DOI:10.19652/j.cnki.femt.2304724

基于双预训练 Transformer 和交叉注意力的多模态谣言检测*

蒋保洋[1,2] 但志平[1,2] 董方敏[1,2] 张洪志[1,2] 刘致远[1,2]
(1.三峡大学计算机与信息学院 宜昌 443002;2.三峡大学水电工程智能视觉监测湖北省重点实验室 宜昌 443002)

摘 要:社交平台上文本和图像相结合的多模态谣言比纯文本谣言更易于误导用户,因此研究多模态的谣言检测方法具有重要意义。现有方法大多只是对各个模态特征直接进行向量拼接,忽略了模态间联系,不能充分利用多模态信息。为了解决上述问题,提出了一种基于双预训练 Transformer 和交叉注意力机制的多模态谣言检测模型:首先使用预训练的 Transformer(BERT 和 ViT)分别提取文本单词和图像的特征,克服了训练样本小的局限性;然后使用交叉注意力机制将文本和视觉特征进行特征融合,充分地学习到两种模态间的潜在联系;最后将得到的多模态融合特征输入谣言检测模块进行分类。实验结果表明,该模型在 Twitter 和微博数据集上的检测性能均高于多模态基准模型,有效性和泛化性进一步提升。

关键词:多模态;谣言检测;注意力机制;深度学习
中图分类号:TP183 **文献标识码**:A **国家标准学科分类代码**:520.604

Multimodal rumor detection method based on dual pre-trained transformer and cross attention mechanism

Jiang Baoyang[1,2] Dan Zhiping[1,2] Dong Fangmin[1,2] Zhang Hongzhi[1,2] Liu Zhiyuan[1,2]
(1. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China;
2. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, China)

Abstract: Rumors with text and images are more misleading and harmful than text-only rumors. Therefore, multimodal rumor detection has become a new hot-spot issue. In addition, most of the existing methods simply concatenate unimodal features without considering inter-modality relationships. Therefore, this paper proposes a multimodal rumor detection method based on dual pre-trained transformers (BERT and ViT) to extract features of text words and images respectively, and then uses cross attention mechanism to fuse text and visual features, plus text semantic features extracted by text CNN. Finally, the obtained multimodal fusion features are input into the rumor detection module for multimodal rumor classification. The model uses the pre-trained model for feature extraction, which has been trained on large-scale datasets, and has better performance. This method considers the relationship between multimodes, which can more effectively fuse multimodal features and effectively improve the effectiveness of rumor detection. As the core of the model, cross attention mechanism dynamically adjusts the weight of words by combining the information of text and image modes. Experiments conducted on two public benchmark datasets (Twitter and Weibo) validate the performance of the proposed method for rumor detection.

Keywords: multimodal; rumor detection; attention; deep learning

收稿日期:2023-02-19
* **基金项目**:NSFC-新疆联合基金(U1703261)项目资助

0 引 言

近年来，随着通信技术和智能终端的迅猛发展，社交媒体上的帖子和评论数量突飞猛进。互联网的快速发展导致社交媒体上出现了大量的假新闻和谣言[1]。与纯文字的谣言相比，这些带有图像的谣言包含更丰富的信息，更容易引起人们的关注，因此大量多模态谣言的迅速传播，会在经济、政治、食品安全和公共安全等领域给社会带来更加巨大的危害[2]。因此，研究出可以及时检测包含文本和图像的多模态谣言方法至关重要。

常见的多模态谣言中有特定的文本内容和与之相关的图像。仅根据文本很难确定它们是否是谣言，但通过观察图像很容易发现这是谣言。从这些情况可以明显看出，在检测网络谣言时，利用图像、文字这些多模态数据信息是至关重要的。

基于同时关注谣言中文本以及图像这两种模态信息的多模态检测方法，在谣言检测中取得了比单模态检测方法更好的效果，越来越多的研究人员开始关注多模态谣言检测。Singhal 等[3]分别使用 XLNet[4]和预训练 VGG (visual geometry group)[5]模型进行文本和视觉特征提取。Khattar 等[6]提出了多模态变分自动编码器(multimodal variational autoencoder，MVAE)，以获得多媒体帖子的潜在多模态表示。然而大多数现有的多模态方法[3-7]只是分别从文本和图像中提取特征，并将文本和视觉特征简单拼接用于谣言检测，而未充分考虑各个模态间的关联性，导致这些方法未能充分利用帖子的多模态信息，影响了检测性能进一步提升。

为了解决上述问题，本文提出了一种基于双预训练 Transformer 和交叉注意力机制的多模态谣言检测方法。该方法首先采用文本卷积神经网络模型(text convolutional neural networks，Text-CNN)从文本信息中提取文本的语义特征，再采用双预训练 Transformer (bidirectional encoder rpresentation from transformers，BERT)和预训练的 ViT (vision transformer) 分别从文本和图像中提取特征，这不仅充分发挥了预训练模型能够克服训练样本小、泛化性更好的优势，而且双预训练模型中 Transformer 编码器的使用，也便于后期的多模态特征融合；然后引入交叉注意力机制，通过结合文本和图像模态的信息动态调整单词权重，将文本特征和视觉特征进行特征融合，最后将得到的多模态融合特征输入谣言检测模块，进行谣言分类。

