基于对比学习的多模态注意力网络虚假信息检测方法

李卓远[1,2]，李 军[1,2]

（1. 北京信息科技大学人工智能安全科技创新团队，北京 100192；2. 北京信息科技大学信息管理学院，北京 100192）

摘 要：针对近年来网络空间中大量涌现的多模态虚假信息难以有效检测的问题，重点提出一种基于对比学习预训练和注意力机制的多模态虚假信息检测方法，使用对比学习对不同模态数据之间进行特征对齐和潜在关系学习，并采用注意力机制实现不同模态特征之间的交互，通过特征融合完成模型构建，最终实现对虚假信息的精准检测。所提出的模型对于多模态虚假信息的检测相较于当前主流方法取得了更好的效果，基本能够对虚假信息实现更准确的识别和检测。

关键词：网络空间安全；虚假信息检测；多模态特征融合；对比学习；注意力机制

中图分类号：TP309.2 **文献标志码**：A

文章编号：2095-2783(2023)11-1192-06 **开放科学(资源服务)标识码(OSID)**：

Contrastive learning-based multimodal attention networks for false information detection method

LI Zhuoyuan[1,2], LI Jun[1,2]

(1. *Artificial Intelligence Security Technology Innovation Team, Beijing Information Science and Technology University, Beijing* 100192, *China*; 2. *School of Information Management, Beijing Information Science and Technology University, Beijing* 100192, *China*)

Abstract: Multi-modal false information has intensively emerged in the cyberspace in recent years and its effective detection is difficult. Facing this problem, this paper focused on proposing a multi-modal false information detection method based on contrastive learning pre-training and attention mechanism. Contrastive learning was used to perform feature alignment and potential relationship learning between different modal data, and attention mechanism was used to realize the interaction between different modal features. Finally, the model was constructed by feature fusion to achieve accurate detection of false information. Compared with the current mainstream methods, the proposed model herein achieves better results in detecting multi-modal false information, and can basically achieve more accurate identification and detection of false information.

Keywords: cyberspace security; false information detection; multimodal feature fusion; contrastive learning; attention mechanism

虚假信息具有渠道多、速度快、范围广等特性，其真实性很难及时得到验证。随着互联网的爆炸式发展，网络成为人与人之间沟通的重要平台，尤其是各种社交软件和媒体的出现，在给人们带来便利的同时，也无形中降低了制造虚假信息的成本，为虚假信息的传播创造了温床，这不断地威胁着人民的生活和社会的稳定，造成了恶劣的影响[1]。不法分子可以利用人们所关心的热点话题来制造虚假信息，恶意引导舆论的走向，引发人们的恐慌，从而达到其不良目的。

现如今，随着人们日常生活中网络带宽的不断提升，虚假信息的存在形式逐渐由单模态文本形式向跨模态图文结合形式转变。而随着卷积神经网络研究的不断深入，针对该类问题的解决方案也逐渐从通过文本中提取语言特征进行单模态的虚假信息检测发展为将从文本中提取的特征和从图片中提取的特征相结合来对跨模态的虚假信息进行检测[2]。这些方法考虑到了不同模态数据特征之间的关联关系，通过交互不同模态之间的关联关系来实现对多模态虚假信息的检测。虽然这些方案具有一定可行性，但也存在一些缺点，首先是不同模态的特征向量在不同特征空间下很难形成更好的交互关系[3]；其次，在其他的多模态联合表征中，很多忽略了单一模态内的交互关系而只注重了不同模态间的交互。

为了解决上述问题，本文提出一种基于对比学习预训练的多模态注意力网络模型：首先，利用双向编码器表征预训练模型(bidirectional encoder representation from transformers，BERT)对待测样本中的文本模态数据进行语言特征提取，利用残差网络(residual network，ResNet)对待测样本中的图像模态数据进行图像特征提取，之后再对不同模态数据提取到的特征向量进行对比学习预训练，通过最小化对比损失函数将不同模态特征向量映射到同一特征空间，实现特征对齐。之后，引入跨模态注意力机

