智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于视觉-语言关键线索挖掘的多模态假新闻检测模型

孟想, 王博岳, 高祎菡, 吴广超, 刘易昆, 吕松澄, 尹宝才

引用本文:
孟想, 王博岳, 高祎菡, 等. 基于视觉-语言关键线索挖掘的多模态假新闻检测模型[J]. 智能系统学报, 2026, 21(1): 109-119.
MENG Xiang, WANG Boyue, GAO Yihan, et al. Visual-language key clue discovery-based multimodal fake news detection model[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(1): 109-119.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505007

DOI: 10.11992/tis.202505007
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20251222.1406.003

基于视觉–语言关键线索挖掘的多模态假新闻检测模型

孟想, 王博岳, 高祎菡, 吴广超, 刘易昆, 吕松澄, 尹宝才
(北京工业大学 信息科学技术学院, 北京 100124)

摘　要: 为了解决现有模型在应对虚假新闻时往往忽视具有判别性的局部细节且难以准确捕捉图文间关键矛盾关系的问题, 本文提出一种基于视觉–语言关键线索挖掘的多模态假新闻检测模型 (visual-language key clue discovery-based multi-modal fake news detection model, VKC-MFND), 这也是一种具有决定性区域/位置感知的多尺度交互模型。该模型包含多尺度特征提取模块、关键特征信息提取模块以及多尺度特征对齐模块 3 个关键模块。具体而言, 多尺度特征提取模块用于提取文本与图像在不同尺度层面的特征, 包括句子级/描述级的全局特征和词级/目标框级的局部特征, 从而全面理解多模态数据, 增强信息的表达能力; 关键特征信息提取模块借助注意力机制, 在细尺度特征之间进行交互, 以发现具有判别性的关键内容, 并与全局语义进行对齐, 实现对图文间关键线索的有效融合; 多尺度特征对齐模块通过联合分类损失与对齐损失进行优化, 进一步挖掘全局语义特征, 实现语义空间的一致性。实验结果表明, 所提出的模型在 Weibo、Weibo-19 及 Pheme 等多个主流多模态假新闻数据集上均优于现有先进方法, 展现出更优的检测性能。消融实验进一步验证了各子模块在整体模型中的有效性和必要性。本研究的结论可为未来多模态假新闻检测模型的设计与优化提供指导。

关键词: 多模态虚假新闻检测; 多尺度特征交互; 关键线索发现; 细尺度表示; 跨模态注意力; 全局特征对齐; 记忆增强机制; 语义不一致检测
中图分类号: TP391.1　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)01−0109−11

中文引用格式: 孟想, 王博岳, 高祎菡, 等. 基于视觉–语言关键线索挖掘的多模态假新闻检测模型 [J]. 智能系统学报, 2026, 21(1): 109–119.
英文引用格式: MENG Xiang, WANG Boyue, GAO Yihan, et al. Visual-language key clue discovery-based multimodal fake news detection model[J]. CAAI transactions on intelligent systems, 2026, 21(1): 109–119.

Visual-language key clue discovery-based multimodal fake news detection model

MENG Xiang, WANG Boyue, GAO Yihan, WU Guangchao, LIU Yikun, LYU Songcheng, YIN Baocai
(School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Multimodal fake news detection aims to enhance the reliability of authenticity assessment by integrating diverse modalities such as text, images, videos, and audio. However, existing models often overlook discriminative local details and struggle to capture the critical inconsistencies between textual and visual content. To address these challenges, this study proposes a novel multimodal fake news detection model, termed the visual-language key clue discovery-based multimodal fake news detection model (VKC-MFND), which is designed to discover key visual-linguistic cues. The model comprises three main components: a multi-scale feature extraction module, a key feature information extraction module, and a multi-scale feature alignment module. Specifically, the multi-scale feature extraction module captures both global features (sentence-level or description-level) and local features (word-level or object box-level) from text and images, thereby enriching the diversity of information representation. The key feature information extraction module utilizes attention-based interactions among fine-grained features to uncover discriminative clues and aligns them with global semantic representations, facilitating the fusion of critical cross-modal information. Meanwhile, the multi-scale feature alignment module optimizes the model using both classification and alignment losses, enhancing semantic consistency in the shared feature space. Extensive experiments conducted on three benchmark multimodal fake news datasets—Weibo, Weibo-19, and Pheme—demonstrate that the proposed model significantly outperforms state-of-the-art approaches. Further ablation studies confirm the effectiveness and necessity of each component in the model.

Keywords: multimodal fake news detection; multi-scale feature interaction; key clue discovery; fine-grained representation; cross-modal attention; global feature alignment; memory-enhanced mechanism; semantic inconsistency detection

在数字时代, 社交媒体和互联网的普及极大促进了信息传播的速度和范围。虚假新闻不仅影响公共健康、政治、经济和社会安全, 还可能造成

收稿日期: 2025−05−16.　**网络出版日期**: 2025−12−22.
基金项目: 国家自然科学基金项目 (92370102).
通信作者: 尹宝才. E-mail: ybc@bjut.edu.cn.

深远的负面影响[1]。因此，开展高效、精准的虚假新闻检测显得尤为重要。

虚假新闻检测的目标是通过分析新闻内容、社交传播背景和外部知识库来评估新闻的真实性[2]。早期研究表明，新闻的语言风格、篡改的图像以及图文信息的不一致性等特征有助于区分真假新闻[3]。例如，虚假新闻常用耸人听闻的标题和情绪化表达来吸引关注，并配以篡改或误导性图片。文本和图像之间可能存在语义不匹配，图像和文本内容的不一致性是有效的检测线索[4]。通过机器学习、深度学习、自然语言处理和计算机视觉等技术，研究者可以高效筛查大量新闻[5]并识别潜在虚假信息，从而减少虚假新闻的传播，推动更可靠的信息环境建设。

现有的多模态检测模型[6-7]通常利用独立的视觉编码器和文本编码器分别提取图像和文本的全局特征[8]，随后在整体层面上进行特征融合[9]。然而，这种粗尺度的融合方式往往忽视了模态内部的局部细节信息，导致细尺度特征的利用率不足，难以准确捕捉决定新闻真实性的关键线索[10-11]。例如，在虚假新闻检测场景中，图像中的特定局部区域如修改过的痕迹[12]、重要的图标信息或文本中的关键短语如夸张性的表述[13]、与图片矛盾信息的词语往往对检测结果具有重要的决定性作用，但这些细尺度特征在现有模型的全局特征融合过程中容易被忽略[14-15]。

在典型的虚假新闻案例中，其图文信息通常蕴含着明显的矛盾线索[16]。传统的研究方法主要依赖于对整张图片和整段文本的全局特征进行分析[17]，这种方法虽然能够捕捉到宏观层面的信息，但往往忽视了具有判别性的局部细节[18]。这种图文细尺度信息之间的不一致性，恰恰为识别新闻真伪提供了重要依据。由此可见，加强对局部信息的关注和细尺度特征的分析[19]，能够显著提升模型识别虚假新闻的能力，特别是在处理此类包含局部矛盾信息的案例时[20]，而本文方法通过细尺度分析[21]直接聚焦于图文不一致的局部关键证据[22-23]，从而提高虚假新闻检测的准确性和可靠性。

