计算机科学
Computer Science
ISSN 1002-137X,CN 50-1075/TP

《计算机科学》网络首发论文

题目：改进的跨模态关联歧义学习的虚假信息检测方法研究
作者：段钰潇，胡艳丽，郭浩，谭真，肖卫东
收稿日期：2023-09-15
网络首发日期：2024-02-22
引用格式：段钰潇，胡艳丽，郭浩，谭真，肖卫东．改进的跨模态关联歧义学习的虚假信息检测方法研究[J/OL]．计算机科学.
https://link.cnki.net/urlid/50.1075.TP.20240221.1555.010

改进的跨模态关联歧义学习的虚假信息检测方法研究

段钰潇　胡艳丽　郭浩　谭真　肖卫东

国防科技大学信息系统工程重点实验室　长沙　410073
（duanyuxiao19@nudt.edu.cn）

摘　要　近年来，随着互联网及多媒体技术的迅猛发展，人们获取信息更加方便快捷，然而虚假信息在网络上的传播也日益严重,负面影响不断扩大。为了增强信息的可信度和欺骗性，虚假信息呈现多模态发展趋势，使得检测工作面临更大挑战。现有的多模态虚假信息检测方法大多关注多模态特征的形成，对于跨模态歧义和不同模态特征在检测中贡献率的研究尚不完善，忽略了不同模态特征间固有差异性对虚假信息检测的影响。为解决该问题，提出了构建改进的跨模态关联歧义学习的虚假信息检测模型，通过对文本和图像特征进行跨模态歧义学习，利用歧义得分更新单模态与融合特征的权重，自适应地拼接单模态与融合特征；同时采用网格搜索动态分配文本、图像特征权重，提高检测准确率。在 Twitter 数据集上对该模型的有效性进行验证，相比于基线模型准确率提高了 6%，相比于未进行动态权重分配的检测方法性能提升了 1.6%。

关键词：　虚假信息检测；多模态；跨模态关联；歧义学习；融合特征

中图分类号　TP391　　DOI：10.11896/jsjkx.230900087

Study on Improved Fake Information Detection Method Based on Cross-modal Correlation Ambiguity Learning

DUAN Yuxiao，HU Yanli，GUO Hao，TAN Zhen and XIAO Weidong

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China

Abstract　In recent years, with the rapid development of the Internet and multimedia technology, it is more convenient for people to obtain information, but the spread of fake information on the Internet is also increasingly serious, and the negative impact is constantly expanding. In order to enhance the credibility and deception, fake information presents a multi-modal development trend, which makes the detection work face greater challenges. The existing multi-modal fake information detection methods pay more attention to the formation of multi-modal features. The research on the contribution rate of cross-modal ambiguity and different modal features in detection is not perfect, ignoring the impact of inherent differences among different modal features on fake information detection. To solve the problem, this paper proposes to construct an improved fake information detection model based on cross-modal correlation ambiguity learning. Through cross-modal ambiguity learning of text and image features, the weights of unimodal features and fused features are updated by the ambiguity score. The unimodal features and fused features are combined adaptively, and the weights of text and image features are dynamically assigned by grid search to improve the detection accuracy. The effectiveness of the model is verified by experiments on the Twitter dataset. The accuracy is improved by 6% compared with the baseline model and 1.6% compared with the detection without dynamic weight assignment.

Keywords　Fake news detection，Multimodal，Cross-modal correlation，Ambiguity learning，Fusion features

1　引言

社交平台已经成为人们获取信息的重要渠道，它提供了便捷的环境，用户可以轻松地创作、加工和分享信息。然而，社交平台的开放性和无限制的信息传播机制降低了虚假信息的产生和传播门槛。近年来，在社交富媒体化潮流下，互联网上用户发布的内容形式由传统文本向图文并茂类型发展。虚假新闻的发布者开始利用具有

到稿日期：2023-09-15　　返修日期：2023-11-06
基金项目：国家自然科学基金（62272469，72301284）；国家重点研发计划（2022YFB3102600）；湖南省科技创新计划（2023RC1007）
This work was supported by the National Natural Science Foundation of China(62272469,72301284),National Key R&D Program of China (2022YFB3102600) and Science and Technology Innovation Program of Hunan Province(2023RC1007).
通信作者：谭真（tanzhen08a@nudt.edu.cn）

误导性或经过篡改的图片来吸引读者的注意，加剧其传播，对经济发展、社会稳定等领域产生了严重的负面影响[1]。例如，虚假新闻宣称大量饮酒可以给身体消毒，导致 COVID-19 爆发期间近 800 人因此死亡[2]。多模态虚假信息的检测在互联网环境治理、稳定舆论导向等方面均发挥着至关重要的作用，成为学术界和工业界的研究热点[3]。