收稿日期：2023-03-22

基金项目：国家自然科学基金资助项目(U1936111)；北京信息科技大学 2021“勤信拔尖人才”培育计划项目(QXTCPB 202104)；“慧眼行动”资助项目(20E08F79)；国防科工局基础科研项目(JCKY2022405C010)

第一作者：李卓远(1996—)，男，硕士研究生，主要研究方向为网络空间安全、虚假信息检测

通信作者：李军，副教授，主要研究方向为网络空间安全、人工智能应用及安全，lijun@bistu.edu.cn

制来捕捉不同模态之间的高层交互关系，根据不同模态间学习到的相关性权重来更新视觉模态和语言模态的特征向量。同时通过模态内自注意力机制来获得单模态内不同特征向量之间的复杂关系。最后，融合不同模态的特征得到最终联合表征，将最终联合表征投射到二分类空间从而实现对待测信息的真伪性判断。

1　相关工作

现阶段，针对虚假信息的检测已成为国内外研究的热点，主要采用机器学习和深度学习的方法，并已取得了有效的研究成果。在单模态虚假信息检测方面，主要是使用深度学习模型来解决语言模态的虚假信息检测问题。Ma 等[4]提出了使用循环神经网络(recurrent neural networks，RNN)模型来进行虚假信息检测，通过使用 RNN 模型对单位时间段内获取到的媒体信息文本向量进行特征表示。Bian 等[5]提出了使用双向卷积神经网络(bi-directional graph convolutional networks，Bi-GCN)来自顶向下和自底向上挖掘虚假信息的传播模式，从而实现对虚假信息的检测。这些方法都能够对语言模态的虚假信息进行有效的识别检测，但其检测范围单一，因而具有一定的局限性。

与本文更相关的是多模态虚假信息检测方面，主流方法是获得不同模态特征后，通过特征融合形成联合表征后进行虚假信息检测[6]。刘金硕等[7]通过构建 MSRD 模型实现对网络虚假信息的检测，通过 VGG19 网络对图片进行特征提取，接着通过 DenseNet 提取图片中的内嵌文字，使用长短期记忆(long short-term memory，LSTM)网络提取内嵌文本文字的特征，将提取到的文本语言特征和图片视觉特征进行拼接，并通过全连接层获取语言和视觉层面共享表示的均值和方差向量，最终实现对内嵌文字图片的图文匹配度检测，实现对虚假信息的甄别。Khattar 等[8]通过自动编码器对数据进行重构得到共享表示，再对其进行二分类实现真伪性鉴定。以上 2 个方法只是单纯地通过特征拼接来实现联合表征，忽略了不同模态特征之间的关系，在执行下游检测任务时准确度较差。孟杰等[9]建立了基于多模态深度融合的虚假信息检测网络。通过建立多分支的卷积-循环神经网络(CNN-RNN)来提取图片的多层次特征，使用 Word2Vec 和双向门控循环单元来获得文本的特征，然后使用注意力机制实现模态间和模态内特征交互，并拼接形成最终联合表征进行下游检测任务。Gao 等[10]通过 Faster-RCNN 和门控循环单元分别提取图像数据和文本数据的特征向量，建立动态的模内和模间注意力流框架(dynamic fusion with intra-and inter-modality attention flow，DFAF)对不同模态的特征向量进行交互，并多次迭代模内和模间注意力流框架，使得不同模态特征向量深度融合，得到联合表征，并以此来实现下游任务。上述 2 个方法均采用注意力机制捕捉不同模态间的关系，但不同模态的特征向量分属不同的特征空间，采用不同特征空间的特征向量融合后的联合表征进行虚假信息检测时，其性能仍有较大提升空间，尤其在数据正样本相关性较弱时，其表现往往较差。

2　基于对比学习的多模态注意力网络框架

2.1　问题定义

给定事件组 $U=\{u_1,u_2,\cdots,u_n\}$，集合 U 由 n 个元素组成，集合中每个元素代表固定事件，如 u_i 代表第 i 个事件，在这个事件中包含与事件相关的文本类数据 t 和相对应的图片类数据 p，u_i 表示为 $u_i=\{t,p\}$。对社交媒体等平台进行虚假信息检测工作，本质上就是对多模态事件的二分类任务。给定对应标签组 $L=\{l_1,l_2,\cdots,l_n\}$，在标签组中每个元素代表对应事件的标签 l_i，这些标签将对应的事件 u_i 进行了标注，分为了虚假信息和非虚假信息。