本文采用 BERT 和 ViT 进行文本和图像的特征提取，不仅解决了训练样本不足的问题，而且收敛更快、效果更好。同时，双 Transformer 编码器的使用，也解决了使用不同预训练模型进行特征融合引起的特征向量不兼容的问题。引入交叉注意力方法挖掘文本特征和图像特征之间的潜在联系，能够更充分地融合语言和视觉特征进行谣言检测，从而解决现有模态特征简单拼接所导致的局限性。对两个公共基准数据集(Twitter 和微博)中的特殊字符、乱码和图像尺寸进行标准化预处理，在实验中更能充分体现各个模型性能的真实性。

1 相关工作

谣言检测作为自然语言处理(natural language processing，NLP)应用的一个分支，研究的任务与一些其他的 NLP 分类检测任务类似，如垃圾邮件检测[8-9]、假新闻检测[10-12]和讽刺信息检测[13-14]。与其他 NLP 分类检测不同的是，谣言检测的研究更关注于社交平台帖子信息的虚假性[15-16]，通过谣言检测方法检测那些由人伪造并且可验证为虚假的帖子内容并对检测到内容进行分类判断。其中可以被验证为虚假的帖子被标记为谣言，其他帖子被标记成非谣言。现有的谣言检测根据社交平台帖子内容包含的信息，分为单模态谣言检测和多模态谣言检测。

1.1 单模态谣言检测

单模态谣言检测侧重于从帖子的文本、图像或传播结构中提取单模态特征。Yang 等[17]使用手工设计的文本特征来检测谣言。Castillo 等[18]提出了一种利用标点符号和表情符号数量进行谣言检测的方法。由于手工构建文本特征浪费时间和精力，研究人员开始使用基于深度学习的方法。Liu 等[19]使用卷积神经网络(convolutional neural networks，CNN)从帖子文本中提取特征。Qi 等[20]提出了一种基于 CNN 的模型融合频域和空间域的视觉信息用于谣言检测。Ma 等[21]建立了一个动态时间序列模型，以从谣言生命周期的时间序列中获取传播结构特征。Wu 等[22]使用注意力机制和图神经网络(graph neural network，GNN)从传播结构中提取特征。虽然这些方法在谣言检测任务中取得了良好的效果，但随着社交平台多模态帖子的增加，这些单模态谣言检测方法暴露了其不能充分利用多模态信息的局限性。

1.2 单模态谣言检测

相比传统的纯文本的帖子，包含文本和图像的多模态帖子具有更丰富的信息，也更容易吸引读者的注意力。谣言帖子经常包含带有强烈情感的文字，同时配有引人注意的图片。这些多模态谣言更容易造成危害，而且很难被人工辨别。因此，研究人员开始利用多模态特征融合方法来检测谣言。Singhal 等[23]用 BERT 获取文本特征，并用预训练的 VGG-19 获取视觉特征，然后将文本和视觉特征进行拼接作为多模态融合特征。Khattar 等提出了用于谣言检测的 MVAE。他们使用双向长短期记忆网络(bidirectional long short-term memory，Bi-LSTM)提取文本特征，使用 VGG-19 提取图像特征。文本和视觉特征被拼接，然后输入到解码器进行谣言检测。Jin 等[24]提出了一种基于注意力的模型，该模型使用循环神经网络(recurrent neural network，RNN)和预训练的 VGG 从文本和图像中提取特征。它使用注意力机制来学习文本、图像和传播结

构之间的潜在表示。Wang 等[25]提出了一种用于谣言检测的事件对抗式神经网络(event adversarial neural networks,EANN)。这些多模态谣言检测方法均优于单模态的方法,且取得了不错的效果。

然而,目前多模态方法大多分别提取文本和视觉特征,然后通过直接拼接文本和视觉特征来简单地融合特征。这些研究把工作重点集中在特征提取阶段,而忽略了模态间关系的复杂性和相关性,这未能充分发挥多模态的优势。近些年,注意力机制在特征提取和特征融合等领域有着明显的效果[26-30],在多模态分类领域取得了不错的成果。因此本文借鉴了注意力机制的相关方法,克服这些多模态谣言检测方法的局限性,提出了一种基于双预训练 Transformer 和交叉注意力的多模态谣言检测方法。

2 本文模型

本文方法的核心思想是通过预训练 Transformer 更好地提取单一模态的特征,同时使用交叉注意力机制充分挖掘文本和视觉特征之间的相关性和潜在联系,从而进行更好地特征融合。

本文方法包括 3 个部分(图 1),分别为特征提取模块、多模态特征融合模块和谣言分类模块。首先,特征提取模块使用 Text-CNN 从文本信息中提取文本的语义特征,使用预训练的 BERT 从文本信息进行提取文本的词特征,使用预训练的 ViT 从图像中提取视觉特征。然后多模态特征融合模块使用交叉注意力机制将文本和视觉特征进行特征融合,得到多模态融合特征。最后,将得到的多模态融合特征输入谣言检测模块,进行多模态谣言分类。