多模态虚假信息检测旨在考虑文本、图像等多种模态特征以及跨模态间的关联关系，并融合多模态特征和跨模态关联关系，进行真实性检测。在不失一般性的前提下，本文聚焦多模态虚假信息的文本模态及图像模态的特征。现有多模态虚假信息检测方法大多关注多模态特征的生成，如基于变分编码器的多模态融合框架[4]，以及端到端事件对抗性神经网络[5]。这类方法能够较充分地利用多模态信息，学习不同模态间的共享特征表示，在该领域取得了较大突破，一定程度上提高了检测准确率。然而，现有模型缺乏对单模态特征和融合特征在检测中贡献率的研究。当图文匹配程度低，即跨模态歧义大时，依靠跨模态相关性就能够精准判断信息的真假，此时融合特征发挥的作用较大；当图文匹配程度高时，融合特征发挥的作用较小。因此，融合特征的重要性与跨模态歧义性紧密关联，基于跨模态关联歧义性进行各个模态特征及融合特征的自适应结合，提高虚假信息检测准确率是本文的研究重点。

针对上述问题，本文提出了一种改进的跨模态关联歧义学习的虚假信息检测模型（Improved Cross-modal Correlation Ambiguity Learning for Multimodal Fake News Detection，IC^{2}LFD），首先利用跨模态歧义学习框架计算图、文模态之间的歧义值；基于学到的歧义值确定单模态和融合模态特征的权重；并运用网格搜索方法动态分配文本、图像单模态特征的权重，拼接生成新特征用于真实性检测。本方案能够在一定程度上降低模态间固有歧义对检测的影响，填补传统方法中对于跨模态歧义学习以及单模态贡献度差异的研究空缺。通过大量实验分析证明了本文提出的改进方法的性能优异性。本文的主要贡献总结为以下 3 点：

（1）采用基于网格搜索的动态权重分配方法，基于跨模态关联歧义学习的歧义值，动态结合文本模态和图像模态对虚假信息检测的贡献度，进一步提升了检测准确率。

（2）模型的有效性在多模态 Twitter 新闻数据集上得到了验证，与现有模型相比，IC^{2}LFD 取得了最优的检测效果。

（3）通过消融实验验证了对齐模块、歧义学习模块和融合模块的有效性。

本文其余部分的组织结构如下：第 2 节讨论相关工作；第 3 节介绍 IC^{2}LFD 模型；第 4 节描述实验和结果；最后总结全文并展望未来。

2 相关工作

文本类虚假信息检测方向的研究起源早，效果也在不断优化。如 Popat 等[6]设计的端到端言论验证模型，结合了 Bi-LSTM 和注意力机制，运用新闻言论语句与外部证据进行检测。有研究者考虑到虚假信息的时间特征[7]，利用虚假新闻的动态传播结构，采用图神经网络方法对虚假信息进行检测。还有一些学者利用社交网络的传播特征构建模型进行虚假信息检测[8]，同样提高了检测效果。

近年来，发布者为了提高虚假信息的可信度和欺骗性，发布的内容呈多模态趋势，因此，研究者也开始聚焦多模态虚假信息检测。Khattar 等[4]提出了一种基于变分自编码器的多模态融合

框架，用于学习不同模态间共享的潜在表示。Wang 等[5]提出了一种用于假新闻检测的端到端事件对抗性神经网络，核心思想是设计一种学习事件不变特征，并保留所有事件间共享特征的方法，以便对新出现的未知事件进行假新闻检测。Xue 等[9]使用 BERT 对文本信息进行建模，使用 ResNet 对视觉信息进行建模，以计算它们之间的相似度。Zhou 等[10]通过全连接层将文本特征和图像特征映射到同一向量空间中，通过二者相似度的高低来判定信息真假。Zhang 等[11]运用多种深度学习模型将图像映射为语义标签，计算新闻文本、图像内容与图像语义之间的一致性。Singhal 等[12]采用修正的对比损失来建立不同模态之间的关联。Cao 等[13]考虑了图像嵌入文字、图像内实体和图像的特征信息，利用 co-attention 注意力机制融合实现实体增强。Qi 等[14]为了更好理解深层语义，提出了语义增强的多模态虚假新闻检测方法。Chen 等[15]使用跨模态对齐模块将异构单模态特征转换为共享语义空间。Hua 等[16]在利用 BERT 提取文本特征后，进一步利用 BERT 提取文本和视觉特征，以增强两种模态特征之间的相互强化。Ying 等[17]创新性地提出了多模态粗判断机制，运用 MLP 进行跨模态一致性学习，通过 iMMoE 网络融合不同模态特征。

现有研究表明多模态虚假信息检测在多方面都取得了一定成果。但客观来看，这些研究成果尚未成熟，还有很多值得改进的地方。

3 模型介绍

本研究旨在验证跨模态歧义性和单模态贡献率差异在虚假信息检测中的重要性，提高多模态虚假信息检测准确率。图 1 展示了所提出的 IC²LFD 模型的基本流程，该模型分为 6 个模块，分别为特征提取、跨模态特征对齐、跨模态特征融合、分类器。