2.2　模型介绍

多模态注意力训练网络主要用于对多模态信息进行特征的提取、交互学习和特征融合，模型会根据全连接层 softmax 函数来判定信息是否为虚假信息。

多模态注意力训练网络主要结构如图 1 所示。多模态待检测信息主要由文本语言类数据和图形视觉类数据组成。为了实现文本和图片等多模态数据的关键要素提取，建立了多模态注意力机制。首先，使用 BERT 预训练模型来实现文本数据的特征提取，再通过 ResNet 网络来对图像模态数据进行特征提取[11]，之后对这 2 个模态的特征向量进行对比学习预训练，再通过多模态注意力机制来捕获语言和视觉模态之间的高层交互，之后进行特征拼接形成多模态的联合表征。在最终表征基础上，通过损失函数和梯度下降法构建分类器权重。多模态注意力机制结构主要由 4 个部分组成：文本特征提取、图片特征提取、特征融合机制、信息分类器。

2.3　语言和视觉数据表征提取

2.3.1　BERT 提取语言数据表征

自然语言处理分为上游任务和下游任务。上游任务负责对数据进行预训练[12]。目前相关方案设计中，上游任务使用频率较高的模型包括 ELMo、Word2Vec 等。其中，ELMo 模型的核心思想体现在深度上下文关系中，通过双向语言模型，根据具体输入来得到上下文依赖的词向量表示，增强了向量特征表示能力，但工作效率较低[13]。Word2Vec 模型将句子中的词语通过高维空间映射成为词向量，可以对单词之间的关系进行表征。Word2Vec 处理后的每个单词与固定的词向量对应，但其没有考虑词序问题，且不能考虑整个长句中上下文相关性，无法

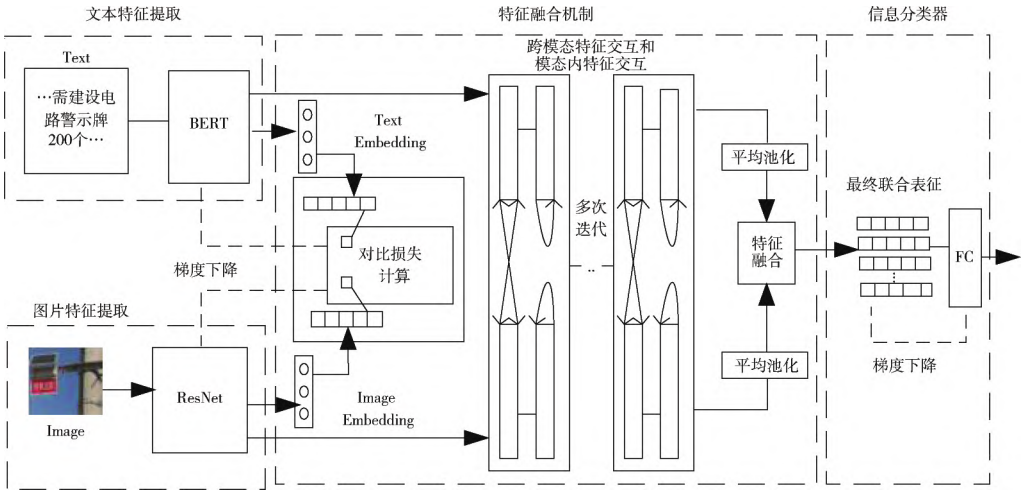

图1 多模态注意力训练网络

Fig. 1 Multimodal attention training network

理解上下文的语义。

BERT模型通过对大规模的未标注语料数据进行训练来获得语言信息的表达。该模型具有并行性好、效率高的特点，可以将数据转化为具有一致性维度的句向量，因此本文采用预训练过的BERT模型来对语言模态数据进行特征向量的提取。将对应文本的长度截断和填充为50，表示为 $\boldsymbol{T}=[T_0,T_1,T_2,\cdots,T_{50}]$，其中 T_0 表示为[CLS]嵌入，再输入BERT模型中提取词特征向量 $\boldsymbol{T}\in\mathbf{R}^{50\times 2\,048}$。