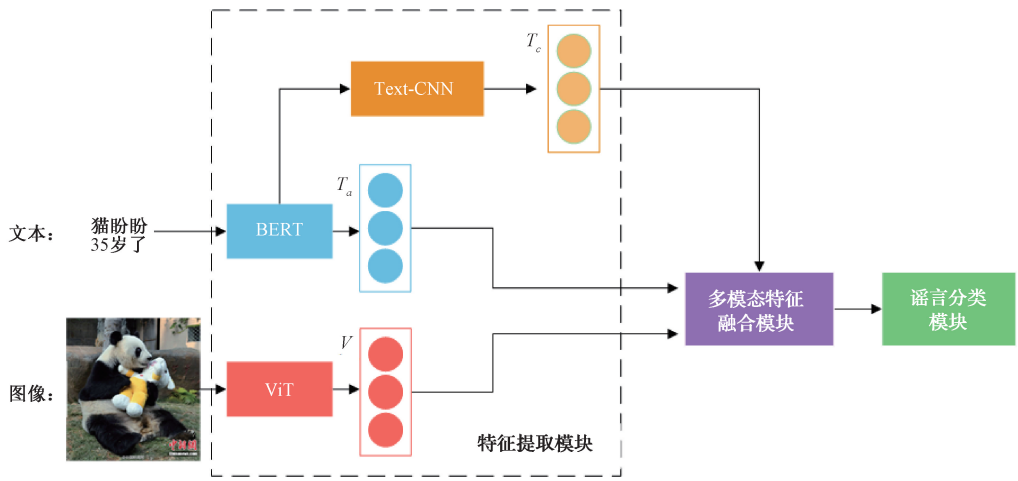


图 1 整体框架图

2.1 文本特征提取器

文本特征提取是谣言检测中的关键。为了从帖子中提取文本信息中的潜在信息和上下文特征。本文使用 BERT 模型来提取词向量,而不是使用 NLP 中常用的 Word2Vec[31]模型。BERT 是一种基于 Transformer 双向编码器的语言模型,其在大量无标注文本上进行了预训练,能够从文本中提取丰富的语义特征,在文本分类领域取得了良好的效果。BERT 本质是由多个 Transformer 中的编码器组成的双向编码器。BERT 和 ViT 都是基于 Transformer 编码器,BERT 提取的特征便于后续和 ViT 提取的特征进行多模态融合。

首先将输入文本转换为输入向量。其中,文本的第 i 个词表示为 $\boldsymbol{e}_i \in \boldsymbol{R}^k$,$k$ 为词向量的维度。因此,n 个单词的输入句子表示如下:

$$\boldsymbol{e}_{1:n} = [e_0, e_1, e_2, \cdots, e_n] \tag{1}$$

式中:n 表示句子中的单词数;e_0 是添加在每个输入示例前面的特殊符号 $[CLS]$。

将由 12 个编码器层组成的预训练 BERT-base 模型记为 f_{Bert}。如图 2 所示,将 $e_{0:n}$ 输入 f_{Bert} 后,得到给定句子的特征向量,计算如下:

$$T_o = f_{Bert}(e_{0:n}) \tag{2}$$

CNN 在图像和文本特征领域也有着强大的效果,因此除了使用 BERT 提取的词向量用于多模态特征融合,同时还使用 Text-CNN 来提取文本的语义特征,如图 3 所示。

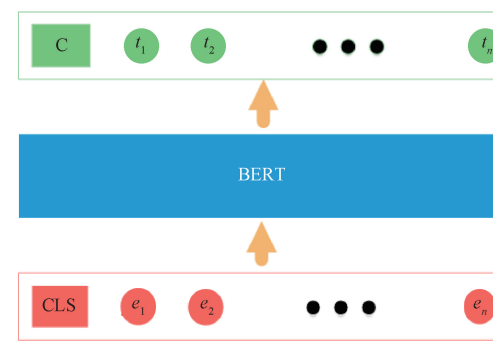


图 2 BERT 模型

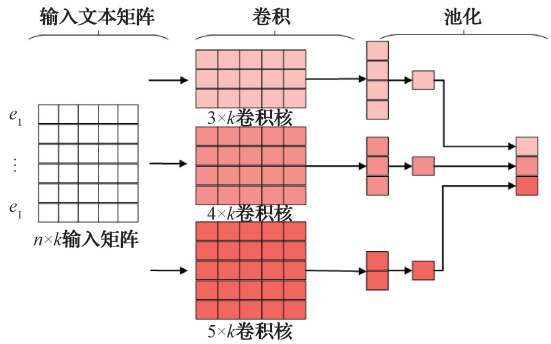


图 3 Text-CNN 模型

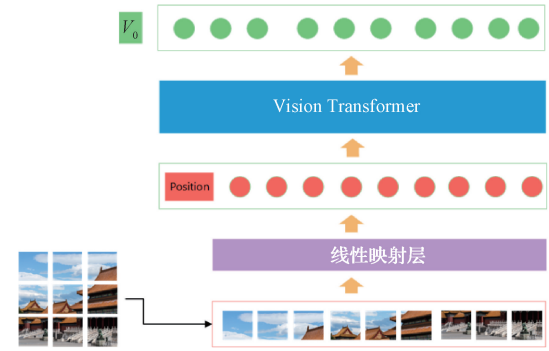


图 4 ViT 模型

将连续的单词序列表示为矩阵 $\boldsymbol{e}_{i:n}$。利用不同窗口大小的卷积核 $\boldsymbol{w}\in\boldsymbol{R}^{l\times k}$ 从矩阵中提取信息。卷积核对每个输入矩阵的不同位置的单词序列矩阵进行卷积运算，提取新的特征值。将一个帖子的第 i 个单词到第 $i+j-1$ 个单词的单词序列矩阵进行卷积，提取特征值如下：