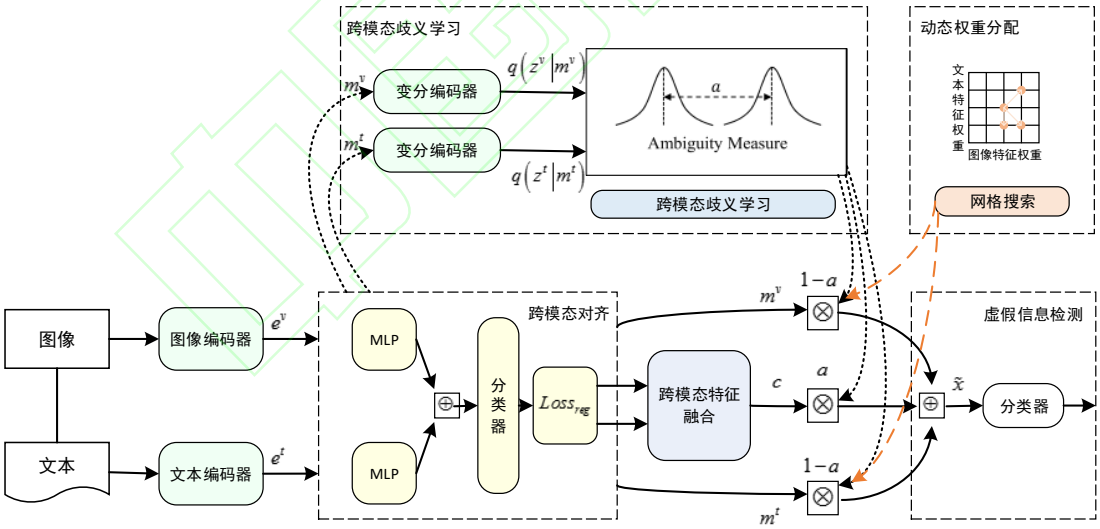

图 1 IC²LFD 模型结构

Fig.1 Framework of IC²LFD model

3.1 特征提取

3.1.1 文本特征

文本是信息的重要组成部分，包含丰富语义信息，在虚假新闻检测中发挥着不可替代的作用。本研究运用 BERT 语言模型提取文本语义信息。BERT 通过双向编码表示，通过有效捕捉潜在语义和上下文信息来表示单词和句子关系。本模型使用基础版 BERT 模型，其包含 12 个编码

层，隐藏层大小为 768，结构如图 2 所示。

对于互联网上的文本信息 x^t，首先对其进行预处理，将文本内容分词为序列 $W=[w_1,w_2,\ldots,w_n]$，进一步转化为 BERT 模型所需的输入序列结构，包括词嵌入、位置嵌入、分割嵌入 3 部分。词嵌入是对每个单词进行向量化处理，位置嵌入是对每个单词的位置信息编码，分割嵌入是对句子间关系的编码，三者拼接。模型充分利用以上信息，经过多层 Transformer 编码器处理后，输出隐状态序列 $\boldsymbol{O}_{\text{text}}=[o_1,o_2,\ldots,o_n]$。每个单词对应的隐状态 o_j 均包含全局的上下文信息。调整维度，将其转化为文本特征序列 $\boldsymbol{R}_{\text{text}}=[t_1,t_2,\ldots,t_n]$，其维度为 $d\times n$，每个文本特征向量 t_j，维度为 $d\times 1$。使用平均池化方法将文本特征序列 $\boldsymbol{R}_{\text{text}}$ 的维度转化为 $d\times 1$，作为最终文本特征表示 e_t.

$$e_t = AVG\_POOLING(R_{\text{text}}) = \frac{t_1 \oplus t_2 \oplus \ldots \oplus t_n}{n} \quad (1)$$

其中，$\oplus$ 表示向量对应位置的元素相加。

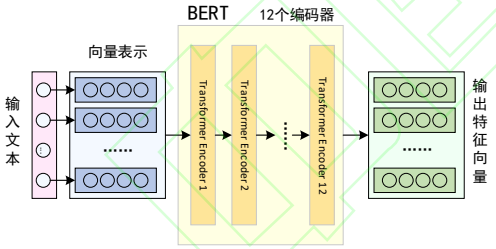

图 2 文本特征提取器

Fig.2 Text feature extractor

3.1.2 图像特征

图像一般包括像素级模式特征和语义特征，有学者将两种特征分开提取，但是模式特征和语义特征之间存在较强的相关性，分开提取会给后续特征对齐与融合任务造成冗余，因此本研究采用 ResNet-34 方法提取图像整体特征。

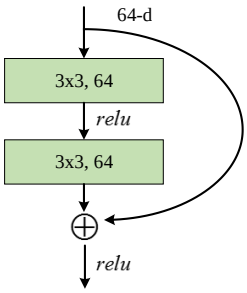

图 3 ResNet-34 残差单元

Fig.3 ResNet-34 residual unit

每一个残差单元包含两个 3×3 的卷积层，如图 3 所示。本文中给定图像 x^v，经过 ResNet-34 网络的处理，得到图像特征表示，维度为 512，通过全连接层转换该特征表示得到 e^v。

