计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010110] [CSTR: 32024.14.csa.010110]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

细粒度对齐与优势模态增强的多模态假新闻检测①

白书铭, 曹霑懋, 刘　聪, 李俊斌

(华南师范大学 计算机学院, 广州 510631)
通信作者: 曹霑懋, E-mail: caozhanmao@scnu.edu.cn

摘　要: 现有的多模态假新闻检测方法仍存在以下不足: 在跨模态语义对齐过程中, 仅对全局特征进行对齐, 难以建立图像局部区域与对应文本片段之间的细粒度语义对齐; 在模态融合阶段通常简单采用等权融合策略, 未能充分发挥信息更丰富的优势模态的作用, 从而限制了模型性能. 鉴于此, 提出了一种细粒度对齐与优势模态增强的多模态假新闻检测模型. 所提模型中的细粒度对齐模块利用 FG-CLIP 模型的细粒度对齐能力, 引导新闻的图像与文本的深层语义特征建立精确对应, 有效抑制无关区域的干扰. 提出用置信度来判定优势模态, 该置信度根据单模态特征与其类别原型之间的距离计算得出. 同时, 引入原型交叉熵损失以增强优势模态的表征能力, 使其在融合过程中发挥主导作用. 在 Weibo 和 GossipCop 数据集上的实验结果表明, 该模型在多数评估指标上优于基线模型, 验证了其在虚假新闻检测任务中的有效性与鲁棒性.

关键词: 多模态假新闻检测; 多模态学习; 细粒度语义对齐; 原型网络; 优势模态增强

引用格式: 白书铭,曹霑懋,刘聪,李俊斌.细粒度对齐与优势模态增强的多模态假新闻检测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10110.html

Multimodal Fake News Detection via Fine-grained Alignment and Dominant Modality Enhancement

BAI Shu-Ming, CAO Zhan-Mao, LIU Cong, LI Jun-Bin

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Existing multimodal fake news detection methods still suffer from the following limitations. During cross-modal semantic alignment, only global features are typically aligned, failing to establish fine-grained semantic correspondences between local image regions and their relevant text fragments. In the modality fusion stage, an equal-weight combination strategy is usually adopted, which prevents the more informative dominant modality from being fully utilized, thereby limiting model performance. To address these issues, this study proposes a multimodal fake news detection model integrating fine-grained alignment and dominant modality enhancement. The fine-grained alignment module leverages the capability of the FG-CLIP model to guide precise correspondence between deep semantic features of news images and texts, effectively suppressing interference from irrelevant regions. Moreover, the dominant modality is determined by a confidence score, which is computed based on the distance between single-modality features and their corresponding class prototypes. A prototype cross-entropy loss is introduced to enhance the representational capacity of the dominant modality, enabling it to play a leading role in the fusion process. Experimental results on the Weibo and GossipCop datasets demonstrate that the proposed model outperforms baseline methods on most evaluation metrics, verifying its effectiveness and robustness in fake news detection tasks.

Key words: multimodal fake news detection; multimodal learning; fine-grained semantic alignment; prototype network; dominant modality enhancement

① 收稿时间: 2025-09-10; 修改时间: 2025-10-10; 采用时间: 2025-10-29; csa 在线出版时间: 2026-01-19

随着社交媒体的快速发展, 公众获取与发布信息变得日益便捷. 然而, 互联网信息的低成本、易获取与快速传播等特点也加剧了虚假新闻的泛滥, 给社会舆论环境造成了负面影响[1]. 因此, 设计一种能够高效、自动的虚假新闻检测方法具有重要的现实意义. 多模态虚假新闻检测方法同时利用新闻的图像、文本信息进行检测, 能够全面捕获不同模态之间的信息关联, 并通过多模态融合实现各模态的优势互补, 从而提高模型对新闻的全面理解能力与分类能力, 其性能优于基于单模态的检测方法[2].

现有多模态虚假新闻检测方法通常使用预训练模型分别提取文本特征和图像特征, 并将二者映射至共享语义空间以实现跨模态语义对齐与融合, 然后使用融合特征进行新闻分类[3]. 其中, 跨模态语义对齐是关键环节, 其核心目标是建立图像内容与文本之间的语义关联, 从而实现图像语义互补与协同理解[4]. 然而, 已有方法采用的是粗粒度的全局对齐策略, 难以在复杂新闻场景中实现图文之间关键区域的细粒度对齐, 从而限制了模型的检测性能.

此外, 有研究指出, 多模态虚假新闻检测模型可能会面临跨模态歧义问题, 即不同模态间给出相互矛盾的预测结果[5], 若用同一个目标函数训练融合特征, 易导致性能下降. 本文针对跨模态歧义问题进一步分析, 发现新闻的文本和图像并不都包含有效的新闻内容. 一些新闻的图片仅具装饰性, 无法提供有价值的语义信息, 新闻文本才包含主要新闻内容; 而在另一些新闻中, 文本内容仅具有博人眼球的作用, 只有现场照片才能反映事件真相. 由此可见, 在一些新闻中, 包含新闻主要内容的优势模态在虚假新闻检测过程中起到了主导作用, 而弱势模态仅仅提供了少量的语义互补信息, 甚至可能带来误导性干扰[6]. 因此, 赋予模型动态识别优势模态的能力, 并使得优势模型能够主导虚假新闻检测的过程可能会进一步提升模型对虚假新闻检测准确率.

针对以上问题, 本文提出了细粒度对齐与优势模态增强的多模态假新闻检测模型 FADM (fine-grained alignment and dominant modality enhancement multimodal fake news detection model). 该模型包含两个关键模块: 细粒度语义对齐模块和优势模态识别与增强模块. 细粒度语义对齐模块引入 FG-CLIP 模型[7], 这是一个经过大规模跨模态语义对齐任务训练的多模态大模型, 能够对图像局部区域进行精准感知, 并与文本中特定语句建立对应关系, 从而实现了细粒度的语义对齐. 通过对 FG-CLIP 进行迁移学习, 能够将这种能力迁移至本文提出的 FADM 模型中. 通过该模块的处理, FADM 模型能够精确捕捉图文新闻中局部语义的一致性和不一致性, 从而实现图文模态相互增强, 有效提高了对虚假新闻检测的准确率. 另外, 本文提出增强优势模态的多模态虚假新闻的方法. 该方法受到原型网络[8]的启发, 其核心思想源自原型网络的非参数分类机制. 首先根据每个模态特征与其所属类别原型之间的距离计算置信度, 将置信度更高的模态判定为优势模态, 另一个则称之为劣势模态. 然后, 通过最小化优势模态的原型交叉熵损失值的方式, 将优势模态的特征向量向其对应类别的原型方向进行聚类, 从而增强了优势模态的表征能力, 使其能够主导检测过程, 同时抑制了劣势模态带来的干扰.