$$v_i = f(\boldsymbol{w}\cdot\boldsymbol{e}_{i:i+l-1}+b) \tag{3}$$

式中：f 为双曲切线激活函数；b 为偏差项。该卷积核 w 可以应用于帖子中的每个单词序列，提取特征向量如下：

$$\boldsymbol{v} = [v_1, v_2, \cdots, v_{n-l+1}] \tag{4}$$

再将特征向量输入池化层，进行最大池化操作，从而选择最大值作为与该卷积核对应的最重要的特征值。

$$\hat{v} = \max\{\boldsymbol{v}\} \tag{5}$$

然后将从不同大小的卷积核中得到的特征合并，得到如下特征矩阵：

$$\boldsymbol{m} = [\hat{v};\hat{v}';\hat{v}''] \tag{6}$$

式中：$\hat{v}$、$\hat{v}'$、$\hat{v}''$ 分别为 3 个不同维度的卷积核所提取和汇集的特征矩阵。

2.2 图像特征提取器

图像信息对谣言检测任务起着至关重要的作用。大量研究表明，ViT 网络具有强大的图像特征提取能力[32-33]，在图像分类等计算机视觉领域问题上取得了令人瞩目的效果。同时，预训练模型在图像融合方面有着不错的效果[34]，本文采用已经在 ImageNet 数据集上预训练过的 ViT 网络作为图像提取模块。BERT 和 ViT 都是基于 Transformer 编码器，ViT 提取的特征便于后续和 BERT 提取的特征进行多模态融合。

对每一条帖子附带的图像进行缩放和标准化等预处理操作。预处理后的图像作为输入送到预训练过的 ViT 模型中进行图像信息提取。如图 4 所示，ViT 将输入图片分割成多个相同大小的图片块，再将每个图片块投影为固定长度的向量送入 Transformer 编码器，后续 Transformer 编码器的操作和原始 Transformer 中一致。

首先将视觉特征表示为 v，将预训练的 ViT-base 模型记为 f_{Vit}，则图像特征提取器中最后一层的操作可以表示为：

$$v_o = f_{Vit}(v) \tag{7}$$

2.3 多模态特征融合模块

多模态特征融合模块负责融合文本和图像特征来获得包含文本和图像特征的多模态特征。本文使用一种基于交叉注意力机制的方法来融合文本和图像特征，而不是简单地拼接本和图像的特征。

为了捕捉到文本序列和图像之间的特征相关性，一个直接的方法是通过获取每个文本单词和图片块的相似性得分。本文使用采取交叉注意力矩阵来表示这种相关性，文本单词和图片块关联性越强，相似性得分越高。

在得到文本特征和视觉特征后，为了使文本和视觉信息进行完全交互，本文将它们输入到多模态注意力融合模块中，通过结合不同模态中单词的表现来调整单词的权重。在得到多模态注意力融合模块的输出后，使用残差拼接来保持数据的原始尺寸，同时再与 Text-CNN 提取得到文本的语义特征进行拼接。最后，本文可以得到多模态特征的表示，使用它作为多模态融合特征并输入到全连接层中以产生最终的预测结果。多模态注意力融合作为模型的核心，旨在利用视觉模态的信息来帮助文本模态调整单词的权重，并微调预先训练好的 BERT 模型，从而通过充分发掘模态间的相关性和潜在联系来更好地利用多模态信息。多模态融合模块的结构如图 5 所示。

首先，本文评估了不同模态下每个单词的权重。文本模态的查询向量 $\boldsymbol{Q}_t$ 和键向量 $\boldsymbol{K}_t$ 被定义为：

$$\boldsymbol{Q}_t = \boldsymbol{K}_t = \boldsymbol{T}_o \tag{8}$$

式中：$\boldsymbol{T}_o$ 是文本特征提取模块中经 BERT 提取的单词特征矩阵。

视觉模态的查询向量 $\boldsymbol{Q}_t$ 和键向量 $\boldsymbol{K}_t$ 被定义为：

$$\boldsymbol{Q}_v = \boldsymbol{K}_v = \boldsymbol{V}_o \tag{9}$$

式中：$\boldsymbol{V}_o$ 是图像特征提取模块中经 ViT 提取的图像特征矩阵。然后将文本注意矩阵 $\boldsymbol{\alpha}$ 和视觉注意矩阵 $\boldsymbol{\beta}$ 定义为：

$$\boldsymbol{\alpha} = \partial(\boldsymbol{Q}_t\boldsymbol{K}_t^{\mathrm{T}}) \tag{10}$$

$$\boldsymbol{\beta} = \partial(\boldsymbol{Q}_v\boldsymbol{K}_v^{\mathrm{T}}) \tag{11}$$

为了通过文本与视觉模态的交互来调整每个单词的权重，本文将文本注意矩阵 $\boldsymbol{\alpha}$ 和视觉注意矩阵 $\boldsymbol{\beta}$ 求和加权，加权融合注意矩阵 $\boldsymbol{\gamma}$ 计算为：