针对以上问题，本文提出一种基于视觉−语言关键线索挖掘的多模态假新闻检测模型 (visual-language key clue discovery-based multi-modal fake news detection model, VKC-MFND)。该模型创新性地引入了细尺度特征挖掘机制，分别从视觉和文本模态中提取全局粗尺度特征及局部细尺度特征，统一在一个框架中[24]。与现有的多模态假新闻检测方法不同，首先，引入细粒度注意力机制，从图像和文本中分别提取局部关键区域和短语的特征，精准捕捉图文不一致性；其次，提出动态跨模态局部对齐策略，显式建模图像局部区域与文本短语的语义关联，直接识别矛盾关键证据；此外，通过多层级特征聚合模块将局部与全局特征动态结合，增强模型对判别性局部线索的敏感性；最后，以全局均方误差 (mean squared error, MSE) 对齐损失约束粗尺度语义一致，并结合可视化对齐结果，不仅提升了模型的可解释性，还显著增强了鲁棒性。本文的主要贡献如下：

1) 提出了 VKC-MFND 模型，该模型融合了全局和局部特征，分别从视觉和文本模态中提取不同尺度的信息。这种方法增强了模型对细尺度特征的识别能力，尤其是在虚假新闻检测任务中，能够更精准地捕捉关键细节。

2) 与现有模型依赖全局特征不同，VKC-MFND 强调了局部细尺度特征的提取，通过关注图像中的目标框特征和文本中的关键词，能够发现图文间不一致的局部信息，这对于识别虚假新闻至关重要。通过全局特征对齐模块，VKC-MFND 有效地缩小了不同模态之间的语义差异，从而增强了图像和文本模态之间的互补性，提高了模型的整体理解能力。

3) 通过在多个公开数据集 (Weibo、Weibo-19、Pheme) 上的实验验证了本文模型的优越性，给出了其在多模态虚假新闻检测任务中的卓越性能，相比于其他现有方法有明显的性能提升。

1　多模态假新闻检测的相关工作

1.1　多模态虚假新闻检测方法

近年来，多模态虚假新闻检测方法取得了显著进展，这些方法主要关注如何有效融合视觉与文本特征，以提升对假新闻的判别能力。为了应对不同模态之间信息的关联性不足，Wang 等[25]提出的事件不变神经网络 (event adversarial neural networks, EANN) 模型通过引入事件分类作为辅助任务，从而提升了模型对突发新闻的判别能力。在特征深度挖掘方面，Dhruv 等[26]提出的多模态变分自编码器 (multimodal variational autoencoder, MVAE) 模型采用变分自编码结构，通过对视觉与文本模态信息的重构，借助原始数据与重建数据之间的差异来识别虚假新闻。

此外，Singhal 等[27]提出的 SpotFake 模型融合了 VGG-19 提取的视觉特征与 BERT 提取的文本特征，并通过分类器进行真假判别；在此基础上，SpotFake+[28]模型进一步采用 XLNet，捕捉更丰富

的语义信息, 提升了检测性能。为加强跨模态特征的交互与一致性, Song 等[29] 提出的跨模态注意残差多通道网络 (cross-modal attention residual multi-channel network, CARMN) 模型将跨模态注意力残差网络与多通道卷积网络相结合, 既保留了各模态的独特信息, 又抑制了跨模态噪声。Zhou 等[30] 则提出的基于相似性感知多模态虚假新闻检测 (similarity-aware fake news detection, SAFE) 方法通过将视觉信息转化为文本, 并基于余弦相似度进行联合判别, 成功缓解了模态异构带来的信息鸿沟。

在进一步的多模态信息融合方面, Zheng 等[31] 提出的多模态特征聚合网络 (multi-modal feature aggregation network, MFAN) 模型通过将文本、视觉和社交属性等多源异构数据纳入基于图注意力网络的统一框架, 深入挖掘了不同数据源之间的互补关系。针对模态不一致所导致的误分类问题, Chen 等[32] 设计了跨模态歧义学习框架 (cross-modal ambiguity learning framework, CAFE) 模型, 该模型采用具有模糊性的多模态方法, 并引入单模态公共语义共享辅助任务, 从而提升了模型的稳健性。此外, Wang 等[33] 提出的 COOLANT 模型通过对比学习与跨模态一致性约束, 优化了视觉与语言特征的融合表达, 有效提升了虚假新闻的判别能力。

尽管上述方法在多模态特征提取、融合以及模态一致性等方面取得了显著进展, 但大多数现有研究侧重于全局语义或宏观一致性, 对图文之间细粒度矛盾证据的关注较少。针对这一问题, 本文提出了一种面向细粒度图文不一致的局部语义建模方法, 旨在充分挖掘和利用决定新闻真实性的细节证据, 从而进一步提升虚假新闻检测的准确性和可解释性。

1.2　特征提取

特征提取是多模态虚假新闻检测任务的关键步骤, 不同尺度的特征在新闻真伪判别中具有不同的重要性。现有的多模态假新闻检测方法通常利用不同的特征提取方法进行视觉与文本模态的信息表达。例如, EANN 模型使用 VGG-19 卷积神经网络提取图像的全局特征, 文本特征则采用文本卷积神经网络 (text convolutional neural network, Text-CNN) 模型进行提取; MVAE 模型通过 VGG-19 提取视觉特征, 使用双向长短期记忆网络 (bidirectional long short-term memory, Bi-LSTM) 提取文本特征, 并通过变分自编码器提取视觉与文本的潜在特征; Spotfake 及其增强版 Spotfake+则分别使用 VGG-19 网络与 XLNet 或 BERT 模型提取视觉与文本的全局语义特征; CARMN 模型通过 ResNet-152 或 Inception-v3 提取视觉特征, 通过双向门控循环单元 (bidirectional gated recurrent unit, Bi-GRU) 或卷积神经网络 (convolutional neural network, CNN) 提取文本特征, 并通过跨模态注意力残差网络提取模态之间的交互特征; SAFE 则使用预训练的 Image2Sentence 模型将视觉信息转化为文本特征; MFAN 使用 ResNet-101 提取图像特征、BERT 提取文本特征, 并通过图注意力网络整合文本、视觉与社交属性特征; CAFE 模型使用 VGG-16 提取视觉特征, Bi-GRU 提取文本特征, 通过跨模态辅助学习任务提取公共语义特征; COOLANT 使用 ResNet-50 和 BERT 模型分别提取视觉和文本特征, 利用跨模态对比学习进行特征提取与融合。