这两个模块均将多模态数据映射到一个共享语义空间之中. 在该空间中, 细粒度语义对齐模块通过拉近配对的图像和文本特征向量之间的距离, 实现跨模态语义对齐; 而优势模态识别与增强模块则是将优势模态的特征向量向其类别原型方向进行聚类, 从而产生更易区分的决策边界, 最终增强其表征学习能力. 两个模块在共享语义空间中协同优化, 实现了语义对齐与模态增强的统一, 从而显著提升多模态虚假新闻检测的准确性与鲁棒性.

本文主要贡献如下.

(1) 提出一种能够实现跨模态细粒度对齐的多模态虚假新闻检测模型. 该模型引入多模态大模型 FG-CLIP, 引导新闻文本与图像进行细粒度语义对齐.

(2) 设计一种基于原型的优势模态评估与增强方法. 该方法基于原型网络判断出包含更多信息的优势模态, 并通过最小化优势模态的原型交叉熵损失, 增强其表征学习能力.

(3) 在两个主流基准数据集 Weibo 与 GossipCop 上进行实验, 结果表明, 本文提出的方法在多项指标上优于现有方法, 验证其有效性.

1 相关工作

1.1 多模态虚假新闻检测

多模态虚假新闻检测任务的主流方法是使用预训练的模型分别提取新闻的图像特征与文本特征, 然后

进行多模态特征融合, 使用融合特征进行新闻分类. Khattar 等人[9]提出将文本特征与图像特征映射至共享语义空间, 通过变分自编码器捕捉模态间的潜在相关性. Zhang 等人[10]使用预训练的 BERT[11]和 VGG-19 分别提取文本和图像的深层次语义信息, 通过对抗学习对齐源域和目标域的特征空间, 缓解领域偏移问题. BMR[12]模型使用了一种多视图自举机制, 能够充分挖掘和融合来自不同模态和不同语义视角的信息, 以增强模型对虚假新闻的辨别能力. Wu 等人[13]设计了多层协同注意力机制增强图像的空间域、频域特征与文本特征之间的交互, 从而充分发挥不同模态特征之间的语义互补优势. 然而, 依赖特征拼接或注意力融合的方法难以对模态间语义关系进行充分建模, 为了解决这一问题, Jing 等人[14]提出了一种渐进式融合网络, 从低层到高层逐步融合模态特征, 使模型能够捕获模态间的多层次交互. 而 Shen 等人[15]提出的 MCOT 模型将对比学习[16]与最优传输理论[17]相结合, 从特征分布层面建模图像与文本模态间的语义关系, 以提升跨模态表示对齐的精度.

近年来, 随着多模态大模型的发展, CLIP[16]等多模态学习模型为跨模态语义对齐提供了新思路. CLIP 模型在大规模的图像-文本对上通过对比学习进行训练, 能够更好地建立图像与文本之间的关联. 因此, 有研究通过引入预训练的 CLIP 模型帮助图文新闻实现跨模态语义对齐. Zhou 等人[18]分别用 CLIP 模型的图像编码器和文本编码器提取新闻的文本和图像特征, 并进行拼接, 生成包含语义关联信息的多模态融合特征. 然而, 现有的多模态学习模型大多停留在粗粒度语义对齐层面, 仅能捕捉全局语义信息, 难以有效建模新闻图文之间的细粒度关联. 为解决这一问题, 研究者提出了 FG-CLIP 模型[7], 它在 CLIP 的基础上通过大规模数据训练、区域级标注构建以及细粒度负样本引入, 显著提升了图文语义的细粒度对齐能力.

1.2 基于原型的优势模态评估与增强方法

新闻的图像与文本并不总是包含有效信息, 包含少量语义信息的弱势模态会引入额外的噪声, 干扰模型对新闻类别的判断, 此时优势模态包含主要的新闻内容, 应该让其主导新闻检测. 然而, 如何识别优势模态与劣势模态, 并增强优势模态的表征学习能力, 使其主导新闻的分类面临着挑战.

原型网络[19]最初用于少样本和零样本学习, 其核心思想是为每个类别构建原型表征, 并通过计算样本与各类别原型的距离进行分类. 由于能够有效表示类别的共性特征, 该方法近年来被广泛应用于多种任务. 针对多模态学习的模态不平衡问题, Fan 等人[20]提出利用原型网络评估每种模态的表现, 并通过熵正则化增强弱势模态的学习速率, 避免优势模态过早收敛. 这些研究表明, 原型网络能够有效评估模态表现.

为了解决多模态虚假新闻检测任务面临的无法有效进行细粒度对齐, 以及无法充分利用优势模态中的信息的问题, 本文提出了细粒度语义对齐和优势模态主导的多模态虚假新闻检测方法, 通过引入 FG-CLIP 模型[7], 引导新闻图像与文本的细粒度语义对齐. 同时, 使用原型网络动态评估新闻中图像与文本模态贡献, 增强优势模态的表征学习能力, 减少弱模态的干扰, 提升整体检测性能.

2 细粒度对齐与优势模态主导模型

本文将虚假新闻检测任务视为二分类问题, 目标是使用本文提出的模型自动判断给定图文新闻的真伪. 模型整体结构如图 1 所示. 该模型主要包含 4 个关键模块: 多模态特征提取模块 (包含图像特征提取器与文本特征提取器)、细粒度语义对齐模块、多模态特征聚合模块和优势模态主导模块. 下面将对本文模型进行详细介绍.