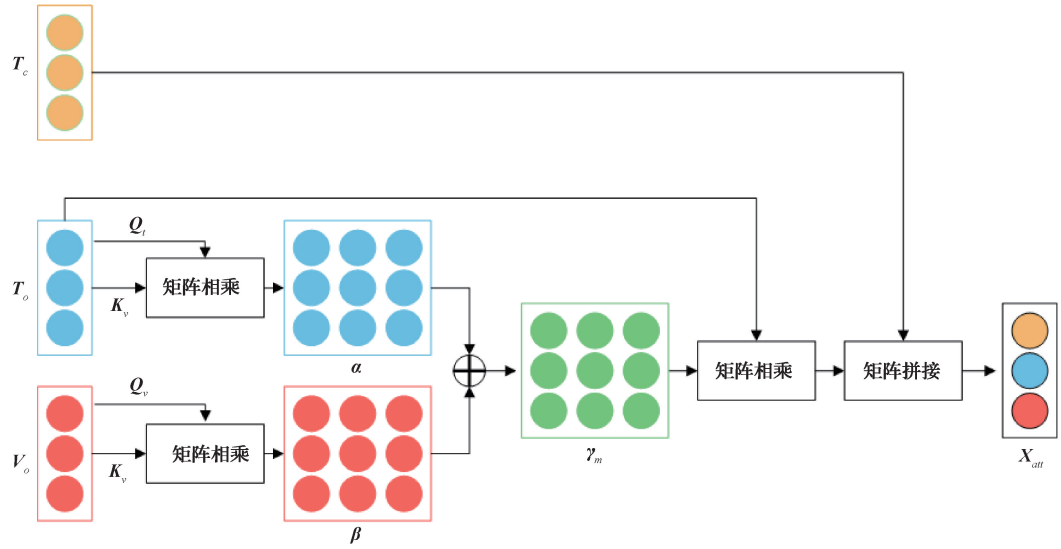


图 5　多模态特征融合模块

$$\boldsymbol{\gamma} = w_t\boldsymbol{\alpha} + w_v\boldsymbol{\beta} + b \tag{12}$$

式中：w_t 和 w_v 分别表示文本和视觉模态的权重；b 为偏置值。然后定义多模态注意矩阵 $\boldsymbol{\gamma}$ 为：

$$\boldsymbol{\gamma}_m = \delta(\boldsymbol{\gamma}) \tag{13}$$

式中：δ 为归一化指数函数。

在得到多模态注意矩阵 $\boldsymbol{\gamma}_m$ 后，本文将 $\boldsymbol{T}_o$ 与多模态注意矩阵 $\boldsymbol{\gamma}_m$ 的值相乘，得到多模态融合特征的输出：

$$\boldsymbol{X}_{att} = \boldsymbol{\gamma}_m\boldsymbol{T}_o \tag{14}$$

式中：$\boldsymbol{T}_o$ 是文本特征提取模块中经 BERT 提取的单词特征矩阵。

2.4　谣言分类模块

谣言分类模块使用多模态融合特征作为输入，通过一个全连接层和归一化层来将帖子分类为谣言或非谣言，公式如下：

$$\hat{y} = \delta(\boldsymbol{w}\boldsymbol{X}_{att} + b) \tag{15}$$

式中：$\boldsymbol{w}$ 为全连接层的权重矩阵；b 为偏置值；δ 为谣言分类模块中使用的归一化指数函数；$\hat{y}$ 为预测概率。本文把谣言标记为 1，非谣言标记为 0，同时采用交叉熵函数作为模型的目标函数，公式如下：

$$L = -\sum_{i=1}^{N} y_i\log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \tag{16}$$

3　实验相关设置

3.1　数据集

本文实验使用从 Twitter 和微博收集的两个公开多模态谣言数据集来评估本文提出的方法。表 1 为两个数据集的详细信息。

Twitter 数据集是来自 MediaEval[35] 的数据集，用于检测 Twitter 上的虚假内容。数据集分为两部分（训练集和测试集），本文保持不变，使用原始的数据集划分。此数据集的内容包含文本、关联图像和上下文信息。在本文的实验中，本文只关注图像和文本信息。

表 1　数据集统计

数据集	Twitter	微博
谣言	7 898	4 749
非谣言	6 026	4 779
总计	13 924	9 528

微博数据集来自 Jin 等[24] 的谣言检测方法。该数据集中的真实新闻来自新华社等权威媒体，数据集中的谣言由微博的官方辟谣系统正式确认。本文也只关注数据集中的图像和文本信息，忽略其他信息。

3.2　实验设置

在将数据送入网络训练前，需要对数据集的文本和图像进行预处理，具体处理方法如下。

Twitter 数据集中含有一些和谣言检测无关的特殊字符和乱码。本文删除了帖子中的特殊字符和乱码。关于图像的预处理，本文只对图像进行了缩放和标准化操作，使图像都具有相同的大小（224×224），方便送入模型。

对于微博数据集，本文使用 BERT 预训练模型自带的分词器对中文文本进行分词。与 Twitter 数据集的预处理类似，本文还删除了帖子的特殊字符和乱码，并对图像进行缩放和标准化。

实验在 CentOS 7.8 系统中进行，使用 Python 3.9.12 编程语言，深度学习框架为 pytorch 1.11.0。本文的文本词嵌入维度为 100，图片尺寸统一缩放为 224×224×3。

3.3　基线模型和评价指标

为了评估上述方法，本文选择准确率、精准率、召回率和 F1 分数作为评估指标，这些指标被广泛用于评估谣言

检测方法。

为了评估该模型的有效性，本文使用了几种单模态和多模态谣言检测方法进行了对比实验。在单模态方法中，本文选择常见的文本和图像模型进行比较。另外本文选取了主流的几种多模态基线方法和本文模型进行对比，这些方法均关注文本和关联图像两种信息。

VQA[36]，视觉问答模型旨在解决视觉语言理解和推理问题，稍作修改可用于检测多模态谣言。

EANN，事件对抗式神经网络旨在解决多模态谣言检测问题。EANN 由 3 部分组成：特征提取器、谣言检测器和事件鉴别器。

MVAE，多模态变分自动编码器旨在解决多模态谣言检测问题。它使用 Bi-LSTM 提取文本特征，使用 VGG-19 提取图像特征。

多模态融合网络（MFN）[37]，旨在解决多模态谣言检测问题。MFN 采用自关注融合机制进行特征级融合，采用潜在主题记忆网络存储语义信息。

基于注意力机制的多模态融合网络（AMFNN）[38]，在文本和图像模态层面进行高级信息交互，利用注意力机制获取和文本相关的视觉特征，同时使用自适应注意力机制来辅助约束内部信息流动。