3.2 跨模态特征对齐

为便于后续特征融合与歧义学习，本研究首先将特征变换到同一语义空间，即语义正则化。

构建跨模态特征对齐模型，包括特定模态的多层感知机 MLP 和一个模态共享层 D，以共享语义。将 MLP+D 处理后的联合特征送入平均池化层，全连接层为二元分类器。整体模型训练采用余弦相似度衡量损失值，具体损失函数表示为：

$$L_{\text{reg}} = \begin{cases} 1-\cos\left(e^t,e^v\right), y_2=1 \\ \max\left(0,\cos\left(e^t,e^v\right)-d\right), y_2=0 \end{cases} \quad (2)$$

模型训练的目标是使真实新闻，即语义正相关的文本-图像对的特征余弦相似度最大化，而负相关的余弦相似度最小化，直至达到特定边界。最终得到在共享语义空间对齐的单模态表示 m^t 和 m^v，将其作为后续跨模态歧义学习和跨模态融合的输入向量。

3.3 跨模态歧义学习

在得到语义对齐的单模态表示后，引入一种歧义学习方法，通过评估两个模态的变异自编码

概率分布的 KL 散度来表示单模态间的差异性分数。

对单模态特征的生成过程进行建模，从具有各向同性的高斯先验分布的潜在空间 $\mathbb{R}^d$ 中抽取得到单模态特征 m^t 和 m^v。

具体而言，单模态观测值的变分后验值表示为：

$$q(z|m)=N\left(z|\mu(m),\sigma(m)\right) \tag{3}$$

其中均值 μ 和方差 σ 能够从特定模态编码器中获取到。

更进一步，对于每一个具有对齐文本特征 m_i^t 和图像特征 m_i^v 的样本数据 x_i 来说，文本特征和图像特征的变分后验值分别为：

$$q\left(z_i^t|m_i^t\right)=N\left(z_i^t|\mu\left(m_i^t\right),\sigma\left(m_i^t\right)\right) \tag{4}$$

$$q\left(z_i^v|m_i^v\right)=N\left(z_i^v|\mu\left(m_i^v\right),\sigma\left(m_i^v\right)\right) \tag{5}$$

考虑到在整个数据集上的分布，计算其均值：

$$q\left(z^t\right)=E_{P\left(m^t\right)}\left[q\left(z^t|m^t\right)\right]=\frac{1}{N}\sum_{i=1}^{N}q\left(z_i^t|m_i^t\right) \tag{6}$$

$$q\left(z^v\right)=E_{P\left(m^v\right)}\left[q\left(z^v|m^v\right)\right]=\frac{1}{N}\sum_{i=1}^{N}q\left(z_i^v|m_i^v\right) \tag{7}$$

然后，计算单模态分布的平均 KL 散度值来衡量数据集 x_i 上不同模态的模糊度：

$$a_i^1=\frac{KL\left(q\left(z_i^t|m_i^t\right)\middle\|q\left(z_i^v|m_i^v\right)\right)}{KL\left(q\left(z^t\right)\middle\|q\left(z^v\right)\right)} \tag{8}$$

$$a_i^2=\frac{KL\left(q\left(z_i^v|m_i^v\right)\middle\|q\left(z_i^t|m_i^t\right)\right)}{KL\left(q\left(z^v\right)\middle\|q\left(z^t\right)\right)} \tag{9}$$

$$a_i=\mathrm{sigmoid}\left(\frac{1}{2}\left(a_i^1+a_i^2\right)\right) \tag{10}$$

其中 a_i 为歧义得分。根据 a_i 的大小，可以直观地评估该样本的文本特征和图像特征的分布差异性。a_i 作为跨模态融合特征的权重，用于调节融合特征和单模态特征的比重。

3.4 跨模态特征融合

跨模态关联可以捕捉不同模态之间的语义交互作用，为虚假信息检测提供补充特征。因此本文设计了跨模态特征融合模块，学习跨特征关联。在前文基础上，首先计算单模态特征之间的注意力权重值 InterC。

InterC 的计算方法为对初始特征映射进行平方根归一化，再经 softmax 函数处理，得到两组模间权值（文-图，图-文）：

$$InterC_{t\leftarrow v}=\mathrm{softmax}\left(\left[m^t\right]\left[m^v\right]^T\Big/\sqrt{\dim}\right) \tag{11}$$

$$InterC_{v\leftarrow t}=\mathrm{softmax}\left(\left[m^v\right]\left[m^t\right]^T\Big/\sqrt{\dim}\right) \tag{12}$$