2.1 多模态特征提取模块

FADM 模型首先使用特征提取器分别对新闻文本和图像进行特征提取.

在文本特征提取器中, 首先对原始输入文本x_{T}进行分词, 得到的单词序列: $\{x_1, x_2, \cdots, x_m\}$. 然后, 在单词序列的开头添加[CLS]向量, 用于分类任务中聚合全局语义; 结尾加上[SEP], 用于分隔语句. 最终映射为 *token* 序列: $\left\{t_0^{[\mathrm{CLS}]}, t_1, t_2, \cdots, t_m, t_{m+1}^{[\mathrm{SEP}]}\right\}$. 其中, m 表示新闻文本中的词数, t_i表示第 i 个词对应的 *token* 序列. 该过程可表示为公式:

$$token = Tokenizer(x_{\mathrm{T}}) \tag{1}$$

然后, 将 *token* 序列输入 BERT 模型[11]进行文本特征提取. BERT 由 12 个堆叠的 Transformer[21]编码器组成, 每一个编码层输出一个特征向量h, $h^{(l)}$表示第 l 层的特征向量:

$$h^{(l)} = BERT(tokens)[l] \tag{2}$$

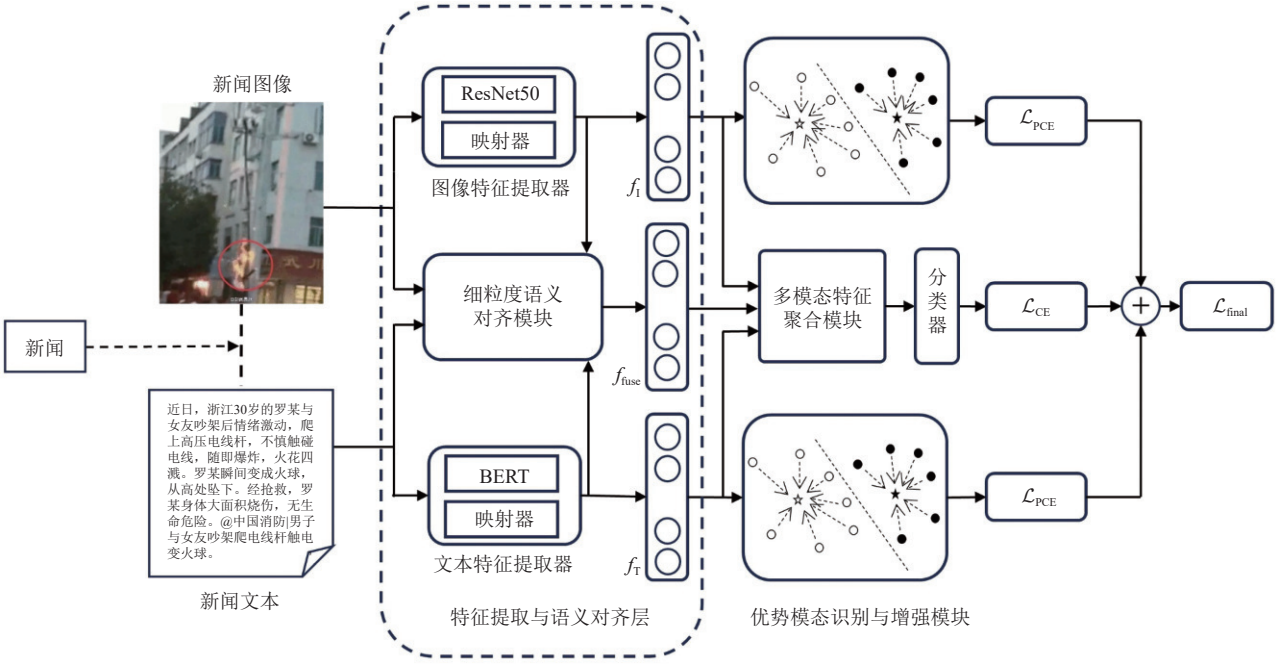


图 1　细粒度对齐与优势模态主导的模型

为了充分利用多层次的文本语义信息, 模型提取每一层的[CLS]向量$h_{[CLS]}^{(l)}$, 并通过层级注意力机制对其加权融合. 首先, 为每一层分配一个可学习的权重参数$w^{(l)}$, 并使用 Softmax 函数计算层注意力权重α_l:

$$\alpha_l = \frac{\exp\left(w^{(l)}\right)}{\sum_{j=0}^{L-1}\exp\left(w^{(j)}\right)} \tag{3}$$

最终的文本特征向量由所有层的[CLS]向量加权求和得到:

$$h_t = \sum_{l=0}^{L-1}\alpha_l \cdot h_{[CLS]}^{(l)} \tag{4}$$

在获得文本特征向量h_t后, 本文引入映射器P_{Text}, 将其映射为低维向量f_{T}, 以去除文本中的冗余信息, 并保留关键语义特征. 该映射器P_{Text}由多个全连接层组成, 通过逐层降维的方式压缩输入特征, 得到处理后的文本特征:

$$f_{\text{T}} = P_{\text{Text}}(h_{\text{T}}) \tag{5}$$

在新闻现场拍摄的图像中通常蕴含丰富的视觉信息, 如特定任务、物体与场景等, 这些信息可以与文本形成互补, 共同判断新闻的真实性.

首先, 对原始输入的新闻图像x_{I}进行预处理, 将其缩放至 224×224 的固定像素大小, 然后输入 ResNet50 模型[22], 提取深层图像特征:

$$h_{\text{I}} = ResNet(x_{\text{I}}) \tag{6}$$

为了进一步突出与新闻语义相关的关键信息, 并减少无关图像背景信息的干扰, 在 ResNet50 模型之后加入了轻量级卷积注意力模块 CBAM[23].

最后, 使用与P_{Text}相同结构的图像映射器P_{Img}将图像特征映射至与文本特征相同的语义空间, 得到图像特征:

$$f_{\text{I}} = P_{\text{Img}}(CBAM(h_{\text{I}})) \tag{7}$$

2.2　细粒度语义对齐模块

为了将新闻文本与图像的局部细节建立精确的语义关联, 本文提出了细粒度语义对齐模块, 其网络结构如图 2 所示.