4 实验结果与分析

4.1 对比实验

Twitter 和微博数据集上文本、视觉、VQA、EANN、MVAE、MFN、AMFNN 和本文模型的实验结果如表 2 所示。两个数据集上多模态方法的准确率和 F1 值的对比如图 6、7 所示。

表 2 实验结果

数据集	模型	准确率	谣言			非谣言		
			精确率	召回率	F1	精确率	召回率	F1
Twitter	Textual	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Visual	0.596	0.695	0.518	0.593	0.524	0.700	0.599
	VQA	0.631	0.765	0.509	0.611	0.550	0.794	0.650
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MFN	0.806	0.799	0.777	0.785	—	—	—
	AMFNN	0.841	0.820	0.660	0.730	—	—	—
	PTCA	**0.903**	**0.971**	**0.893**	**0.930**	**0.766**	**0.930**	**0.840**
微博	Textual	0.643	0.662	0.578	0.617	0.609	0.685	0.647
	Visual	0.608	0.610	0.605	0.607	0.607	0.611	0.609
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MFN	0.808	0.808	0.806	0.807	—	—	—
	AMFNN	0.862	0.860	0.850	0.850	—	—	—
	PTCA	**0.921**	**0.928**	**0.907**	**0.917**	**0.934**	**0.915**	**0.925**

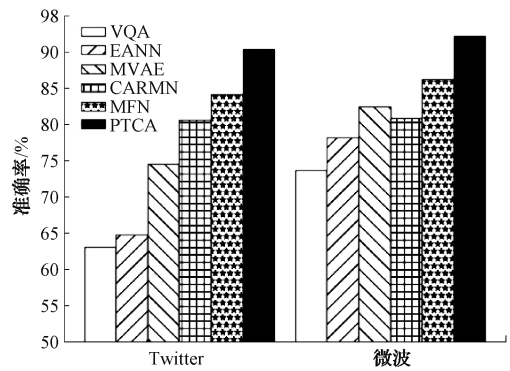


图 6 各模型准确率对比

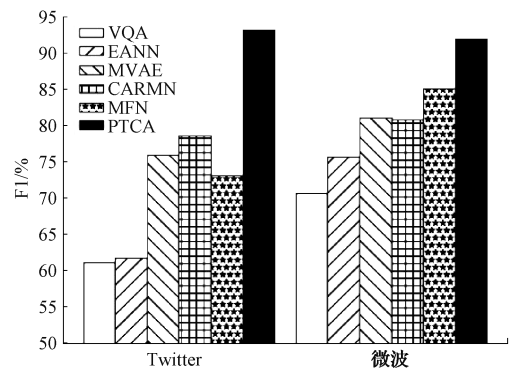


图 7 各模型 F1 对比

由表 2 可以观察到，多模态方法比单模态的单一图像或者单一文本方法具有更好的性能。可以得出结论，融合

图像和文本的多模态方法有利于谣言检测，因为图像联合文本可以包涵更丰富的信息。

由图 6 和 7 看出，MFN、AMFNN 和本文模型 PTCA 性能明显优于 VQA、EANN 和 MVAE。结果表明，本文提出的多模态特征融合方法优于简单的向量拼接方法。这说明 MFN、AMFNN 和本文模型 PTCA 的多模态融合机制可以更好地学习多模态信息的共享表示和潜在关系。

从图 6 和 7 还可以看出，本文模型 PTCA 在两个数据集的多个指标上优于所有基线模型。在 Twitter 数据集上，本文模型 PTCA 在准确率方面分别比 VQA、EANN、MVAE、MFN 和 AMFNN 高了 27.2%、25.5%、15.8%、9.7%和 6.2%。在微博数据集上，本文模型 PTCA 在准确率方面分别比 VQA、EANN、MVAE、MFN 和 AMFNN 高了 18.5%、13.9%、9.7%、11.3%和 5.9%。这些对比结果验证了本文模型 PTCA 方法的有效性，同时说明交叉注意力机制特征融合方法能更好地发掘不同模态间的潜在联系，从而有助于提升谣言检测模型的效果。

4.2 消融实验

为了验证图像信息和交叉注意力融合模块对本文模型所起到的作用，本文设计了消融实验来分析各个模块对模型效果的影响。实验结果如表 3 所示，表中“w/o image”为模型去除图像内容的实验结果，“w/o att”为模型去除交叉注意力模块，对图像和文本特征向量进行直接拼接的实验结果。

表 3 消融实验结果

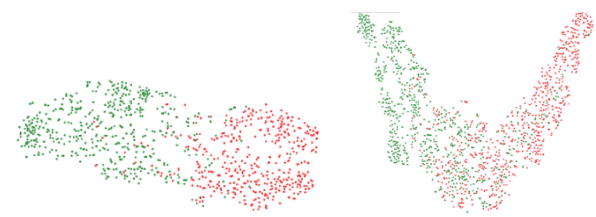
数据集	模型	准确率	精确率	召回率	F1
Twitter	w/o image	0.877	0.867	**0.963**	0.912
	w/o att	0.859	0.860	0.848	0.853
	PTCA	**0.903**	**0.971**	0.893	**0.930**
微博	w/o image	0.876	0.844	**0.911**	0.876
	w/o att	0.900	0.919	0.869	0.893
	PTCA	**0.921**	**0.928**	0.907	**0.917**