上述方法大多侧重于对全局信息的提取和融合, 忽略了视觉和文本模态内部的细粒度信息。这种粗尺度的融合方式使得模型难以精确捕捉虚假新闻中的关键局部线索。针对这一不足, 本文提出了一种多尺度特征提取方法。选择大型语言与视觉助手 (large language and vision assistant, LLaVA)[34] 模型生成图片描述作为视觉模态的粗尺度全局特征, 主要是由于 LLaVA 模型能够生成更为详细且语义丰富的描述, 有效弥合视觉与文本之间的语义鸿沟。

同时, 本文通过 VinVL 模型提取图像目标框特征, 获得视觉模态的细尺度特征表示, VinVL 模型在目标区域识别与描述方面表现突出, 有利于捕捉图像中的细节信息。此外, 使用 BERT[35] 提取文本的关键词特征, 精确挖掘文本模态中的关键短语信息。通过将注意力机制应用于这些细尺度特征, 进一步实现图文之间细尺度特征的交互融合。本文选择上述特征提取方法的主要原因在于, 它们能够有效结合全局语义与局部细节信息, 不仅显著提高了模型对新闻内容的整体理解能力, 也增强了对隐藏于局部特征中的细微矛盾线索的感知能力。

2　VKC-MFND 模型设计

本章提出的基于视觉–语言关键线索挖掘发现的多模态假新闻检测模型整体框架如图 1 所示。模型输入包括了新闻文本、新闻图片和图片描述, VKC-MFND 模型的整体框架由以下 3 个模块组成: 1) 多尺度特征提取模块。该模块从文本与图像这两种模态的数据中提取不同层次的特

征，涵盖局部细节与全局语义。通过提取多层次的特征，模型能够更全面地理解数据，避免仅依赖单一特征尺度所带来的局限性，从而增强特征的多样性。2）关键特征信息提取模块。在多模态数据中挖掘对虚假新闻检测最具决定性的局部信息，如文本中的敏感词或图像中的特定区域，通过融合关键信息降低无关信息的干扰，提升检测精度。3）多尺度特征对齐模块。确保不同模态的全局特征在语义空间中对齐，增强多模态数据的一致性，增强模型对多模态数据的整体理解。

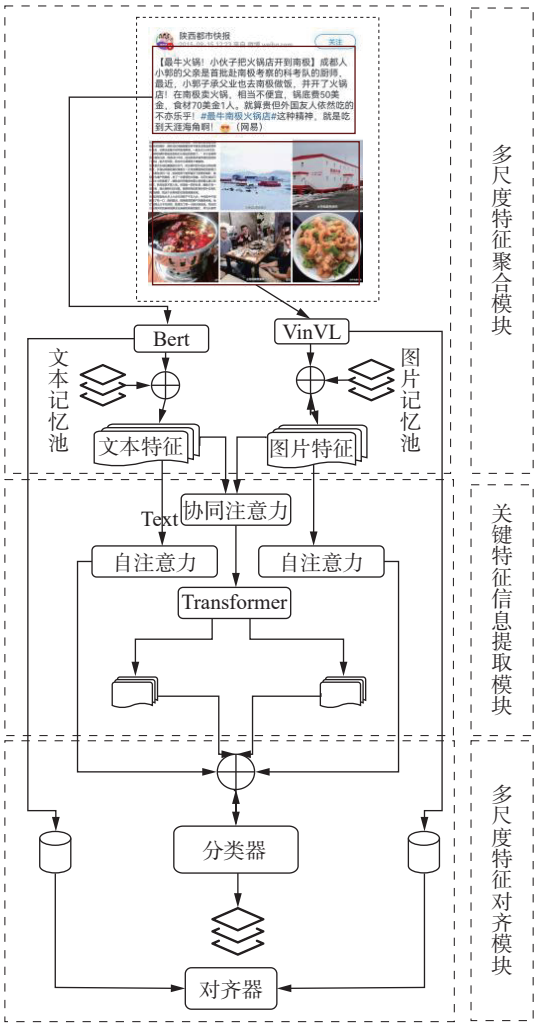


图 1　VKC-MFND 模型整体框架

Fig. 1　Overall frame diagram of VKC-MFND model

2.1　多尺度特征提取模块

给定的每个多模态新闻样本表示为

$$P = [T, V]$$

式中 T 和 V 分别为文本和视觉模态数据。为了有效学习这两种模态的特征，模型采用了预训练的 Transformer 编码器处理文本和图像信息。为了进一步提高特征表达的维度，模型在此基础上增加了一个全连接层将特征转换到适合后续处理的维度空间，从而使得多模态信息能够更全面地进行融合和分析。多尺度特征提取模块包括以下 4 个关键部分：1) 图像描述提取：通过图像描述生成技术，转化图像内容为文本信息；2) 细尺度局部特征提取：专注于捕捉图像和文本中的细节信息；3) 粗尺度全局特征提取：从全局角度提取信息以增强模型对整体情境的理解；4) 记忆信息增强机制：提升模型对长期依赖信息的记忆与处理能力。

1) 图像描述提取

在多模态虚假新闻检测任务中，理解与分析文本与视觉之间的关系非常关键。然而，不同模态之间的语义差异一直是一个挑战。为了缓解这一问题，本文通过为图像生成标题，将视觉内容转化为文本描述。这种方法可以使模型在文本层面处理图像信息，从而弥合视觉与语言模态之间的语义鸿沟。

Liu 等[34] 提出的 LLaVA 模型是一种端到端训练的大规模多模态模型，将视觉编码器与 LLaMA(large language model meta AI) 相结合，实现了对视觉和语言的通用理解。与其他传统的图像标题生成方法如全能基础 (one-for-all, OFA) 模型相比，LLaVA 模型能够提供更丰富且语义更深的句子描述，更好地捕捉图像的细节与上下文，因此，利用 LLaVA 生成图像的文本描述见下式：

$$T_{\text{cap}} = \text{LLaVA}(I) \tag{1}$$

T_{cap} 不仅作为新闻图像的文本描述，用于连接视觉与语言模态，还同时承载了图像的全局语义信息。通过 LLaVA 生成文本描述，模型能够将视觉内容转换为自然语言表达，使得图像信息可以在文本层面进行处理，在很大程度上增强跨模态信息的融合与对齐能力。

2) 细尺度局部特征提取

细尺度局部特征提取包含了文本细尺度特征提取和视觉细尺度特征提取两个主要部分。

①文本细尺度特征提取：采用 BERT 模型来提取词级特征，BERT 是一种基于 Transformer 模型的预训练语言表示方法，能够考虑上下文中词语的双向关联，从而更好地捕获文本的上下文信息。

在 VKC-MFND 模型中，每段文本内容 T 被表示为

$$T = \{w_1, w_2, \cdots, w_O\}$$

式中：w_i 为文本中第 i 个词，O 为词的总数。随后，文本 T 经过 BERT 模型的最后一层编码器进行处理，生成一个包含 n 个词的序列，并通过全连接层提取其细尺度文本表示，以进一步优化语义表达，文本局部特征提取如下式所示：

$$T_f = \boldsymbol{W}_t \times \text{Bert}(T) = \{t_1, t_2, \cdots, t_n\} \tag{2}$$