更新单模态特征，得到显式相关映射如下：

$$\hat{m}^t=InterC_{T\leftarrow V}\times m^t \tag{13}$$

$$\hat{m}^v=InterC_{V\leftarrow T}\times m^v \tag{14}$$

以往研究通常采用简单拼接方式得到跨模态融合特征，这种方法容易忽略不同模态间复杂的交互作用，因此本文采用交互矩阵 $\boldsymbol{c}$ 来融合文本特征和图像特征。

$$\boldsymbol{c}=\hat{m}^t\otimes\hat{m}^v \tag{15}$$

其中，$\otimes$ 定义为外部乘积，最终交互矩阵 $\boldsymbol{c}$ 被平展为平面向量 $\boldsymbol{c}'$，作为跨模态融合特征。

3.5 动态权重分配

在前文基础上，基于歧义得分，连接向量，得到综合特征表示 $\tilde{x}$。

$$\tilde{x}=\left(a_i\times\boldsymbol{c}'\right)\oplus\left(\left(1-a_i\right)\times m^t\right)\oplus\left(\left(1-a_i\right)\times m^v\right) \tag{16}$$

其中，$\oplus$ 表示连接操作，跨模态融合特征权重为 a_i，单模态特征权重为 $1-a_i$。

$\tilde{x}$ 的文本特征和图像特征所占比重一致，然而现实生活中的互联网信息，往往以模态表达方式为主。如图 4 所示，在阅读这一则新闻时往往会先注意到图片所描述的事物，而后阅读文字。此处图像占主要位置，文字起到辅助解释的作用，因而文本特征的权重低于图像特征。因此考虑进一步细化文本和图像特征的权重分配。

为了找到最佳单模态权重参数组合，采用网格搜索方法。具体而言，首先确定待搜索的参数空间，经排列组合生成所有可能的权重组合，动态调整文本和图像特征的权重分配。

Sharks in the mall! After the hurricane sandy!

图 4 多模态新闻实例

Fig.4 Example of multimodal news

针对每一种权重组合，引入模型，通过计算检测的准确率评估性能，选择准确率最高的组合作为最佳权重组合。准确率的计算函数为：

$$f = accuracy\_score(pre\_label, real\_label) \tag{17}$$

动态权重分配的实现算法如算法 1 所示。

算法 1 基于网格搜索的动态权重分配算法

输入：参数空间 M，模型准确率函数 f，数据集 D
输出：最佳权重组合，最高准确率

```
初始化最佳权重组合 q = [1,1]
初始化最高准确率 best_acc = 0.0
for [a, b] in M:
    使用当前权重组合训练模型 IC²LFD([a, b])
    acc = f([a,b])
    if acc > best_acc:
        best_acc = acc
        q = [a,b]
end for
return q, best_acc 返回最佳权重组合和最高准确率
```

3.6 虚假信息分类器

本研究采用 MLP 进行检测，采用 softmax 函数处理，将输出结果转化为概率分布。

$$\tilde{y}_1 = \mathrm{softmax}\left(MLP\left(\tilde{x}\right)\right) \tag{18}$$

考虑到虚假信息检测为一个二分类问题，采用二值交叉熵损失函数作为模型损失值的计算函数。

$$L_{\mathrm{cls}}\left(y_1, \tilde{y}_1\right) = y_1 \log\left(\tilde{y}_1\right) + \left(1-\tilde{y}_1\right)\log\left(1-y_1\right) \tag{19}$$

其中 y_1 为真实值，$\tilde{y}_1$ 为预测值。

跨模态对齐阶段的损失函数对于虚假新闻检测这一任务的帮助并不突出，因此引入新权重 $\beta \in (0,1)$ 来限制其效果。综合以上情况，最终得到总损失函数表达式为：

$$L = L_{\mathrm{cls}} + \beta \cdot L_{\mathrm{reg}} \tag{20}$$

4 实验设计与分析

4.1 数据集

本研究采用由 MediaEval 建立的为帮助开发人员研究跨越主题和事件领域的虚假信息所建立的 Twitter 数据集[18]，具体信息如表 1 所列。

表 1 Twitter 数据集统计信息

Table 1 Twitter dataset statistics

数据集划分	真实新闻	虚假新闻	总计
训练集	6 840	5 007	11 847
测试集	717	689	1 406
总计	7 557	5 696	13 253

该数据集涵盖了跨越多个主题和事件领域的语料，按照文献[18]的方法将其划分为训练集与测试集。在预处理阶段，去除相关用户、发布时间和地理位置等信息，删除不包含文本或图像的新闻。在文本编码器中，设置输入文本长度不超过 200 个字。在图像编码器中，设置输入图像尺寸为 224×224 。

4.2 实验设置

为提高模型的普遍适用性，并参考已有研究方法，本文实验参数的设置选择通用尺寸。在文

本编码器中，将每个文本编码为 256 维的特征向量。跨模态对齐使用 3 个全连接层，每一层包含 64 个隐藏单元。跨模态融合中交互矩阵 $\boldsymbol{C}$ 被平展为维度为 64×64 的向量。损失函数 L_{reg} 中的边界 d 设置为 0.2，L_{cls} 中的超参数 β 设置为 0.5。训练过程中，为增加模型鲁棒性并提高训练效率，使用 Adam 优化器，默认激活函数为 ReLU。模型迭代次数为 100，批处理大小为 25，学习率 1×10^{-4}，采用 dropout 防止过拟合。