由消融实验结果可以看出，去除图像的模型“w/o image”实验结果明显较差，这说明多模态信息比单一模态含有更丰富的有助于谣言检测的重要信息，多模态的谣言检测方法相比较单一模态更能提高检测效果，多模态谣言检测研究具有重要意义；同时，对文本和图像特征用简单拼接方式进行特征融合的“w/o att”方法实验效果比使用了交叉注意力的模型更差。这说明对特征向量进行直接拼接的方法不能够充分利用多模态特征，同时验证了本文交叉注意力模块能够更好地发掘文本和图像特征的内在联系，更好地融合多模态特征，从而提高谣言检测模型的性能；最后，包含所有模块的完整模型谣言检测效果好，这充分验证了本文模型的有效性。

4.3 可视化分析

为了进一步探索分析交叉注意力特征融合模块在整体模型的有效性，本文以微博数据集为例，使用 t-SNE 算法将去除交叉注意力特征融合模块的“w/o att”学习到的多模态特征表示和完整模型学习到的多模态特征表示进行可视化对比，如图 8 所示。使用 t-SNE 算法可以将高维数据映射到二维空间，并且在二维坐标图上进行可视化。

图 8(a)和(b)分别为“w/o att”学习到的多模态特征表示和完整模型学习到的多模态特征表示。从图 8 可以明显看到，本文的完整模型学习到的多模态特征表示相较简单拼接特征的“w/o att”学习到的多模态特征表示效果更好。图 8(a)为“w/o att”学习到的多模态特征表示可以学习到样本的特征，但是不少特征被错误的分类。相反，在图 8(b)可以看到完整模型提取到的多模态特征表示样本点的分布间隔更加明显，聚类更加集中，这表明谣言分类效果更好。这进一步验证了交叉注意力特征融合模块在整体模型的有效性，相较于简单地对文本和图像特征进行拼接的方法，交叉注意力特征融合方法能够更大限度地学习到模态间的潜在联系，模型可以获得优良的多模态特征表示，从而使得谣言检测效果更好。



(a) “w/o att” (b) 完整模型

图 8 “w/o att”和完整模型可视化对比

4.4 超参数分析

为了探索本文模型中超参数对模型效果的影响，本文在 Twitter 和微博两个数据集上，设置了几组对比实验，对学习率进行迭代，最后观察实验结果。如图 9 所示，当学习率为 0.000 5 时，模型在 Twitter 和微博数据集上达到最优的效果，准确率分别达到了 90.3%和 92.1%。因此本文在实验中将模型学习率设置为 0.000 5，以获取最优的实验效果。

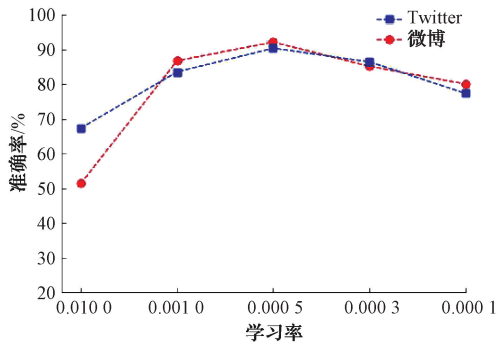


图 9 学习率对实验的影响

5 结 论

肆意传播的谣言会对社会的政治、医疗和经济等造成重大负面影响，而包含文本和图像的多模态谣言比纯文本谣言更具误导性和危害性。因此，研究一种有效的便快速可靠地检测错误信息的多模态谣言检测方法是至关重要的。本文提出了一种基于双预训练 Transformer 和交叉注意力机制的多模态谣言检测方法，该方法利用预训练模型能够克服训练样本小的局限性、泛化性更好的特性，采用预训练的 BERT 和预训练的 ViT 分别从文本和图像中提取特征。同时引入交叉注意力机制，通过结合文本和图像模态的信息来动态调整单词的权重，可以更有效地融合多模态数据，从而提高了谣言检测的性能。最后，本文模型在两个公共真实数据集上进行的实验验证了所提出模型用于谣言检测的有效性。

此外，本文的方法只关注了帖子的文本和图像，一些帖子具有更丰富的多模态信息（如传播结构、视频和音频等）。在未来的工作中，计划使用更多不同类型的信息进行多模态谣言检测。

参 考 文 献

[1] ZHANG X, GHORBANI A A. An overview of online fake news: Characterization, detection, and discussion[J]. Information Processing & Manage-ment, 2020,57(2): 102025.

[2] NAEEM S, BHATTI R, KHAN A. An exploration of how fake news is taking over social media and putting public health at risk[J]. Health Information & Libraries Journal, 2021, 38(2): 143-149.

[3] SINGHAL S, KABRA A, SHARMA M, et al. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(10): 13915-13916.

[4] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. 33rd Conference on Advances in Neural Information Processing Systems, 2019.

[5] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014, DOI: 10.48550/arXiv.1409.1556.

[6] KHATTAR D, GOUD J S, GUPTA M, et al. Mvae: Multimodal variational autoencoder for fake news detection[C]. World Wide Web Conference, 2019: 2915-2921.

[7] AZRI A, FAVRE C, HARBI N, et al. Rumor classification through a multimodal fusion framework and ensemble learning[J]. Computer Science, 2022,DOI: 10.48550/arXiv.2302.05289.