式中: $\boldsymbol{W}_t$为文本全连接线性层的权重矩阵, $t_i \in \mathbf{R}^{d_t}$为预训练 BERT 模型中第i个标记最后隐藏状态的输出, d_t为词嵌入的维度。

②视觉细尺度特征提取: 为了提取每篇文章的细尺度图像特征, 同样使用预训练的 Transformer 编码器来获取对象目标框特征表示。Zhang 等[19] 提出的 VinVL 是一种前沿的深度学习视觉语言模型, 通过利用大规模图文配对数据集, 深度理解和表示图像中的视觉信息。VinVL 能够在生成更详细准确的图像描述、回答关于图像内容的问题以及支持其他各种视觉语言任务中表现出色。因此, 使用 VinVL 预训练模型作为细尺度视觉特征提取的主干网络, 最后一层输出经过全连接层处理后, 被表示为视觉信息的局部特征。

$$V_f = \boldsymbol{W}_v \cdot \text{VinVL}(I) = \{v_1, v_2, \cdots, v_m\} \tag{3}$$

式中: $\boldsymbol{W}_v$为视觉全连接线性层的可学习权重矩阵, $\boldsymbol{v}_i \in \mathbf{R}^{d_v}$为 VinVL 中第 i 个图像目标框对象区域的向量表示, d_v为细尺度视觉特征的维度, m为对象区域的数量。

3) 粗尺度全局特征提取

粗尺度全局特征提取包含文本粗尺度特征提取和视觉粗尺度特征提取两个部分。

①文本粗尺度特征提取: 在预训练的 BERT 模型中, [CLS] 标记起着关键作用。对于所有输入序列, [CLS] 标记位置的输出代表整个句子的表示, 其维度是固定的。[CLS] 标记能够捕获整个句子的语义信息, 从而对上下文有全面的理解。这种全局语义表示对于诸如文本分类、句子相似性计算等自然语言处理任务具有重要作用。

在本章中, 利用 BERT 模型对每段新闻文本 T 进行隐藏层计算, 并从其最后一层隐藏状态中提取 [CLS] 标记对应的特征向量, 作为新闻文本的粗尺度表示, 用于捕捉全局语义特征, 以支持后续的多模态虚假新闻检测任务。具体表示见下式:

$$T_c = \boldsymbol{W}_t \cdot \text{BERT}(T) = t_{\text{cls}} \tag{4}$$

式中$t_{\text{cls}} \in \mathbf{R}^{d_t}$为预训练 BERT 模型中 [CLS] 标记最后隐藏状态的输出。

②视觉粗尺度特征提取: 类似于文本的全局信息提取, 视觉信息的最终粗尺度特征由通过对图像的标题进行 BERT 编码的 [CLS] 特征输出来表示。

$$V_c = \boldsymbol{W}_t \cdot \text{BERT}\left(T_{\text{cap}}\right) = t_{\text{cap}}^{\text{cls}} \tag{5}$$

4) 记忆信息增强机制

随着模型规模的不断扩大, 参数微调变得日益复杂, 而提示学习则成为了一种重要的策略。提示学习的目的是保持模型结构的固定, 并使得下游任务能够与预训练模型对齐, 这种方法以其简单高效、概念明确和显著的效果而受到广泛关注。

受到这一思想的启发, 本文提出了一种可学习的记忆信息增强机制, 使得模型能够灵活地适应特定任务和数据分布。通过为图像和文本的局部特征提供可学习的记忆信息, 模型可以被引导去关注图像和文本中一些关键的细节, 例如图像中的背景、物体的细节或文本中的情感表达等。这些细节对于判断信息的真实性至关重要, 细尺度的理解帮助模型在分析新闻内容时更为准确地捕捉到可能影响判断的微小特征。

具体来说, 设$M_t \in \mathbf{R}^{n \times d_m}$和$M_v \in \mathbf{R}^{m \times d_m}$分别为文本和视觉模态局部表示$T_f$和$V_f$的记忆信息, 其中$d_m$是可学习记忆向量的长度。通过记忆向量增强的文本和视觉细尺度特征表示如下式所示:

$$T_f^m = \text{concat}\left(T_f, M_t\right) \tag{6}$$

$$V_f^m = \text{concat}\left(V_f, M_v\right) \tag{7}$$

式中: $T_f^m \in \mathbf{R}^{n \times d_{tm}}$, $d_m = d_t + d_m$为记忆信息增强后的文本局部细尺度特征的维度; $V_f^m \in \mathbf{R}^{m \times d_{vm}}$, $d_{vm} = d_v + d_m$为记忆信息增强后的视觉局部细尺度特征的维度。

2.2　关键特征信息提取模块

在关键特征信息提取模块中, 利用注意力机制对每对图像和文本的目标对象区域和单词信息进行对齐和突出显示, 使提取的局部表示更具判别力。对于给定的局部序列输入$L = \{l_1, l_2, \cdots, l_j\}$, 第$i$个头的相关性矩阵计算为

$$M = (H_1 \oplus H_2 \oplus \cdots \oplus H_h)\boldsymbol{W}_o \tag{8}$$

$$H_i = \text{Attention}\left(Q_i^L, K_i^L, V_i^L\right) = \text{softmax}\left(\frac{Q_i^L K_i^L}{\sqrt{d_L}}\right) V_i^L \tag{9}$$

式中: l_j为第j个局部区域, 具体指目标框级别的视觉信息和词级别的文本内容; $\boldsymbol{W}_O \in \mathbf{R}^{d_L \times d_L}$为权重矩阵; Q_i^L、K_i^L、$V_i^L \in \mathbf{R}^{j \times d_L}$分别为第$i$个头 ($H$) 的查询($Q$)、键($K$)和值($V$), 其具体表示为

$$Q_i^L = LW_i^q, K_i^L = LW_i^k, V_i^L = LW_i^v \tag{10}$$

在模态内的局部信息交互分支中, 通过设置记忆增强的文本局部特征T_f^m作为 L, 可以获得自注意力文本表示T_f; 同样, 对V_f^m进行相同操作得到局部自注意力视觉特征V_f。