为了评估算法性能，本文采用了广泛使用的评估指标，包括准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 值。由于 F1 值能够综合精确率和召回率，体现算法在正负样本中的效果，因此我们重点关注 F1 值。其次关注准确率，因为考虑到数据集相对均衡，而准确率能够反映模型性能。

本模型由 Pytorch1.7.1 构建，代码运行系统为 Ubuntu，采用型号为 NVIDIA GeForce RTX 3090 的 GPU 进行模型训练。

4.3 动态权重分配实验

首先进行动态权重分配实验，来确定虚假信息分类器最适用的文本特征与图像特征权重比值。运用网格搜索方法，控制变量，遍历文本特征和图像特征权重在一定范围内的所有组合，不改变其他参数。考虑到计算资源消耗与现实新闻特征，本文主要聚焦于 [0,2] 空间上，步长 0.1。针对每一组合，采用准确率评估性能，选择最佳权重组合。

实际实验遍历了步长 0.1 的所有情况，为突出展示，本文选取 [0,2] 上具有代表性的权重值，更新具体结果如表 2 所列。

表 2 权重更新实验结果

Table 2 Results of weight updating experiment

文本	图像	*Accuracy*	*F*1
1	1	0.806	0.803
1	0.5	0.770	0.740
1	0.8	0.802	0.807
1	1.5	0.714	0.737
1	1.8	0.699	0.739
1	2	0.608	0.704
0.5	1	0.754	0.745
0.8	**1**	**0.813**	**0.819**
1.5	1	0.776	0.733
1.8	1	0.710	0.760
2	1	0.697	0.712

表中最优解用加粗表示，次优解用下划线表示。对比准确率和 F1 值，能够发现当文本与图像权重之比为 1:0.8，1:1，0.8:1 时，准确率和 F1 值均高于 0.8，而 0.8:1 时效果更好。分析得到这一权重比值的原因：不同数据集中文本和图像发挥作用的重要程度不同，一些新闻数据以文字为主导，也有一些以图像为主导。在 Twitter 数据集中，图像特征发挥的作用略大于文本特征，检测效果更佳。因此 IC²LFD 模型的文本、图像单模态特征的权重设定为 0.8:1。

4.4 对比实验

为了进一步凸显 IC²LFD 模型的先进性，本文选择 Att_RNN[19]，MVAE[4]，MCAN[20] 和 CAFE[21]模型进行对比实验。

在 IC²LFD 模型基础上，去除跨模态歧义学习，不改变融合特征权重，不更新单模态权重分配，作为基线模型。

Att_RNN 模型是运用注意力机制融合文本、图像和社交网络特征信息进行检测。为保证测试变量一致，这里删去了社交网络特征信息。

MVAE 模型学习文本和图像的共享表示，使用变分自编码器和二进制分类器进行分类。

MCAN 模型采用多注意力层融合模态特征，先融合图像的频域特征和空域特征，再融合文本特征，从最后共同关注层得到的融合特征表示用于虚假信息检测。

CAFE 模型考虑了不同模态间的差异性，引入跨模态歧义学习，而文本、图像单模态特征权重值均为 1。对比实验结果如表 3 所列。

表 3 模型性能对比

Table 3 Performance comparison of models

模型	*Accuracy*	*Precision*	*Recall*	*F*1
Att_RNN	0.779	0.778	0.799	0.788
MVAE	0.745	**0.801**	0.719	0.758
MCAN	0.809	0.732	0.871	0.795
CAFE	0.806	0.807	0.799	0.803
基线模型	0.720	0.512	0.893	0.651
IC²LFD	**0.813**	0.739	**0.917**	**0.819**

结果显示，IC²LFD 模型的 F1 值为 0.819，准确率为 0.813，均为最高值。相比于基线模型，F1 值提高了 0.168，准确率提高了 0.061；相比于 CAFE 模型，F1 值提高了 0.016，准确率提高了 0.007。与目前流行的模型进行对比，验证了所提模型的先进性，以及跨模态关联歧义学习和单模态贡献差异的重要性。

MVAE 模型准确率和 F1 值均为 4 种模型中最低，原因在于该模型设计受到需要较好领域知识的限制；Att_RNN 模型对于语义关联性的处理能力低于 IC²LFD 模型；而 MCAN 模型性能在已有模型中最佳，其优势可能在于进一步提高了处理多模态信息的能力，但对跨模态差异的关注度不够；CAFE 模型引入了歧义学习，但是忽略了文本、图像特征发挥作用大小不等的问题。