[8] SAEED R, RADY S, GHARIB T F. An ensemble approach for spam detection in Arabic opinion texts[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(1): 1407-1416.

[9] ASGHAR M Z, ULLAH A, AHMAD S, et al. Opinion spam detection framework using hybrid classification scheme[J]. Soft Computing, 2020, 24(5): 3475-3498.

[10] ZHU Y, SHENG Q, CAO J, et al. Memory-guided multi-view multi-domain fake news detection [J]. IEEE Transactions on Knowledge and Data Engineering, 2022,DOI: 10.1109/TKDE.2022.3185151.

[11] CUI J, KIM K, NA S H, et al. Meta-path-based fake news detection leveraging multi-level social context information[C]. Proceedings of the 31st ACM International Conference on Information & Knowledge Management,2022: 325-334.

[12] DAVOUDI M, MOOSAVI M R, SADREDDINI M H. DSS: A hybrid deep model for fake news detection using propagation tree and stance network[J]. Expert Systems with Applications, 2022, 198: 116635.

[13] DU Y, LI T, PATHAN M S, et al. An effective sarcasm detection approach based on sentimental context and individual expression habits[J]. Cognitive Computation, 2022, 14(1): 78-90.

[14] LI L, LEVI O, HOSSEINI P, et al. A multi-modal method for satire detection using textual and visual cues[J]. Computer Science, 2020, DOI: 10.48550/arXiv.2010.06671.

[15] HE Z, LI C, ZHOU F, et al. Rumor detection on social media with event augmentations[C]. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval,2021:2020-2024.

[16] WU Y, ZHAN P, ZHANG Y, et al. Multimodal fu-

sion with co-attention networks for fake news detection[C]. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021: 2560-2569.

[17] YANG F, LIU Y, YU X, et al. Automatic detection of rumor on sina weibo[C]. Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2012: 1-7.

[18] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on Twitter[C]. Proceedings of the 20th International Conference on World Wide Web, 2011: 675-684.

[19] LIU Z, WEI Z, ZHANG R. Rumor detection based on convolutional neural network[J]. Journal of Computer Applications, 2017, 37(11): 3053.

[20] QI P, CAO J, YANG T, et al. Exploiting multi-domain visual information for fake news detection[C]. 2019 IEEE international conference on data mining (ICDM). IEEE, 2019: 518-527.

[21] MA J, GAO W, WEI Z, et al. Detect rumors using time series of social context information on microblogging websites[C]. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015: 1751-1754.

[22] WU Z, PI D, CHEN J, et al. Rumor detection based on propagation graph neural network with attention mechanism[J]. Expert Systems with Applications, 2020, 158: 113595.

[23] SINGHAL S, SHAH R R, CHAKRABORTY T, et al. Spotfake: A multi-modal framework for fake news detection[C]. 2019 IEEE 5th International Conference on Multimedia Big Data (BigMM). IEEE, 2019: 39-47.

[24] 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.

[25] WANG Y, MA F, JIN Z, 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, 2018: 849-857.

[26] YANG K, XU H, GAO K. Cm-bert: Cross-modal bert for text-audio sentiment analysis[C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 521-528.

[27] 刘华玲,陈尚辉,乔梁, 等. 多模态混合注意力机制的虚假新闻检测研究[J/OL]. 计算机工程与应用:1-11 [2023-03-30]. http://kns.cnki.net/kcms/detail/11.2127.TP.20220818.1240.002.html.

[28] 周燕. 基于GloVe模型和注意力机制Bi-LSTM的文本分类方法[J]. 电子测量技术, 2022, 45(7): 42-47.

[29] 程德强,陈杰,寇旗旗,等. 融合层次特征和注意力机制的轻量化矿井图像超分辨率重建方法[J]. 仪器仪表学报, 2022, 43(8): 73-84.

[30] 梁继然,陈壮,董国军,等. 结合注意力机制和密集连接网络的车辆检测方法[J]电子测量与仪器学报, 2022, 36(3): 210-216.

[31] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013, DOI: 10.48550/arXiv.1301.3781.

[32] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]. Computer Vision and Pattern Recognition, 2020.

[33] LANCHANTIN J, WANG T, ORDONEZ V, et al. General multi-label image classification with transformers[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 16478-16488.

[34] 刘小利. 基于深度学习算法的图像融合[J]. 国外电子测量技, 2020, 39(7): 38-42.

[35] BOIDIDOU C, PAPADOPOULOS S, DANG-NGUYEN D, et al. Verifying multimedia use at MediaEval 2016[C]. MediaEval 2016 Workshop, 2016.

[36] ANTOL S, AGRAWAL A, LU J, et al. Vqa: Visual question answering[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 2425-2433.

[37] CHEN J, WU Z, YANG Z, et al. Multimodal fusion network with contrary latent topic memory for rumor detection[J]. IEEE MultiMedia, 2022, 29(1): 104-113.

[38] 戚力鑫,万书振,唐斌,等. 基于注意力机制的多模态融合谣言检测方法[J]. 计算机工程与应用, 2022, 58(19): 209-217.

作 者 简 介

蒋保洋,硕士研究生,主要研究方向为自然语言处理。E-mail:byjiang@ctgu.edu.cn

但志平(通信作者),博士,教授,硕士生导师,主要研究方向为自然语言处理、计算机视觉和模式识别。E-mail:zp_dan@ctgu.edu.cn

董方敏,博士,教授,博士生导师,主要研究方向为计算机图形图像处理和智能信息处理。