2.3.2 ResNet提取视觉数据表征

ResNet网络相较于传统的VGG网络在特征提取和目标分类等方面有着更好的表现。尤其是在网络层次更深的情况下，ResNet网络解决了传统VGG网络出现的网络退化问题。图片特征提取器负责对图像特征进行捕获。给定输入数据 p，p 为待检测数据对中的图像数据，通过ResNet101网络对图像数据 p 进行基础特征提取，最终获得视觉特征表示为 $\boldsymbol{I}\in\mathbf{R}^{2\,048}$。

2.4 特征融合机制

特征融合机制主要由对比学习预训练、模间注意力交互机制、模内注意力交互机制及多模态融合器4个部分组成。

2.4.1 对比学习预训练

对比学习的宗旨是将相似的数据拉近，同时将不相似的数据进行互斥，从而更好地实现数据表征[14]。本部分的目标为使单模态的特征向量在进行特征交互前达到更好的单模态表示，以增强正样本不同模态数据之间的相关性。

在不同模态特征进行交互前，采用图文对比学习损失函数(image-text contrast loss，ITC)[15]对齐图像和文本的特征。通过文本特征提取和图片特征提取得到文本特征向量和图片特征向量后，通过学习相似度函数得到相似度得分。相似度函数为

$$s=[\boldsymbol{g}_v(\boldsymbol{v})]^{\mathrm{T}}\boldsymbol{g}_w(\boldsymbol{w})。\tag{1}$$

式中，g_v 和 g_w 为将图片特征嵌入(v)和文本特征嵌入(w)映射到标准化后在低维度空间的线性表示。之后计算图文之间的相似度，其中式(2)为视觉模态对语言模态特征的相似度，式(3)为语言模态对视觉模态特征的相似度：

$$y_i(\boldsymbol{I})=\frac{\mathrm{e}^{s(\boldsymbol{I},\boldsymbol{T}_i)/\tau}}{\sum_{i=1}^{M}\mathrm{e}^{s(\boldsymbol{I},\boldsymbol{T}_i)/\tau}},\tag{2}$$

$$y_i(\boldsymbol{T})=\frac{\mathrm{e}^{s(\boldsymbol{T},\boldsymbol{I}_i)/\tau}}{\sum_{i=1}^{M}\mathrm{e}^{s(\boldsymbol{T},\boldsymbol{I}_i)/\tau}}。\tag{3}$$

式中：τ 为可学习的温度参数；$y_i(\boldsymbol{I})$ 和 $y_i(\boldsymbol{T})$ 为ground truth的one-hot相似性，负样本对的概率为0，正样本对的概率为1。

图文对比学习的交叉熵损失定义为

$$\mathcal{L}=\frac{1}{2}\mathrm{E}_{(\boldsymbol{I},\boldsymbol{T})\sim D}[H(y(\boldsymbol{I}),y(\boldsymbol{I}))+H(y(\boldsymbol{T}),y(\boldsymbol{T}))]。\tag{4}$$

在本模型中，采用了计算对比学习损失时最常用的InfoNCE损失函数。在训练过程中，通过最小化InfoNCE损失函数值来拉近图像文本正样本对的特征，具体函数形式如下：

$$\mathcal{L}_{\boldsymbol{I}\sim\boldsymbol{T}}=-\frac{1}{2}\mathrm{E}_{(\boldsymbol{I},\boldsymbol{T})}\left[\log\frac{\mathrm{e}^{s(\boldsymbol{I},\boldsymbol{T})/\tau}}{\sum_{i=1}^{M}\mathrm{e}^{s(\boldsymbol{I},\boldsymbol{T}_i)/\tau}}\right],\tag{5}$$

$$\mathcal{L}_{\boldsymbol{T}\sim\boldsymbol{I}}=-\frac{1}{2}\mathrm{E}_{(\boldsymbol{T},\boldsymbol{I})}\left[\log\frac{\mathrm{e}^{s(\boldsymbol{T},\boldsymbol{I})/\tau}}{\sum_{i=1}^{M}\mathrm{e}^{s(\boldsymbol{T},\boldsymbol{I}_i)/\tau}}\right],\tag{6}$$

$$\mathcal{L}=\mathcal{L}_{\boldsymbol{I}\sim\boldsymbol{T}}+\mathcal{L}_{\boldsymbol{T}\sim\boldsymbol{I}}。\tag{7}$$