在模态间的局部信息融合分支中, 通过协同注意力机制生成多模态特征。具体地, 为了获得融合的视觉特征V_f^t, 将T_f^m替代 L 来计算查询矩阵, 将V_f^m替代 L 来计算键和值矩阵, 类似地, 可以以相同方式获得融合的文本特征T_f^v。每篇新闻文章的关键特征信息提取模块的输出表示为单模态特征和局部交互跨模态特征的拼接。

$$F = \text{concat}\left(T_f, T_f^v, V_f^t, V_f\right) \tag{11}$$

式中$F \in \mathbf{R}^{d_F}$,d_F为最终细尺度融合表示的维度。

2.3　全局特征对齐模块

不同模态的表示通常存在显著的语义差异,因此需要将不同模态嵌入转换到一个共同的空间,以弥合这些差异并对齐各模态的特征。具体来说,对于每篇新闻文章的文本−图像对,通过多层感知机 (multilayer perceptron MLP) 将其粗尺度文本特征T_c和粗尺度视觉特征 V_c转换为同一模态特征空间后得到T_c'和V_c',具体过程如下式所示:

$$T_c' = \boldsymbol{W}_{t2} \times \sigma(\boldsymbol{W}_{t1} \times T_c) \tag{12}$$

$$V_c' = \boldsymbol{W}_{v2} \times \sigma(\boldsymbol{W}_{v1} \times V_c) \tag{13}$$

式中:$\sigma(\cdot)$为 ReLU 激活函数,它使神经网络模型具有非线性特性;$\boldsymbol{W}_{t1}$、$\boldsymbol{W}_{t2}$和$\boldsymbol{W}_{v1}$、$\boldsymbol{W}_{v2}$分别为文本和视觉全局特征的第一层和第二层 MLP 的可学习权重矩阵参数。之所以选取第一层和第二层的权重参数,是因为两层 MLP 结构能够对输入特征进行由浅入深的非线性变换,第一层用于初步特征映射和信息抽取,第二层则进一步对特征进行压缩和高阶表达。这样的分层设计不仅有助于增强特征之间的复杂关系建模能力,提升多模态特征的对齐灵活性,还有利于模型梯度的有效传播,保证训练稳定性与性能表现。

在这种对齐空间中,即使文本和图像的原始特征具有不同的语义和分布特性,对齐后的特征可以减少这种差异对模型性能的影响,更好地利用两种模态的互补优势,从而提高对多模态内容的理解能力。为了实现全局模态对齐,使用 MSE 均方误差损失函数进行模态对齐任务,其计算公式为

$$\mathcal{L}_{\text{aligned}} = \frac{1}{N}\sum_{i=1}^{N}\left(T_c' - V_c'\right)^2 \tag{14}$$

式中 N 为新闻文章的总数。在此过程中,这两个全局特征被整合以捕捉和利用来自不同模态的互补信息,全局文本特征封装了文本数据的内容信息,而全局视觉特征通过生成文本描述封装了视觉的语义本质,它们的结合旨在增强模型在识别和理解多模态内容方面的性能。

式中$V_c \in \mathbf{R}^{d_v}$。通过这种方式,利用文本描述作为中介,有助于模型在统一的表示环境中理解图像与文本之间的语义联系,这对于分析复杂的图文关联至关重要。

2.4　总损失函数

通过上述一系列操作,得到了融合的细尺度特征F,、精炼的粗尺度文本对齐特征T_c'和图像对齐特征V_c',这些特征将用于最终的多模态新闻分类器。

在结合这些多尺度文本−图像对表示之前,对融合的细尺度特征 F 使用平均池化操作,该操作对所有段进行平均,将局部特征矩阵转换为固定长度的向量,从而将特征维度映射到低维空间,池化过程为

$$F' = \text{Average Pooling}(F)$$

新闻的最终表示为粗尺度全局特征与融合细尺度局部特征的结合。

$$F_{tv} = \text{concat}\left(T_c', V_c', F'\right) \tag{15}$$

分类损失$\mathcal{L}_{\text{cls}}$为由F_{tv}经过多层带有激活函数的多层感知机预测得到的结果$\hat{\boldsymbol{y}}$与真实标签$\boldsymbol{y}$的交叉熵损失,具体计算公式为

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\text{MLP}\left(F_{tv}\right)\right) \tag{16}$$

$$\mathcal{L}_{\text{cls}} = \boldsymbol{y}\log(\hat{\boldsymbol{y}}) + (1-\boldsymbol{y})\log(1-\hat{\boldsymbol{y}}) \tag{17}$$

最终的目标函数表示为对齐损失$\mathcal{L}_{\text{aligned}}$与分类损失$\mathcal{L}_{\text{cls}}$的加权和,加和过程为

$$\mathcal{L} = \alpha\mathcal{L}_{\text{aligned}} + \beta\mathcal{L}_{\text{cls}} \tag{18}$$

式中α和β为可调节的参数。整体算法的伪代码如下:

输入:多模态新闻数据 P=[T, V] 包括文本 T 和图像 V、新闻的标签 y 以及迭代次数 epoch;

输出:分类结果 $\hat{\boldsymbol{y}}$。

1) For i = 1 to epoch do;

2) 通过式 (1) 生成图像的文本描述T_{cap};

3) 分别通过式 (2) 和式 (3) 生成文本细尺度特征T_f与视觉细尺度特征V_f;

4) 分别通过式 (4) 和式 (5) 生成文本粗尺度特征T_c与视觉粗尺度特征V_c;

5) 通过式 (6) 和式 (7) 生成记忆机制增强的文本局部特征T_f^m和视觉局部特征V_f^m;

6) 通过式 (8) ~ (11) 生成融合的细尺度特征F;

7) 通过式 (12) ~ (14) 计算全局特征对其损失$\mathcal{L}_{\text{aligned}}$;

8) 通过式 (15) 结合多尺度特征生成最终的新闻表示F_{tv};

9) 通过式 (16) 预测新闻的分类结果$\hat{\boldsymbol{y}}$;

10) 通过式 (17) 计算新闻的分类交叉熵损失$\mathcal{L}_{\text{cls}}$;

11) 通过式 (18) 结合对齐损失和分类损失计算最终目标损失$\mathcal{L}$;

12) End for

3　实验设置与结果分析

3.1　实验设置与结果分析

1) 数据集

使用 Weibo[36]、Weibo-19[37] 和 Pheme[38] 这 3 个公开的多模态虚假新闻检测数据集。

2) 实验设置

3 个数据集按 7∶1∶2 的比例进行分割分别用于模型的训练、验证和测试。使用的 3 个数据集都通过所提出的模型进行了独立的训练和测试，共进行了 3 组实验。批次大小设置为 64 用于训练，50 用于测试，该模型使用 Adam 优化器，通过多次的参数调节，将 Weibo 数据集和 Weibo-19 数据集的学习率为设置为 0.000 5，Pheme 数据集的学习率设置为 0.000 9 来达到最好的测试效果。在所有实验中，模型在单个 NVIDIA RTX 3090 GPU 上训练 30 个 epoch，并且随机种子是固定的，以确保在参数相同的情况下每次训练和测试的结果能够保持一致。

表 1 中进一步给出了 VKC-MFND 模型在 3 个不同数据集下的训练批大小、学习率等实验详细配置。

表 1　VKC-MFND 模型在不同数据集的实验配置

Table 1　Experimental configuration of different data sets

参数名称	Weibo	Datasets Weibo-19	Pheme
批次样本大小	64	64	64
学习率	5×10^{-4}	5×10^{-4}	9×10^{-4}
训练批次	30	30	30
训练集构成	Train+valid	Train+valid	Train+valid
数据集样本数	6 844	1 467	2 018