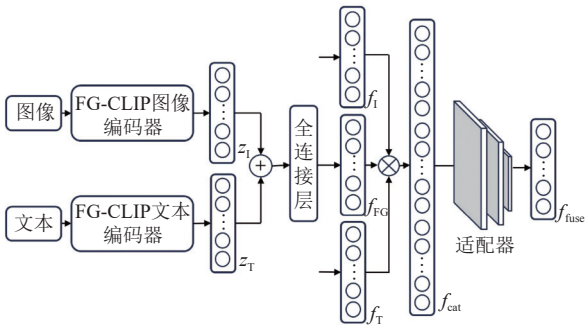


图 2　细粒度语义对齐模块

该模块的核心思想在于构建一个共享语义空间, 并在该空间内实现图像与文本特征的对齐. 为此, FG-CLIP 引入了区域级对比学习 (regional contrastive learning) 和硬负样本 (hard fine-grained negative sample) 学习机制, 以在图像内部的候选区域与其对应的文本描述之间建立细粒度对齐, 同时提升模型区分语义上非常相近但不一致的能力. 在区域对齐过程中, 从输入图像中抽取多个候选区域, 并将输入文本段落与这些区域建立对应关系. 然后, 利用 FG-CLIP 模型的文本编码器E_{Text}和图像编码器E_{Img}, 分别将这些配对的文本段落和图像区域映射到一个共享的语义向量空间中, 并提取跨模态语义特征向量:

$$z_{\text{T}} = E_{\text{Text}}(x_{\text{T}}) \tag{8}$$

$$z_{\text{I}} = E_{\text{Img}}(x_{\text{I}}) \tag{9}$$

在该语义空间中, 通过计算图像特征向量z_{I}与文本特征向量z_{T}之间的余弦相似度来衡量两种模态的语义一致性:

$$s_{ij} = \frac{z_{\text{T}}^{i} \cdot z_{\text{I}}^{j}}{\left\|z_{\text{T}}^{i}\right\| \left\|z_{\text{I}}^{j}\right\|} \tag{10}$$

其中, s_{ij} 表示第 i 个文本与第 j 个图像区域的相似度. 为实现多模态特征的有效对齐, FG-CLIP 采用对比学习策略构建损失函数, 通过最大化匹配样本之间的相似度、最小化非匹配样本之间的相似度, 从而在共享语义空间中实现对齐优化:

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ij}/\tau)} - \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ji}/\tau)} \right] \tag{11}$$

其中, N为批量大小, τ为温度系数, 用于调节分布平滑度. 该损失函数通过双向约束确保每个图像与其对应文本在语义空间中保持最高相似度, 而与不相关文本或图像的相似度尽可能低. 通过最小化上述对比损失, 模型能够在共享语义空间中实现图像与文本的细粒度对齐.

硬负样本策略引入那些与正样本语义高度相似但实际不匹配的区域-文本对, 从而迫使模型学习更细粒度的语义差异. 这不仅增强了模型区分正负样本的能力, 也提高了对微小语义差异的判别精度.

本文使用预训练的 FG-CLIP, 并通过迁移学习实现多模态新闻的对齐. 首先, 通过逐元素求和操作以实现初步的跨模态语义对齐:

$$f_{\text{FG}} = FC(z_{\text{T}} + z_{\text{I}}) \tag{12}$$

其中, FC 表示全连接层, 用于进一步捕捉融合特征内部的语义交互关系, 并在特征层面进行细粒度语义调优, 最终得到具有细粒度语义对齐能力的融合表示f_{FG}.

随后, 将f_{T}、f_{I}和f_{FG}在通道维度进行拼接, 得到整合后的多模态表示:

$$f_{\text{cat}} = concat(f_{\text{T}}, f_{\text{FG}}, f_{\text{I}}) \tag{13}$$

其中, $concat$表示拼接操作, 这不仅保留了各模态中原始的判别性特征, 还显式引入了 FG-CLIP 所建模的细粒度语义对齐信息, 从而提升整体的跨模态表征能力.

最后, 将f_{cat}输入适配器模块以实现多模态特征的深度对齐与融合. 该适配器结构由多个串联的全连接层、ReLU 激活函数以及 Dropout 组成:

$$f_{\text{fuse}} = Adapter(f_{\text{cat}}) \tag{14}$$

其中, $Adapter$表示适配器模块, 在训练阶段的参数可学习, 而 FG-CLIP 主体结构保持冻结, 从而实现了参数高效的微调策略. 该适配器不仅提供了非线性建模能力与正则化效果, 同时借助 FG-CLIP 的细粒度语义对齐能力, 使f_{cat}中的局部语义交互信息得到充分利用, 从而有效抑制无关噪声区域的干扰, 建立更精准的跨模态细粒度语义对应.

2.3 多模态特征聚合与新闻分类

在获得多模态融合特征f_{fuse}后, 进一步将其与图像和文本的单模态特征进行聚合. 考虑到在新闻图文语义一致时, 融合特征f_{fuse}可能包含冗余信息, 进而导致特征过拟合问题. 因此, 模型在多模态特征聚合阶段引入单模态特征, 增强模型对不同语义一致性的适应能力.

首先将f_{T}、f_{fuse}以及f_{I}分别通过全连接层FC_1、FC_2、FC_3, 映射为二维 logit 向量. 然后, 将 3 个 logit 向量逐元素求和, 得到最终聚合特征f_{final}:

$$f_{\text{final}} = FC_1(f_{\text{T}}) + FC_2(f_{\text{FG}}) + FC_3(f_{\text{I}}) \tag{15}$$

随后, 经过 $Sigmoid$ 激活函数, 得到预测概率:

$$\hat{y} = Sigmoid(f_{\text{final}}) \tag{16}$$

其中, $\hat{y}$表示模型对输入新闻的预测概率.