IC²LFD 模型相比于以上四者，对多模态特征融合的思考更细致，更加关注跨模态歧义和单模态特征的重要性差异，在准确率、召回率以及 F1 值上都明显优于基线模型和已有模型，充分说明了跨模态歧义学习和单模态权重差异在虚假信息检测中的重要性。歧义学习能够在不同歧义水平下权衡单模态特征和跨模态特征，自适应地调整权重；单模态特征在检测中发挥作用不同，动态权重分配确定最佳权重组合能够提高检测准确率。

4.5 消融实验

4.5.1 单模态检测

本研究设计单模态检测实验来探究单模态特征在检测中发挥的作用。对比方法如下;

（1）BERT：仅使用文本作为模型输入，采用 BERT 模型提取文本特征，输出基于文本的检测结果。

（2）ResNet-34：仅使用图像作为模型输入，采用 ResNet-34 提取图像特征，输出基于图像的检测结果。

表 4 单模态检测结果

Table 4 Unimodal detection results

模型	Accuracy	Precision	Recall	F1
BERT	0.758	**0.752**	0.769	0.760
ResNet-34	0.684	0.400	**0.953**	0.564
基线模型	0.720	0.512	0.893	0.651
IC²LFD	**0.813**	0.739	0.917	**0.819**

结果如表 4 所列，表中每项评价指标最优值用加粗表示，次优值用下划线表示。准确率和 F1 值能够更直观地反映性能。可见，相比于 IC²LFD 模型，文本或图像模态特征的单独预测准确率和 F1 值均较低，说明单一模态进行虚假信息检测利用的信息较片面，效果欠佳。而基于文本的 BERT 模型的检测结果在准确率、精确率和 F1 值上均优于基线模型，可见该情况下，将文本、图像特征重要性视为一致的多模态检测不如直接采用文本的单模态检测，只有合理利用多模态信息才能够提高虚假信息检测效果。

4.5.2 模块消融

为进一步研究 IC²LFD 模型各个组件的有效性，进行了 3 组消融实验，分别为去除跨模态对齐模块、去除跨模态歧义学习模块、去除跨模态融合模块。

（1）IC²LFD w/o Alignment：去除跨模态对齐模块，使用单模态嵌入学习进行特征融合。

（2）IC²LFD w/o Ambiguity：去除跨模态歧义学习模块，即本模型的基线模型。

（3）IC²LFD w/o Fusion：去掉跨模态融合模块，使用新闻文本和图片作为模型输入，经跨模态对齐处理后直接拼接，输入分类器中。结果如表 5 所列。

通过对比实验结果可知，去除模态对齐和模态歧义学习模块对检测结果的影响较大，准确率分别下降了 0.124, 0.093，F1 值分别下降了 0.146, 0.168。

表 5 IC²LFD 模型消融实验

Table 5 IC²LFD model ablation experiment

模型	*Accuracy*	*Precision*	*Recall*	*F*1
IC²LFD	**0.813**	**0.739**	0.917	**0.819**
w/o Alignment	0.689	0.574	0.814	0.673
w/o Ambiguity	0.720	0.512	0.893	0.651
w/o Fusion	0.792	0.642	**0.929**	0.759

由此可知，不同模态特征的对齐处理、权衡单模态和融合模态特征都能够显著提高检测性能。而去除跨模态融合模块使得检测准确率下降 0.021，F1 值下降 0.06，精确率下降较多，达 0.097，召回率却上升 0.012。考虑到检测的随机性影响，这一实验结果反映了跨模态融合比简单拼接单模态特征来进行虚假信息检测更有效。整体来看，去除模型的任意一个模块，都会导致检测效果降低，验证了每个模块的必要性。

结束语

本文围绕多模态虚假信息检测这一中心任务，以“图文”歧义为切入点，提出了 IC²LFD 模型；对单模态特征的权重比例进行探究，得到最佳文本图像权重组合为 0.8:1。采用 Twitter 数据集设计实验，IC²LFD 模型的 F1 值为 0.819，准确率为 0.813。F1 值相比基线模型提升了 0.168，相比未进行动态权重分配的 CAFE 模型提升了 0.016。实验不仅验证了跨模态歧义学习的有效性和先进性，更证明了在多模态虚假信息检测中，文本和图像的重要性并不相同。进一步地，经消融实验，验证了各模态特征和各个模块的必要性。

互联网更新迭代飞快，大语言模型的出现和普及深刻影响着人类生活，多模态虚假信息不断复杂化，虚假信息检测仍存在许多亟待解决和完善的地方。在本文研究的基础上，未来多模态检测可以增加信息传播特征和用户多维度特征，进一步挖掘更多模态间的深层关联，借助预训练模型、对比学习、知识驱动等技术，设计更高效、更便捷的检测模型，阻断虚假信息的传播，为网络安全提供科技支撑。

参考文献

[1] JIN Z, CAO J, WANG B, et al. Research on social multimedia rumor detection technology integrating multi-modal features [J]. Journal of Nanjing University of Information Science and Technology (Natural Science Edition), 2017, 9 (6): 583-592. (in Chinese)

金志威，曹娟，王博，王蕊，张勇东.融合多模态特征的社会多媒体谣言检测技术研究[J].南京信息工程大学学报(自然科学版),2017,第 9 卷(6): 583-592

[2] ISLAM S, SARKAR T,KHAN S H, et al. COVID-19-related infodemic and its impact on public health: A global social media analysis[J].Am. J. Trop. Med. Hyg,2020, 103 (4), 1621–1629.