式中，$\mathcal{L}$ 为图文对比学习损失函数的值。式(5)为视觉模态对语言模态之间的InfoNCE损失函数计算过程，式(6)为语言模态对视觉模态之间的InfoNCE损失函数计算过程。

最后，根据损失函数 $\mathcal{L}$ 进行梯度下降优化，实现将2个独立特征空间下特征进行特征对齐，从而映射到同一特征空间，增强向量之间的特征相关性，为

之后的跨模态特征交互做准备。

2.4.2　模间注意力交互

模间注意力模块首先使用注意力机制计算不同模态之间的相关性，从而通过所学习到的相关性权重更新文本和图片的特征向量。模间注意力计算公式为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}。\quad (8)$$

式中：Attention(·)为注意力运算函数；$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为 query 矩阵、key 矩阵和 value 矩阵；d 作为防止分子点积值过大的比例因子，其值为输入特征的维度。

将文本特征向量 Text Embeding（以 $\boldsymbol{T}$ 代表，$\boldsymbol{T} \in \mathbf{R}^{50\times 2\,048}$）和图片特征向量 Image Embeding（以 $\boldsymbol{I}$ 代表，$\boldsymbol{I} \in \mathbf{R}^{2\,048}$）输入到模间注意力机制模块中，注意力运算函数过程如下：

$$\boldsymbol{T}_{\text{inter}} = \text{Attention}(\boldsymbol{Q}_{\boldsymbol{T}},\boldsymbol{K}_{\boldsymbol{I}},\boldsymbol{V}_{\boldsymbol{I}}),\quad (9)$$

$$\boldsymbol{I}_{\text{inter}} = \text{Attention}(\boldsymbol{Q}_{\boldsymbol{I}},\boldsymbol{K}_{\boldsymbol{T}},\boldsymbol{V}_{\boldsymbol{T}})。\quad (10)$$

经过计算得到更新后的文本特征矩阵 $\boldsymbol{T}_{\text{inter}}$ 和图片特征矩阵 $\boldsymbol{I}_{\text{inter}}$，作为模内注意力机制的输入矩阵。

2.4.3　模内注意力交互

模内注意力模块是利用注意力计算公式对单模态内进行建模交互，模内注意力机制是文本模态和图片模态交互关系的进一步延伸和补充。其计算过程和模间注意力机制类似，只不过 3 个参数向量均来自单模态内交互，具体计算过程如下：

$$\boldsymbol{I}_{\text{intra}} = \text{Attention}(\boldsymbol{Q}_{\boldsymbol{I}},\boldsymbol{K}_{\boldsymbol{I}},\boldsymbol{V}_{\boldsymbol{I}}),\quad (11)$$

$$\boldsymbol{T}_{\text{intra}} = \text{Attention}(\boldsymbol{Q}_{\boldsymbol{T}},\boldsymbol{K}_{\boldsymbol{T}},\boldsymbol{V}_{\boldsymbol{T}})。\quad (12)$$

经过模内注意力机制后，得到更新后的文本特征 $\boldsymbol{T}_{\text{intra}}$ 和图片特征 $\boldsymbol{I}_{\text{intra}}$。