在训练阶段, 采用二元交叉熵损失函数对模型进行优化. 设批量大小为 N, 真实标签为y_i, 预测概率为$\hat{y}_i$,

则损失函数为:

$$\mathcal{L}_{\mathrm{CE}}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\left(\hat{y}_i\right)+\left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right] \tag{17}$$

该设计在保留多模态互补信息的同时, 有效缓解了特征冗余带来的过拟合问题, 从而提升虚假新闻检测的准确性与鲁棒性.

2.4 优势模态识别与增强模块

该模块旨在突出判别能力更强的优势模块, 抑制劣势模态的干扰, 从而提升虚假新闻检测的准确率和泛化能力.

首先, 借鉴原型网络的思想, 对图像和文本模态进行性能评估. 具体而言, 从训练集中按照一定比例随机取出部分新闻, 分别提取每条新闻的图像特征f_{I}和文本特征f_{T}. 对每个类别k, 计算其原型向量:

$$\begin{cases} c_{\mathrm{T}}^{(k)}=\dfrac{1}{N_k}\displaystyle\sum_{i=1}^{N_k}f_{\mathrm{T}}^{(k_i)} \\ c_{\mathrm{I}}^{(k)}=\dfrac{1}{N_k}\displaystyle\sum_{i=1}^{N_k}f_{\mathrm{I}}^{(k_i)} \end{cases} \tag{18}$$

其中, $k\in\{0,1\}$分别表示“假”和“真”两类新闻, N_k为类别k下的样本数量.

然后, 计算样本特征与原型的欧氏距离, 并通过Softmax 函数将距离转换为置信度:

$$\begin{cases} p_{\mathrm{T}}^{(i)}\left(y=y_i\right)=\dfrac{\exp\left(-D\left(f_{\mathrm{T}}^{(i)},c_{\mathrm{T}}^{(y_i)}\right)\right)}{\displaystyle\sum_{j=0}^{K}\exp\left(-D\left(f_{\mathrm{T}}^{(i)},c_{\mathrm{T}}^{(j)}\right)\right)} \\ p_{\mathrm{I}}^{(i)}\left(y=y_i\right)=\dfrac{\exp\left(-D\left(f_{\mathrm{I}}^{(i)},c_{\mathrm{I}}^{(y_i)}\right)\right)}{\displaystyle\sum_{j=0}^{K}\exp\left(-D\left(f_{\mathrm{I}}^{(i)},c_{\mathrm{I}}^{(j)}\right)\right)} \end{cases} \tag{19}$$

其中, y_i为样本标签, 而$D(A,B)=\sqrt{\displaystyle\sum_{i=1}^{n}(A_i-B_i)^2}$表示向量$A$与向量$B$之间的欧氏距离. 随后, 对批量大小为$N$的样本进行累加, 得到模态置信度:

$$\begin{cases} R_{\mathrm{T}}=\displaystyle\sum_{i=0}^{N}p_{\mathrm{T}}^{i} \\ R_{\mathrm{I}}=\displaystyle\sum_{i=0}^{N}p_{\mathrm{I}}^{i} \end{cases} \tag{20}$$

当R_{T}>R_{I}时, 文本模态为优势模态; 反之, 图像模态为优势模态.

完成模态评估后, 引入原型交叉熵损失以增强优势模态的表征能力.

$$\mathcal{L}_{\mathrm{PCE}}=-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(-E\left(f^{(i)},c^{(y_i)}\right)\right)}{\displaystyle\sum_{j=0}^{K}\exp\left(-E\left(f^{(i)},c^{(j)}\right)\right)}\right) \tag{21}$$

其中, f和c分别表示优势模态的特征向量及其对应类别的原型.

最终, 模型的总损失函数由二元交叉熵损失与原型交叉熵损失组成.

$$\mathcal{L}_{\mathrm{final}}=\mathcal{L}_{\mathrm{CE}}+\mathcal{L}_{\mathrm{PCE}} \tag{22}$$

该机制能够自适应地突出优势模态, 强化其判别作用, 从而提升虚假新闻检测的准确率与泛化能力.

3 实验与分析

3.1 数据集

为了全面评估所提出模型的性能, 本文使用两个多模态假新闻检测领域常用的公开数据集: Weibo 和GossipCop. 这两个数据集均来源于社交媒体平台, 涵盖了真实新闻与虚假新闻的图文样本, 具体信息如表 1 所示.

表 1 数据集统计信息

数据集	真实新闻数量	虚假新闻数量	新闻总数量
Weibo	4 799	4 749	9 528
GossipCop	10 259	2 581	12 840

在实验过程中, 按照 7:1:2 的比例将数据集划分为训练集、验证集和测试集, 以保证评估的科学性与可比性.

3.2 实验设置及评价指标

实验环境基于 Ubuntu 20.04 操作系统, 采用NVIDIA RTX 4060 显卡, 并使用 PyTorch 深度学习框架实现模型训练与测试. 训练过程中使用 Adam 优化器更新参数, 初始学习率设置为 0.001, 批量大小为 64, 迭代轮数为 50.

本实验使用多模态虚假新闻检测任务常用的准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall)、$F1$ 分数 ($F1$-score) 作为模型性能的评估指标.

3.3 对比实验与结果分析

表 2 列出了本文提出的 FADM 模型与其他基线模型 SAFE[24]、BDANN[10]、MCAN[13]、CAFE[5]、BMR[12]、FND-CLIP[18], 在 GossipCop 和 Weibo 数据集上的性能对比, 加粗数值为最优结果.