3)VKC-MFND 模型参数描述

对于新闻图像的细尺度特征提取，使用 VinVL 为每张图片提取 30 个目标框区域 (m=30)。

对于新闻的图像全局描述，本文使用 LLaVA 为每张图片生成 5 个字幕，并随机选取一个字幕作为图片的多尺度特征。

本文使用 BERT 预训练模型处理新闻文本以及图像字幕，将文本细尺度局部序列的长度统一到 30 个标记 (n=30)，将文本全局特征以及图像描述表示为 [CLS] 标记，输出维度 d_t=768。

记忆增强机制中，可学习的记忆向量表示维度均设置为 d_m=50。在关键特征信息提取模块中多头注意机制的头数设置为 8(h=8)。在总损失函数中，对齐损失 $\mathcal{L}_{\text{aligned}}$ 与分类损失 $\mathcal{L}_{\text{cls}}$ 的权重 α 和 β 分别设置为 $\alpha=1$ 和 $\beta=0.5$。

3.2　实验方案

为全面评估所提出方法的有效性，论文的对比方法包含 EANN、MVAE、Spotfake、Spotfake+、CARMN、SAFE、MFAN、CAFÉ和 COOLANT 共 9 种具备代表性的先进多模态检测方法，所有方法均在当前环境下进行统一复现，确保结果的公平性和可比性。

3.3　对比实验与分析

在 Weibo、Weibo-19 和 Pheme 数据集上的具体实验结果分别如表 2~4。可以明显地发现，VKC-MFND 模型在多模态虚假新闻检测任务上能够实现检测性能的提升，其在所有 3 个数据集上均超越了复现的先进多模态虚假新闻检测方法。

表 2　在 Weibo 数据集上与其他先进模型对比结果

Table 2　Results compared with other advanced models on Weibo dataset

方法	准确率	虚假新闻			真实新闻		
		精确率	召回率	$F1$分数	精确率	召回率	$F1$分数
EANN	0.770	0.774	0.757	0.766	0.766	0.782	0.774
MVAE	0.715	0.765	0.616	0.682	0.682	0.812	0.742
SpotFake	0.767	0.734	0.834	0.780	0.810	0.701	0.752
SpotFake+	0.731	0.716	0.760	0.737	0.748	0.702	0.725
SAFE	0.812	0.884	0.715	0.790	0.763	**0.907**	0.829
CARMN	0.825	0.837	0.816	0.826	0.814	0.835	0.824
MFAN	0.844	0.863	0.816	0.839	0.828	0.872	0.849
CAFE	0.849	**0.898**	0.810	0.852	0.798	0.891	0.842
COOLANT	0.843	0.826	0.853	0.839	0.859	0.834	0.846
VKC-MFND	**0.859**	0.865	**0.849**	**0.858**	**0.863**	0.869	**0.861**

注: 加黑代表最优效果。

表 3　在 Weibo-19 数据集上与其他先进模型对比结果

Table 3　Results compared with other advanced models on Weibo-19 dataset

方法	准确率	虚假新闻			真实新闻		
		精确率	召回率	$F1$分数	精确率	召回率	$F1$分数
EANN	0.827	0.760	0.819	0.788	0.876	0.832	0.854

续表 3

方法	准确率	虚假新闻			真实新闻		
		精确率	召回率	*F*1分数	精确率	召回率	*F*1分数
MVAE	0.722	0.663	0.595	0.627	0.754	0.804	0.778
SpotFake	0.712	0.707	0.459	0.555	0.714	0.877	0.787
SpotFake+	0.722	0.667	0.586	0.624	0.751	0.810	0.780
SAFE	0.854	0.807	0.828	0.817	0.886	0.872	0.879
CARMN	0.864	0.845	0.817	0.831	0.877	0.897	0.887
MFAN	0.878	0.828	0.871	0.849	0.913	0.883	0.898
CAFE	0.898	0.864	**0.879**	0.872	0.921	0.911	0.916
COOLANT	0.881	0.853	0.846	0.850	0.899	0.904	0.902
VKC-MFND	**0.905**	**0.893**	0.862	**0.877**	**0.913**	**0.934**	**0.923**

注：加黑代表最优效果。

表 4　在 Pheme 数据集与其他先进模型对比结果

Table 4　Results compared with other advanced models on Pheme dataset

方法	准确率	虚假新闻			真实新闻		
		精确率	召回率	*F*1分数	精确率	召回率	*F*1分数
EANN	0.782	0.649	0.558	0.600	0.826	0.875	0.850
MVAE	0.792	0.685	0.540	0.604	0.824	0.897	0.859
SpotFake	0.769	0.636	0.496	0.557	0.808	0.882	0.844
SpotFake+	0.823	0.706	0.681	0.694	0.870	0.882	0.876
SAFE	0.813	0.753	0.540	0.629	0.829	0.926	0.875
CARMN	0.805	0.549	0.721	0.623	0.912	0.829	0.869
MFAN	0.875	0.756	**0.850**	0.800	**0.934**	0.886	0.909
CAFE	0.891	0.826	0.796	0.810	0.917	0.930	0.923
COOLANT	0.894	0.796	0.833	0.814	0.934	0.917	0.925
VKC-MFND	**0.904**	**0.839**	0.832	**0.836**	0.930	**0.934**	**0.932**

注：加黑代表最优效果。

绝大多数对比方法致力于减小不同模态信息差异带来的语义差异，SAFE 使用预先训练的 Image2Sentence 模型，将新闻内容中的视觉信息处理为句子。LLaVA 相比于 Image2Sentence 在图像描述生成中具有显著优势，它通过结合大规模视觉编码器和语言模型，能够生成更丰富、更具语义深度的描述，捕捉图像中的细节、上下文和情感信息，相比之下，Image2Sentence 通常依赖固定模板，描述机械且缺乏灵活性，难以捕捉复杂场景和细节。因此，VKC-MFND 模型利用 LLaVA 进行图片的全局特征提取能够在语义丰富性和细节捕捉方面表现更优，这也使得 VKC-MFND 模型的检测效果优于 SAFE。

此外，MFAN 将文本和社交图特征进行对齐来减小模态差异，CAFE 通过辅助跨模态学习任务将原始新闻嵌入转换到共享空间中，COOLANT 利用图文对比学习来实现多模态特征对齐，但这些方法都局限于对全局信息的提取和交互。在本章提出的 VKC-MFND 方法中，不仅利用图像描述和 MSE 对齐损失缩小了图文不同模态的语义差异，在此基础上整合了目标框级图像区域信息以及词级的文本细尺度特征并进行有效交互，使得 VKC-MFND 模型在检测任务上取得了相较于其他先进方法更卓越的性能。