2.4.4　多模态融合器

在上述过程中，为保证模型准确性，会多次迭代模间注意力机制和模内注意力机制[16]。得到文本特征向量 $\boldsymbol{T}_{\text{intra}}$ 和图像特征向量 $\boldsymbol{I}_{\text{intra}}$ 之后，再对 $\boldsymbol{T}_{\text{intra}}$ 和 $\boldsymbol{I}_{\text{intra}}$ 进行平均池化，得到最终表征 $\boldsymbol{T}_{\text{final}}$ 和 $\boldsymbol{I}_{\text{final}}$，平均池化过程如下：

$$\boldsymbol{T}_{\text{final}} = \text{AvgPool}(\boldsymbol{T}_{\text{new}}),\quad (13)$$

$$\boldsymbol{I}_{\text{final}} = \text{AvgPool}(\boldsymbol{I}_{\text{new}})。\quad (14)$$

最后将池化后的最终表征 $\boldsymbol{T}_{\text{final}}$ 和 $\boldsymbol{I}_{\text{final}}$ 进行拼接，得到联合表征 $\boldsymbol{V}$。对 $\boldsymbol{V}$ 进行线性变换[17]后得到最终联合表征，实现最终联合表征提取。

2.5　信息分类器

在虚假信息检测过程中，将对输入数据的真实性进行判断，因此将该问题定义为二分类问题，即根据前面所得到的最终联合表征（Final Embedding），通过全连接层（FC）将其投影到二分类目标空间，再经过 softmax 函数得到虚假信息识别的概率 p。在训练过程中，采用交叉熵损失函数进行监督学习，损失函数公式如下：

$$\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum\nolimits_i -[y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)]。\quad (15)$$

式中：N 为样本总数，y_i 为第 i 个样本的标签；p_i 为第 i 个样本为虚假信息的概率。

3　实　验

3.1　数据集

为了评估多模态注意力训练网络模型，本实验使用的数据集来自微博平台和 Twitter 平台，该数据集包含从微博社区管理中心和 Twitter 平台爬取的 2017 年 9 月—2021 年 7 月的部分数据，包含来自于娱乐、时政、科技及社会生活等多个领域的新闻信息。去重后得到了微博平台的 7 608 对真实信息和 7 085 对虚假信息以及 Twitter 平台的 6 279 对真实信息和 6 211 对虚假信息。数据集数据统计见表 1。

表 1　数据集数据统计

Table 1　Dataset statistics

数据集平台	标签	数目/对	总数/对
微博	真实信息	7 608	14 693
	虚假信息	7 085	
Twitter	真实信息	6 279	12 490
	虚假信息	6 211	

3.2　实验设置

软硬件环境：AMD Ryzen7 5800H 3.2 GHz，RTX-3090 GPU，Python 3.8.12。在整个网络训练中，初始学习率（lr）设置为 0.01，采用余弦退火学习率衰减策略（Cosine annealing），batch-size 设置为 32，epoch 设置为 100，优化器使用 AdamW。实验参数见表 2。

表 2　实验参数

Table 2　Experimental parameters

参数	参数值
batch-size	32
lr	0.01
epoch	100
optimization	AdamW

3.3　基准模型

本文选取以下几个模型作为对比实验的基准模型，意在突出本文模型的性能。

1）DFAF 模型[6]：该模型集成了跨模态的自注意力机制和跨模态的共注意力机制，并且通过注意力机制建立模间注意流，实现了对不同模态的数据进行特征提取和特征融合。

2）MVAE 模型[4]：该模型通过对文本模态数据和视觉模态数据的共享表示进行学习来对虚假信息进行检测。

3）EANN 模型[18]：该模型通过将图像视觉模态特征和文本语言模态特征进行特征拼接形成多模态

特征，最后通过分类器实现对虚假信息进行检测。

4）VQA（visual question answering）模型[19]：该模型通过将图片模态数据和文本模态数据进行特征提取和拼接形成联合表征，通过二分类实现根据用户给定图片对相应问题进行回答。

3.4 评价方法

本实验选取的常用评价指标为精确率（Precision）、召回率（Recall）、准确率（Accuracy），其中 TP 表示被正确识别为虚假信息的图文对数据数，FP 表示被错误识别的图文对数据数，TN 表示被正确识别的负样本图文对数据数，FN 表示将正样本图文对错误识别为负样本图文对的数据数。

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

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

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{18}$$

$$\text{F1}=\frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}}。 \tag{19}$$