表 2 FADM 模型与基线模型的对比实验结果

数据集	方法	Accuracy	虚假新闻			真实新闻		
			Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
Weibo	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	BDANN	0.851	0.869	0.836	0.852	0.832	0.866	0.849
	MCAN	0.910	0.927	0.899	0.912	0.897	**0.922**	0.908
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	FND-CLIP	0.907	0.914	0.901	0.908	0.914	0.901	0.907
	BMR	0.916	**0.924**	0.901	0.912	0.900	0.921	0.920
	FADM	**0.926**	0.912	**0.950**	**0.931**	**0.942**	0.898	**0.920**
GossipCop	SAFE	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	BDANN	0.865	0.705	0.517	0.597	0.891	0.948	0.919
	MCAN	0.877	0.710	0.604	0.653	0.909	0.941	0.922
	CAFE	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	FND-CLIP	0.880	0.761	0.549	0.638	0.899	0.959	0.928
	BMR	0.883	0.717	**0.647**	**0.680**	**0.917**	0.939	0.928
	FADM	**0.886**	**0.777**	0.569	0.657	0.903	**0.965**	**0.931**

通过分析实验结果可知: 本文提出的 FADM 模型在两个主流数据集中的大多数评估指标取得了最优结果, 验证了其有效性与鲁棒性. 在 Weibo 数据集上, FADM 的检测准确率达到了 92.6%, 较次优模型提升了 1%; 在对虚假新闻的检测中, 召回率和 *F*1 分数分别提升了 4.9% 和 1.9%. 在 GossipCop 数据集上, FADM 的准确率为 88.6%, 较次优模型提升了 0.3%, 其中虚假新闻的检测精确率提高了 6%. 虽然整体准确率提升幅度有限, 但在虚假新闻检测中精确率的显著改善更具实际意义, 这表明模型在识别虚假新闻时具有更高的可靠性, 能够更有效地减少误报.

与 FND-CLIP 模型相比, FADM 在大多数评估指标上表现更优, 其原因在于 FND-CLIP 仅从整体语义层面对图文一致性进行建模, 在复杂场景下难以捕获图文之间的关键区域的细粒度语义对应关系, 导致模型在图文局部语义理解方面存在不足, 进而影响其在虚假新闻检测任务中的整体性能表现; 而本文模型提出的细粒度语义对齐模块有效捕捉图文之间的局部对应关系, 从而提升模型对多模态信息的综合理解能力.

与 CAFE 模型相比, FADM 模型对虚假检测性能更优, 原因在于 CAFE 模型仅从信息论角度量化文本与图像之间的歧义性, 并根据歧义性程度自适应选择单模态或跨模态特征. 然而, CAFE 在处理语义歧义时, 对两个单模态特征赋予相同的权重, 这没有考虑到发生语义歧义的其中一个原因是包含较少语义信息的弱模态干扰了模型的检测; 而本文提出由语义信息更丰富的优势模态主导虚假新闻检测的策略, 能够充分利用语义信息更丰富的模态, 显著提升检测效果.

3.4 消融实验与结果分析

为验证本文模型中各个关键模块有效性, 本文进行了消融实验, 实验结果如表 3 所示. 首先通过移除模型中各个关键模块, 得到以下变体模型.

表 3 消融实验结果比较

数据集	方法	Accuracy	虚假新闻			真实新闻		
			Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
Weibo	移除优势模态识别与增强模块	0.902	0.905	0.909	0.907	0.898	0.894	0.907
	移除细粒度语义对齐模块	0.896	0.911	0.889	0.900	0.880	0.903	0.891
	仅使用图像模态	0.660	0.696	0.630	0.661	0.628	0.694	0.659
	仅使用文本模态	0.888	**0.920**	0.863	0.891	0.858	**0.917**	0.886
	完整的FADM模型	**0.926**	0.912	**0.950**	**0.931**	**0.942**	0.898	**0.920**
GossipCop	移除优势模态识别与增强模块	0.873	0.724	0.558	0.630	0.900	0.949	0.924
	移除细粒度语义对齐模块	0.867	0.733	0.484	0.583	0.886	0.958	0.921
	仅使用图像模态	0.735	0.345	0.339	0.342	0.832	0.836	0.834
	仅使用文本模态	0.784	0.478	**0.656**	0.533	0.903	0.817	0.858
	完整的FADM模型	**0.886**	**0.777**	0.569	**0.657**	**0.903**	**0.961**	**0.931**

(1) 移除优势模态识别与增强模块: 不对两个模态的虚假新闻检测能力进行评估与增强, 让两种模态的特征以同等权重进行融合.

(2) 移除细粒度语义对齐模块: 仅使用 BERT 提取的文本特征和 ResNet50 提取的图像特征, 将两者直接拼接后进行虚假新闻检测.

(3) 仅使用文本模态: 仅使用 BERT 模型提取的文本特征进行检测.

(4) 仅使用图像模态: 仅使用 ResNet50 提取的图像特征进行检测.

结果表明, 完整的 FADM 模型在各项指标上均优于其变体, 说明各模块均对整体性能的提升发挥了重要作用.

首先, 多模态方法整体优于单模态方法, 说明同时使用新闻的图像与文本进行检测能够有效利用图文的互补信息. 而仅使用新闻图像进行检测的性能最低, 说明图像模态在虚假新闻检测任务中判别信息有限, 而文本模态的贡献, 应在模型检测过程中充分发挥其主导作用.

在移除了细粒度语义对齐模块后, 模型在 Weibo 和 GossipCop 数据集上的准确率分别下降 3% 和 1.9%, 性能显著下降, 表明 FADM 模型有效迁移了 FG-CLIP 的图文细粒度语义对齐能力, 并在新闻检测中发挥了重要作用. 在去除了优势模态识别与增强模块后, 模型在 Weibo 和 GossipCop 数据集上的准确率分别下降 2.4% 和 1.3%, 验证了所提出的优势模态增强策略能够有效缓解弱模态干扰, 从而进一步提升虚假新闻检测性能.

3.5 子集比例与学习率选择分析

本实验在 Weibo 数据集上进一步分析了子集比例与学习率两个关键超参数对模型性能的影响, 实验结果如表 4 所示. 其中, 对于优势模态识别与增强模块, 从训练集中选择不同比例的子集计算原型会对优势模态的识别与增强效果不同. 本实验设置不同的子集比例进行实验, 其中, 0% 表示不使用该模块.

从表 4 中可以看出, 随着子集比例的增大, 准确率逐渐上升, 在 5% 时达到最好结果 92.6%. 然而, 随着子集比例的增大, 准确率逐渐减小至 91.5% 左右. 这是因为使用过小规模的训练集计算出的原型, 与真实原型有较大的偏差, 使得优势模态的训练效果不佳. 而使用太大规模的训练集计算出的原型会使模型出现过拟合现象, 同样导致优势模态无法得到充分训练, 从而影响模型的整体性能.

