include computer communication networks, clouding computing, and information security.

(责任编辑:何杨)