3.4　消融实验与分析

为了验证 VKC-MFND 模型各个模块的有效性，本章分别对多尺度特征提取模块、关键特征信息提取模块以及全局特征对齐模型这 3 个模块进行消融实验。为了确保实验的严谨性和可比性，设计了 5 种不同的消融方案，这 5 种消融方案配置如下：

VKC-MFND w/o L：移除多尺度特征提取模块中的细尺度特征提取部分，只使用图像描述以及文本句子级粗尺度特征作为新闻表示。

VKC-MFND w/o G：移除多尺度特征提取模块中的图文粗尺度特征以及全局特征对齐模块，

只使用图像目标框特征以及文本词嵌入来进行后续融合及分类。

VKC-MFND w/o M: 去除了多尺度特征提取模块中的记忆信息增强机制, 并且细尺度特征仅由编码器提取来表示。

VKC-MFND w/o C: 去除了关键特征信息提取模块, 采用未经融合的单模态特征进行特征拼接直接用于检测任务。

VKC-MFND w/o A: 去除全局特征对齐模块, 粗尺度特征仅用来学习分类损失。

对 VKC-MFND 模型在 Weibo、Weibo-19 和 Pheme 数据集上的具体消融实验结果如表 5 所示。实验结果表明, 移除任何一个模块都会导致模型性能的显著下降。

表 5　VKC-MFND 模型消融实验结果

Table 5　Ablation experiments of VKC-MFND model

数据集	方法	准确率	*F*1分数	
			虚假新闻	真实新闻
Weibo	VKC-MFND	**0.859**	**0.858**	0.783
	w/o L	0.787	0.790	0.866
	w/o G	0.834	0.837	0.829
	w/o M	0.847	0.845	0.851
	w/o C	0.831	0.832	0.829
	w/o A	0.844	0.842	**0.845**
Weibo-19	VKC-MFND	**0.905**	**0.911**	**0.941**
	w/o L	0.861	0.835	0.880
	w/o G	0.868	0.830	0.892
	w/o M	0.885	0.855	0.904
	w/o C	0.864	0.820	0.891
	w/o A	0.888	0.858	0.908
Pheme	VKC-MFND	**0.904**	**0.833**	**0.932**
	w/o L	0.813	0.702	0.864
	w/o G	0.865	0.752	0.907
	w/o M	0.875	0.762	0.915
	w/o C	0.834	0.667	0.889
	w/o A	0.870	0.773	0.910

注: 加黑代表最优效果。

多尺度特征提取模块的消融实验验证了细尺度和粗尺度特征协同的重要作用。其中, 移除细尺度特征提取部分 (w/o L) 对模型性能在 3 个数据集上的影响都最为显著, 在 Weibo、Weibo-19 和 Pheme 数据集上的准确率分别下降至 0.787、0.861 和 0.813。这表明细尺度特征在捕捉局部细节和关键信息方面具有重要作用。同时, 移除粗尺度特征 (w/o G) 以及全局特征对齐模块 (w/o A) 导致模型准确率的下降说明图像描述和全局特征对齐能够有效减少跨模态语义差异, 提升特征的一致性和互补性, 在跨模态语义对齐和全局信息整合中起到了关键作用。移除记忆信息增强机制后, 模型性能有所下降, 表明记忆信息增强机制能够有效提升细尺度特征的表达能力和任务适应性, 移除该机制会导致模型对细尺度信息的表达能力减弱。此外, 移除关键特征信息提取模块 (w/o C) 的模型准确率明显降低表明关键特征信息提取模块在捕捉跨模态关键信息方面具有不可替代的作用, 通过跨模态注意力机制能够有效融合文本和图像的局部关联信息, 从而提升模型对细微线索的捕捉能力。

总体而言, VKC-MFND 模型的各个模块之间具有显著的协同作用。多尺度特征提取模块通过整合细尺度和粗尺度特征, 能够全面捕捉新闻内容的局部细节和全局语义; 关键特征信息提取模块通过跨模态注意力机制, 有效融合了文本和图像的局部关联信息; 全局特征对齐模块通过最小化跨模态特征差异, 提升了特征的一致性和互补性, 这种模块间的协同作用使得 VKC-MFND 模型在多模态假新闻检测任务中表现出色。

4　结论

1)本文提出了基于视觉–语言关键线索发现的多模态假新闻检测模型 VKC-MFND, 旨在通过多尺度信息交互来提升检测效果。该模型的核心创新在于融合了全局与局部特征, 分别从视觉和文本模态中提取多层次的信息。具体来说, 全局粗尺度特征包括通过 LLaVA 生成的图像描述和文本的句子级表征, 而局部细尺度特征则涉及 VinVL 提取的图像目标框特征和 BERT 提取的文本词汇特征。

2)在此基础上, 模型采用注意力机制来融合这些细尺度的多模态信息, 并通过全局特征对齐模块来减少不同模态间的语义差异, 这种方法不仅整合了从宏观到微观的多层次特征, 还显著增强了模型对关键细节的识别能力。为验证 VKC-MFND 模型的有效性, 本文在 3 个公开的多模态虚假新闻检测数据集上进行了全面实验, 对比实验结果充分验证了模型在虚假新闻检测任务中的优越性能, 消融实验结果进一步证明了所提出的 3 个模块的有效性。

参考文献:

[1]　VOSOUGHI S, ROY D, ARAL S. The spread of true and

false news online[J]. Science, 2018, 359(6380): 1146−1151.

[2] CAO Juan, QI Peng, SHENG Qiang, et al. Exploring the role of visual content in fake news detection[EB/OL]. (2020−03−11)[2025−04−20]. https://arxiv.org/abs/2003.05096.

[3] ZHANG Xichen, GHORBANI A A. An overview of online fake news: Characterization, detection, and discussion[J]. Information processing & management, 2020, 57(2): 102025.

[4] ZHANG Zhenyu, ZHANG Lei, YANG Dingqi, et al. KRAN: knowledge refining attention network for recommendation[J]. ACM transactions on knowledge discovery from data, 2022, 16(2): 1−20.

[5] NAN Qiong, CAO Juan, ZHU Yongchun, et al. MDFEND: multi-domain fake news detection[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Virtual Event: ACM, 2021.

[6] RADFORD A, KIM J K, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. Online: ICML, 2021.

[7] WU Yang, ZHAN Pengwei, ZHANG Yunjian, et al. Multimodal fusion with co-attention networks for fake news detection[C]//Findings of the Association for Computational Linguistics, Stroudsburg: USAACL, 2021.

[8] 王安然, 袁得嵛, 潘语泉, 等. 基于超图双重注意力机制的多模态谣言检测模型[J]. 计算机科学与探索, 2025, 19(11): 3033−3045.
WANG Anran, YUAN Deyu, PAN Yuquan, et al. Multimodal rumor detection model based on hypergraph dual attention mechanism[J]. Journal of frontiers of computer science and technology, 2025, 19(11): 3033−3045.