表 4 不同子集比例下的准确率对比

子集比例 (%)	准确率
0	0.902
1	0.920
5	0.926
10	0.918
20	0.917
50	0.918
100	0.915

学习率的大小同样对模型性能、训练速度和泛化能力有一定的影响. 为了探究不同学习率对模型性能的影响, 本实验分别设置学习率大小为: 0.001、0.000 5、0.000 1 进行实验分析, 实验结果如图 3、图 4 所示.

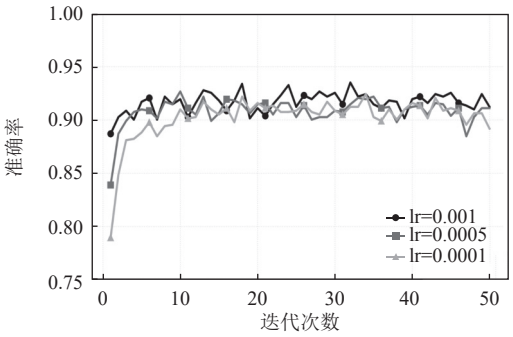


图 3 训练过程中的准确率变化折线图

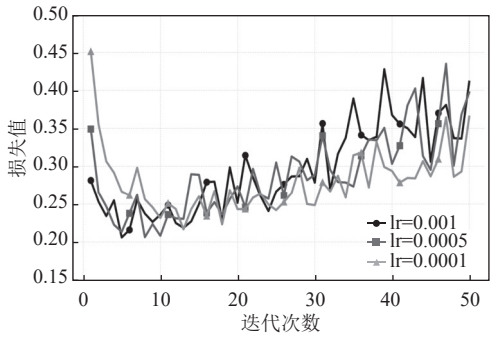
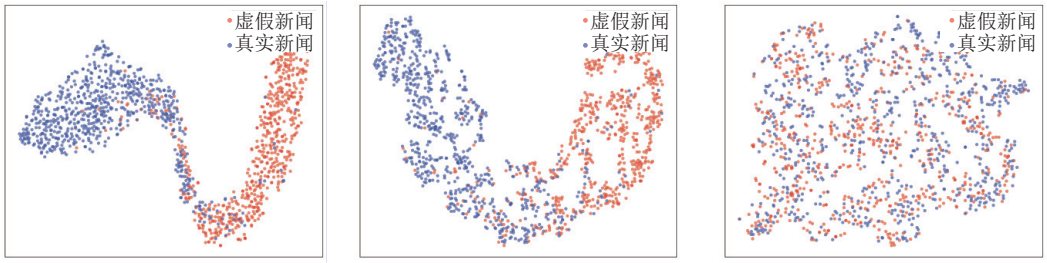


图 4 训练过程中的损失值变化折线图

从图 3 中可以看出, 相较于其他两个学习率, 使用大小为 0.001 的学习率, 模型可以较快地收敛, 并且准确率明显高于其他两个学习率. 从图 4 中可以看出使用 0.001 学习率进行训练的过程中, 其损失值在 15 个迭代轮次之前最低, 但是随着迭代次数的增加, 其损失值波动较为剧烈, 并且超过其他两个学习率, 这是由于迭代次数过多导致模型出现过拟合现象, 而具有更大学习率的模型更容易出现过拟合.

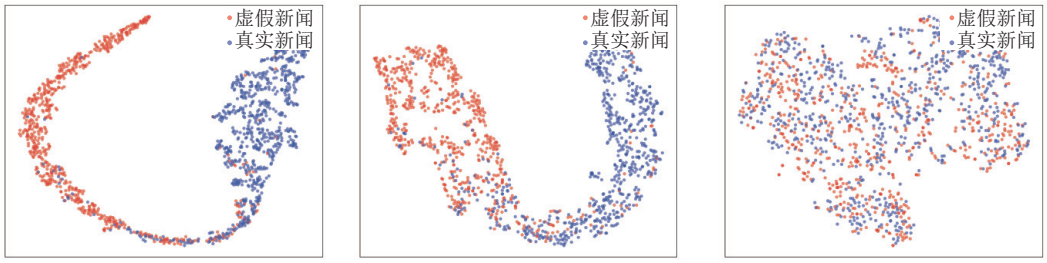
3.6 可视化分析

为进一步分析模型所提取特征的判别能力, 本文使用 Weibo 测试集, 使用 t-SNE 算法对融合特征、文本模态特征、图像模态特征进行二维可视化分析. 其中相同颜色的点表示属于同一类别的特征. 实验结果如图 5、图 6 所示.



(a) 融合特征 (b) 文本模态特征 (c) 图像模态特征

图 5 移除模态增强模块的 FADM 模型的特征可视化



(a) 融合特征 (b) 文本模态特征 (c) 图像模态特征

图 6 FADM 模型的特征可视化

首先, 从整体来看, 融合特征相比于单一模态特征具有更清晰的决策边界, 能够更好地区分真实新闻与虚假新闻, 说明多模态模型有效利用不同模态之间的互补信息, 从而获得更优的性能. 其次, 文本模态特征较图像模态特征更易区分, 表明文本包含更多判别性语义信息, 而图像信息有限且噪声较多. 因此, 文本模态在虚假新闻检测中更具优势.

进一步对比图 5 与图 6 可见, 引入了优势模态识别与增强模块后, 文本模态特征呈现出更高的类内聚集度与更清晰的类间分离度, 使得真实新闻与虚假新闻在嵌入空间中形成了更加明显的边界. 这一效果源于该模块能够自动识别文本模态为优势模态, 并引导其特征向原型方向聚类, 从而增强判别能力. 融合特征的可视化结果也进一步印证了该机制的有效性, 即优势模态的引导能够显著提升整体检测性能与鲁棒性.

3.7 案例分析

为了直观展示模型的检测效果, 本文从 Weibo 数据集中选择了两则具有代表性的多模态虚假新闻样例进行案例分析.