3.5 对比实验

基准模型和多模态注意力训练网络（multimodal attention training network，MATN）模型在微博数据集和 Twitter 数据集上的实验结果见表 3。结果表明，本文 MATN 模型在准确率、F1 等指标方面优于基准模型。使用对比学习预训练对视觉模态和语言模态的数据进行预训练，实现了不同模态之间的特征对齐，从而能够更好地实现不同模态数据的特征向量之间的交互。再经过不同模态间和同一模态内注意力交互以及特征向量融合后，对于多模态虚假信息的检测将有更好的表现。

表 3 MATN 与基准模型在微博和 Twitter 数据集上的对比实验结果

Table 3 Comparative experimental results of MATN and benchmark models on Weibo and Twitter datasets

数据集	基准模型	Accuracy	虚假信息			真实信息		
			Precision	Recall	F1	Precision	Recall	F1
微博	DFAF	0.821	0.818	0.831	0.824	0.820	0.823	0.821
	MVAE	0.764	0.783	0.735	0.758	0.781	0.803	0.792
	EANN	0.753	0.819	0.801	0.810	0.760	0.805	0.782
	VQA	0.621	0.606	0.611	0.608	0.629	0.612	0.620
	MATN(**本文**)	**0.872**	**0.869**	**0.851**	**0.860**	**0.870**	**0.859**	**0.864**
Twitter	DFAF	0.813	0.827	0.839	0.833	0.819	0.822	0.820
	MVAE	0.733	0.787	0.852	0.818	0.800	0.811	0.805
	EANN	0.712	0.818	0.805	0.811	0.811	0.799	0.805
	VQA	0.623	0.623	0.609	0.616	0.623	0.631	0.627
	MATN(**本文**)	**0.878**	**0.874**	**0.860**	**0.867**	**0.877**	**0.880**	**0.878**

注：粗体数字为同列中最优数值。

3.6 消融实验

为了更清晰地了解 MATN 模型各个模块的作用，通过增量的形式，对模型结构各模块进行拆分并逐步组装来由简到繁地进行对比实验，消融实验对比结果见表 4，其中：

1）Text＋Image：由文本特征提取部分、图片特征提取部分组成，得到语言特征向量和视觉特征向量，再进行平均池化后拼接成最终的联合表征。

2）w/o contrastive learning：由文本特征提取部分、图片特征提取部分以及模间/模内注意力交互部分组成。

表 4 消融实验对比结果

Table 4 Comparison of ablation experiments

模型	Accuracy	
	微博	Twitter
Text＋Image	0.637	0.661
w/o contrastive learning	0.806	0.813
w/o inter-att/intra-att	0.775	0.789
MATN(**本文**)	0.872	0.878

3）w/o inter-att/intra-att：由文本特征提取部分、图片特征提取部分以及对比学习预训练部分组成。

4）MATN（**本文**）：包含 MATN 模型的所有模块。

由消融实验结果可以看出：未经过对比学习预训练和注意力机制，只将图片和文本特征经过简单向量拼接的多模态模型对虚假信息数据检测的准确率分别为 63.7%和 66.1%，检测效果较差，证明只是简单的特征拼接很难实现对虚假信息的准确鉴别。当使用跨模态注意力机制后，模型的检测准确率提高了 15%以上，证明模态间注意力和模态内注意力可以有效地实现不同模态之间的交互，提高对跨模态虚假信息的检测效果。在此基础上，对不同模态特征进行对比学习预训练，并经过模间和模内注意力交互，得到的 MATN 模型对虚假信息鉴别的准确率在 87%以上，较单独使用跨模态注意力机制的模型准确率提高了 6.5%以上。证明通过对比学习预训练对不同模态特征实现特征对齐，可以明显提高对虚假信息的检测准确率。

4 结　语

对于网络平台虚假信息的检测任务，通过对比学习预训练对不同模态数据进行特征对齐后，再通过模态间和模态内的注意力机制，能够更好地实现不同模态之间的交互，尤其是通过对比学习对各模态数据进行预训练后，增强了弱相关正样本对之间的关联，极大地提高了模型对于虚假信息甄别的准确度，为虚假信息的检测任务提供了一种新的潜在解决方案。