[3] CAO J, SHENG Q, QI P. Progress and prospect of Internet false information detection [J]. Communication of China Computer Society, 2020,16 (3): 52-57. (in Chinese)

曹娟，盛强，亓鹏.互联网虚假信息检测进展与展望[J].中国计算机学会通讯,2020,第 16 卷(3): 52-57

[4] KHATTAR D, GOND J, GUPTA M, et al. MVAE: Multimodal Variational Autoencoder for Fake News Detection[C]//The World Wide Web Conference. 2019: 2915-2921.

[5] WANG Y Q, MA F L, JIN Z W,et al. EANN: Event Adver-sarial Neural Networks for Multi-Modal Fake News Detection[C]//In ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD). 2018:849–857.

[6] POPAT K, MUKHERJEE S, YATES A, et al. DeClarE: Debunking Fake News and False Claims Using Evidence-Aware Deep Learning[C]// Proceeding of the 2018 Conference on Empirical

Methods in Natural Language Processing (EMNLP), Brussels, Belgium. USA: ACL. 2018:22-32.

[7] SONG C, SHU K, WU B. Temporally evolving graph neural network for fake news detection[J].Information Processing & Management, 2021,58 :102712.

[8] LIU Y, WU Y-F B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]// The 32th AAAI Conference on Artificial Intelligence (AAAI-18). 2018: 354-361.

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

[10] ZHOU X, WU J, ZAFARANI R. Safe: Similarity-aware multi-modal fake news detection[J]. arXiv: 2003.04981,2020.

[11] ZHANG G, LI J. Detecting Social Media Fake News with Semantic Consistency Between Multi-model Contents[J]. Data Analysis and Knowledge Discovery, 2021, 5(5):21-29. (in Chinese)

张国标，李洁．融合多模态内容语义一致性的社交媒体虚假新闻检测[J]．数据分析与知识发现, 2021, 5(5): 21-29.

[12] SHIVANGI S, MUDIT D, RAJIV R S,et al. Inter-Modality Discordance for Multimodal Fake News Detection[C]//In ACM Multimedia Asia (MMAsia'21), December 1-3.2021:1-7.

[13] QI P, CAO J, LI X R,et al. Improving Fake News De-tection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues[C]//ACM MM21.2021:1212-1220.

[14] QI P, CAO J, SHENG Q. Semantics-Enhanced Multi-Modal Fake News Detection[J]. Journal of Computer Research and Development,2021,58(7): 1456-1465. (in Chinese)

亓鹏，曹娟，盛强.语义增强的多模态虚假新闻检测[J].计算机研究与发展,2021,第 58 卷(7): 1456-1465.

[15] CHEN Y, LI D, ZHANG P, et al. Cross-modal ambiguity learning for multimodal fake news detection[C]//Proceedings of the ACM Web Conference .2022:2897–2905.

[16] HUA J, CUI X D, LI X H, et al. Multimodal fake news detection through data augmentation-based contrastive learning[J].Applied. Soft Computing, 2023,136 (C), 1568–4946.

[17] YING Q, HU X, ZHOU Y, et al. Bootstrapping multi-view representations for fake news detection[C]//AAAI .2023.

[18] 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.

[19] JIN Z, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia. 2017: 795-816.

[20] WU Y, ZHAN P, ZHANG Y, et al. Multimodal Fusion with Co-Attention Networks for Fake News Detection[OL[.https://aclanthology.org/2021.findings-acl.226/.

[21] CHEN Y X, LI D S,ZHANG P, et al. Cross-modal Ambiguity Learning for Multimodal Fake News Detection[C]//Proceedings of the ACM Web Conference 2022 (WWW'22), April 25-29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA,2022:2897-2905.

段钰潇，出生于 2001 年，硕士研究生，主要研究方向为知识融合与虚假信息检测。

胡艳丽，出生于 1979 年，博士，副教授，CCF 会员，主要研究方向为自然语言处理知识图谱。

郭浩，出生于 1997 年，博士研究生，研究方向为知识融合与虚假新闻检测。

谭真，出生于 1991 年，博士，副教授，国家自然科学基金获得者，CCF 会员，主要研究方向为知识图谱和智能问答。

肖卫东，出生于 1968 年，博士，教授，CCF 会员，研究方向为数据智能。

DUAN Yuxiao， born in 2001, postgraduate. Her main research interests include knowledge fusion and fake news detection .

TAN Zhen，born in 1991, Ph.D, associate professor. His main research interests include knowledge graph and intelligent Q&A.

（责任编辑：何杨）