[9] QIAN Shengsheng, WANG Jinguang, HU Jun, et al. Hierarchical multi-modal contextual attention network for fake news detection[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Even: ACM, 2021.

[10] 赵梦凡, 张钰涛, 赵铤钊. 社交媒体假新闻检测: 基本理论、方法及研究方向[J]. 软件导刊, 2024, 23(9): 31−40.
ZHAO Mengfan, ZHANG Yutao, ZHAO Tingzhao, et al. Social media fake news detection: basic theories, methods, and research directions[J]. Software guide, 23(9): 31–40.

[11] 朱枫, 张廷辉, 李鹏, 等. 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, et al. Multimodal adaptive fusion-based short video fake news detection[J]. Computer science, 51(11): 39−46.

[12] QI Peng, CAO Juan, LI Xirong, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues[C]//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu: ACM, 2021.

[13] CHEN Yixuan, LI Dongsheng, ZHANG Peng, et al. Cross-modal ambiguity learning for multimodal fake news detection[C]//Proceedings of the ACM Web Conference 2022. Virtual Event: ACM, 2022.

[14] YING Qichao, HU Xiaoxiao, ZHOU Yangming, et al. Bootstrapping multi-view representations for fake news detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023.

[15] ZHENG Jiaqi, ZHANG Xi, GUO Sanchuan, et al. MFAN: multi-modal feature-enhanced attention networks for rumor detection[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Vienna: International Joint Conferences on Artificial Intelligence Organization, 2022.

[16] 彭广川, 吴飞, 韩璐, 等. 基于跨模态交互与特征融合网络的假新闻检测方法[J]. 计算机科学, 2024, 51(11): 23−29.
PENG Guangchuan, WU Fei, HAN Lu, et al. A fake news detection method based on cross-modal interaction and feature fusion network[J]. Computer science, 51(11), 23–29.

[17] 杨书新, 丁祺伟. 基于局部和全局特征聚合的虚假新闻检测方法[J]. 计算机工程与应用, 2025, 61(9): 139−147.
YANG Shuxin, DING Qiwei. False news detection method based on local and global feature aggregation[J]. Computer engineering and applications, 2025, 61(9): 139−147.

[18] LIU Xuannan, LI Peipei, HUANG Huaibo, et al. FKA-owl: advancing multimodal fake news detection through knowledge-augmented LVLMs[C]//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne: ACM, 2024.

[19] ZHANG Pengchuan, LI Xiujun, HU Xiaowei, et al. Vinvl: revisiting visual representations in vision-language models[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021.

[20] 袁玥, 刘永彬, 欧阳纯萍, 等. 基于一对多关系的多模态虚假新闻检测[J]. 中文信息学报, 2023, 37(9): 131−139.
YUAN Yue, LIU Yongbin, OUYANG Chunping, et al. Multimodal fake news detection based on one-to-many relationships[J]. Journal of Chinese information processing, 2023, 37(9): 131−139.

[21] KIM W, SON B, KIM I. Vilt: vision-and-language transformer without convolution or region supervision[C]//In-

ternational Conference on Machine Learning. Online: PMLR, 2021.

[22] WANG Peng, YANG An, MEN Rui, et al. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022.

[23] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[EB/OL]. (2024−03−04)[2025−04−20]. https://arxiv.org/abs/2303.08774.

[24] 周昊玮, 刘勇, 玄萍. 基于预训练和多模态融合的假新闻检测[J]. 计算机工程, 2024, 50(1): 289−295.
ZHOU Haowei, LIU Yong, XUAN Ping. Fake news detection based on pretraining and multimodal fusion[J]. Computer engineering, 2024, 50(1): 289−295.

[25] WANG Yaqing, MA Fenglong, JIN Zhiwei, 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. London: ACM, 2018.

[26] KHATTAR D, GOUD J S, GUPTA M, et al. MVAE: multimodal variational autoencoder for fake news detection[C]//The World Wide Web Conference. San Francisco: ACM, 2019.

[27] 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. Singapore: IEEE, 2019.

[28] SINGHAL S, KABRA A, SHARMA M, et al. SpotFake+: a multimodal framework for fake news detection via transfer learning (student abstract)[C]//34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020.

[29] LI Jun, BIN Yi, ZOU Jie, et al. Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection[C]//Proceedings of the 5th ACM International Conference on Multimedia in Asia. New York: Association for Computing Machinery, 2023.

[30] ZHOU Xinyi, WU Jindi, ZAFARANI R, et al. SAFE: similarity-aware multi-modal fake news detection[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online: ACL, 2020.

[31] ZHENG Jiaqi, ZHANG Xi, GUO Sanchuan, et al. MFAN: multi-modal feature-enhanced attention networks for rumor detection[C]//International Joint Conference on Artificial Intelligence. Vienna: IJCAI, 2022.

[32] CHEN Yanchun, ZHANG Yuan, ZHANG Mengnan, et al. Consumption of coffee and tea with all-cause and cause-specific mortality: a prospective cohort study[J]. BMC medicine, 2022, 20(1): 449.

[33] WANG Longzheng, ZHANG Chuang, XU Hongbo, et al. Cross-modal contrastive learning for multimodal fake news detection[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023.

[34] LI Bo, ZHANG Yuanhan, GUO Dong, et al. LLaVA-onevision: easy visual task transfer[EB/OL]. (2024−10−26)[2025−01−20]. https://arxiv.org/abs/2408.03326.

[35] LIU Yihan, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[EB/OL]. (2019−07−26)[2025−04−20]. https://arxiv.org/abs/1907.11692.

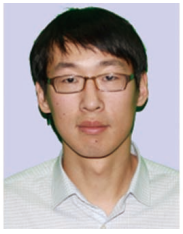
[36] JIN Zhiwei, CAO Juan, GUO Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View: ACM, 2017.

[37] SONG Changhe, YANG Cheng, CHEN Huimin, et al. CED: credible early detection of social media rumors[J]. IEEE transactions on knowledge and data engineering, 2021, 33(8): 3035−3047.

[38] ZUBIAGA A, LIAKATA M, PROCTER R. Exploiting context for rumour detection in social media[C]//The 9th International Conference on Social Informatics. Oxford: Springer International Publishing, 2017.

作者简介:

孟想, 主要研究方向为多模态真假新闻检测。E-mail: mx2005@emails.bjut.edu.cn。



王博岳, 教授, 主要研究方向为跨媒体数据分析、图结构学习。主持国家自然科学基金项目等 10 余项, 发表学术论文 10 余篇。E-mail: wby@bjut.edu.cn。

尹宝才, 教授, 主要研究方向为多媒体技术、跨媒体智能、视频编码。主持国家青年科学基金项目 A 类、国家自然科学基金重大项目课题等多项, 发表学术论文 100 余篇。 E-mail: ybc@bjut.edu.cn。