（由于印刷关系，查阅本文电子版请登录：http://www.paper.edu.cn/journal/zgkjlw.shtml）

[参考文献] (References)

[1] 王剑，王玉翠，黄梦杰. 社交网络中的虚假信息：定义、检测及控制[J]. 计算机科学，2021，48(8)：263-277.
WANG J，WANG Y C，HUANG M J. False information in social networks：definition，detection and control [J]. Computer Science，2021，48(8)：263-277. (in Chinese)

[2] 黄皓，周丽华，黄亚群，等. 基于混合深度模型的虚假信息早期检测[J]. 山东大学学报(工学版)，2022，52(4)：89-98，109.
HUANG H，ZHOU L H，HUANG Y Q，et al. Early detection of fake news based on hybrid deep model [J]. Journal of Shandong University (Engineering Science)，2022，52(4)：89-98，109. (in Chinese)

[3] 王莉. 网络虚假信息检测技术研究与展望[J]. 太原理工大学学报，2022，53(3)：397-404.
WANG L. Development and prospect of false information detection on social media [J]. Journal of Taiyuan University of Technology，2022，53(3)：397-404. (in Chinese)

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

[5] BIAN T，XIAO X，XU T Y，et al. Rumor detection on social media with bi-directional graph convolutional networks [C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York：AAAI，2020：549-556.

[6] JIN Z W，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. New York：ACM，2017：795-816.

[7] 刘金硕，冯阔，PAN J Z，等. MSRD：多模态网络谣言检测方法[J]. 计算机研究与发展，2020，57(11)：2328-2336.
LIU J S，FENG K，PAN J Z，et al. MSRD：multi-modal web rumor detection method [J]. Journal of Computer Research and Development，2020，57(11)：2328-2336. (in Chinese)

[8] KHATTAR D，GOUD J S，GUPTA M，et al. MVAE：multimodal variational autoencoder for fake news detection [C]// WWW'19：the World Wide Web Conference. New York：ACM，2019：2915-2921.

[9] 孟杰，王莉，杨延杰，等. 基于多模态深度融合的虚假信息检测[J]. 计算机应用，2022，42(2)：419-425.
MENG J，WANG L，YANG Y J，et al. Multi-modal deep fusion for false information detection [J]. Journal of Computer Applications，2022，42(2)：419-425. (in Chinese)

[10] GAO P，JIANG Z K，YOU H X，et al. Dynamic fusion with intra- and inter-modality attention flow for visual question answering [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York：IEEE，2019：6632-6641.

[11] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision [EB/OL]. [2021-02-26]. https://arxiv.org/abs/2103.00020.

[12] LI L，YAN C G，CHEN X，et al. Distributed image understanding with semantic dictionary and semantic expansion [J]. Neurocomputing，2016，174：384-392.

[13] NAN Q，CAO J，ZHU Y C，et al. MDFEND：multi-domain fake news detection [C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York：ACM，2021：3343-3347.

[14] ZHANG H，KOH J Y，BALDRIDGE J，et al. Cross-modal contrastive learning for text-to-image generation [C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York：IEEE，2021：833-842.

[15] LI J N，SELVARAJU R R，GOTMARE A D，et al. Align before fuse：vision and language representation learning with momentum distillation [EB/OL]. [2021-10-07]. https://arxiv.org/abs/2107.07651.

[16] QI P，CAO J，LI X R，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. New York：ACM，2021：1212-1220.

[17] 张北辰，李亮，查正军，等. 基于跨模态对比学习的视觉问答主动学习方法[J]. 计算机学报，2022，45(8)：1730-1745.
ZHANG B C，LI L，ZHA Z J，et al. Contrastive cross-modal representation learning based active learning for visual question answer [J]. Chinese Journal of Computers，2022，45(8)，1730-1745. (in Chinese)

[18] WANG Y Q，MA F L，JIN Z W，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. New York：ACM，2018：849-857.

[19] ANTOL S，AGRAWAL A，LU J S，et al. VQA：visual question answering [C]// 2015 IEEE International Conference on Computer Vision (ICCV). New York：IEEE，2015：2425-2433.