图 7(a) 展示了一则虚假新闻样例, 模型预测出了正确的结果. 可以看出, 模型能够将文本中的“东京申奥成功”与新闻图像中包含的人物、场景、旗帜等信息建立语义关联, 进而判断出图文信息在该部分具有一致性, 因此认为该部分新闻为真实信息. 然而, FADM 模型检测出“东京奥运预算为北京 1/10 安倍内阁已宣布将为奥运会投资 45 亿美元, 比起 2008 年北京奥运会花费 430 亿美元, 仅为 1/10.”为谣言. 同时无法从图像中检测出证实这一信息的依据, 因此判断这一消息为假, 模型最终判断这一则新闻为虚假新闻.

图 7(b) 展示了一则真实新闻样例, 但模型误将其判定为虚假新闻. 这是因为新闻图像仅包含一块牛肉, 缺乏充分的语义信息. 因此模型根据模态评估机制将文本模态识别为优势模态, 并通过原型网络机制增强其表征学习能力. 然而, 由于模型对文本语义的理解存在偏差, 误将新闻文本判断为虚假信息, 进而导致整体误判.

通过以上两个案例可以看出, 在多模态数据均包含有效语义信息时, 细粒度语义对齐机制能够有效提升 FADM 模型的判断准确性, 这是由于不同模态间的语义互补在检测过程中发挥了关键作用. 当某一模态

信息缺失或语义不足时, 本文提出的基于原型网络的模态增强机制虽能在一定程度上缓解信息缺失带来的影响, 但仍存在误判的可能性, 这为后续改进提供了方向.

(a) 检测正确案例

(b) 检测错误案例

图 7 FADM 模型案例分析

4 结束语

本文针对多模态虚假新闻检测中跨模态细粒度语义对齐能力不足, 难以捕捉图文间的深层语义关联; 以及在模态信息量不均衡时, 等权融合易使优势模态受弱势模态干扰, 降低检测效果这两大问题, 提出了基于FG-CLIP 的细粒度对齐机制和基于原型网络的优势模态识别与增强模块, 构建了 FADM 模型. 实验结果表明, FADM 在多个指标上优于现有方法, 且通过消融实验、可视化分析与参数敏感实验验证了其有效性与鲁棒性. 未来将进一步结合外部知识与大规模多模态模型, 以提升检测的实时性与准确性.

参考文献

1 Shu K, Sliva A, Wang SH, *et al*. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22–36. [doi: 10.1145/3137597.3137600]

2 Jin ZW, 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: ACM, 2017. 795–816. [doi: 10.1145/3123266.3123454]

3 Singhal S, Shah RR, Chakraborty T, *et al*. SpotFake: A multi-modal framework for fake news detection. Proceedings of the 5th IEEE International Conference on Multimedia Big Data. Singapore: IEEE, 2019. 39–47. [doi: 10.1109/BigMM.2019.00-44]

4 Li JN, Selvaraju RR, Gotmare AD, *et al*. Align before fuse: Vision and language representation learning with momentum distillation. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 742. [doi: 10.5555/3540261.3541003]

5 Chen YX, Li DS, Zhang P, *et al*. Cross-modal ambiguity learning for multimodal fake news detection. Proceedings of the 2022 ACM Web Conference. Lyon: ACM, 2022. 2897–2905. [doi: 10.1145/3485447.3511968]

6 Singhal S, Pandey T, Mrig S, *et al*. Leveraging intra and inter modality relationship for multimodal fake news detection. Companion Proceedings of the Web Conference 2022. Lyon: ACM, 2022. 726–734. [doi: 10.1145/3487553.3524650]

7 Xie C, Wang B, Kong F, *et al*. FG-CLIP: Fine-grained visual and textual align-ment poster. Proceedings of the 2025 International Conference on Machine Learning. 2025. 12345–12356.

8 Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4080–4090. [doi: 10.5555/3294996.3295163]

9 Khattar D, Goud JS, Gupta M, *et al*. MVAE: Multimodal variational autoencoder for fake news detection. Proceedings of the 2019 World Wide Web Conference. San Francisco: ACM, 2019. 2915–2921. [doi: 10.1145/3308558.3313552]

10 Zhang T, Wang D, Chen HH, *et al*. BDANN: BERT-based

domain adaptation neural network for multi-modal fake news detection. Proceedings of the 2020 International Joint Conference on Neural Networks. Glasgow: IEEE, 2020. 1–8. [doi: 10.1109/IJCNN48605.2020.9206973]

11 Devlin J, Chang MW, 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: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/V1/N19-1423]

12 Ying QC, Hu XX, Zhou YM, *et al*. Bootstrapping multi-view representations for fake news detection. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2023. 5384–5392. [doi: 10.1609/aaai.v37i4.25670]

13 Wu Y, Zhan PW, Zhang YJ, *et al*. Multimodal fusion with co-attention networks for fake news detection. Proceedings of the 2021 Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2021. 2560–2569.

14 Jing J, Wu HC, Sun J, *et al*. Multimodal fake news detection via progressive fusion networks. Information Processing & Management, 2023, 60(1): 103120.

15 Shen XR, Huang MW, Hu Z, *et al*. Multimodal fake news detection with contrastive learning and optimal transport. Frontiers in Computer Science, 2024, 6: 1473457. [doi: 10.3389/fcomp.2024.1473457]

16 Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8748–8763.

17 Pramanick S, Roy A, Patel Johns VM. Multimodal learning using optimal transport for sarcasm and humor detection. Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022. 546–556.

18 Zhou YM, Yang YZ, Ying QC, *et al*. Multimodal fake news detection via CLIP-guided learning. Proceedings of the 2023 IEEE International Conference on Multimedia and Expo. Brisbane: IEEE, 2023. 2825–2830.

19 Pahde F, Puscas M, Klein T, *et al*. Multimodal prototypical networks for few-shot learning. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 2643–2652.

20 Fan YF, Xu WC, Wang HZ, *et al*. PMR: Prototypical modal rebalance for multimodal learning. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 20029–20038.

21 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010. [doi: 10.5555/3295222.3295349]

22 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

23 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

24 Zhou XY, Wu JD, Zafarani R. SAFE: Similarity-aware multi-modal fake news detection. Proceedings of the 24th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Singapore: Springer, 2020. 354–367. [doi: 10.1007/978-3-030-47436-2_27]

(校对责编: 张重毅)