情报学报　2023 年 12 月　第 42 卷　第 12 期
Journal of the China Society for Scientific and Technical Information, Dec. 2023, 42(12): 1477-1486

DOI: 10.3772/j.issn.1000-0135.2023.12.007

基于多模态Transformer的虚假新闻检测研究

王震宇，朱学芳

（南京大学信息管理学院，南京　210023）

摘　要　为了减少虚假新闻给社会带来的负面影响，虚假新闻检测一直是自然语言处理中的一个重要领域。现有多模态虚假新闻检测方法通常使用预训练模型充当特征提取器，但是这些方法存在以下不足：①预训练模型参数在模型训练过程中总是会冻结，但预训练模型并不完美；②基于CNN（convolutional neural network）的图像特征提取器结构通常比基于Transformer的文本特征提取器结构更加复杂，图像特征通常被提前存储，使得这些模型的缺点被忽略。为此，本文提出基于端到端训练的多模态Transformer模型，通过使用视觉Transformer代替CNN提取图像特征，统一了不同模态的特征提取过程，利用共同注意力模块实现图像特征和文本特征交叉融合，并且在3个公开数据集上进行了对比实验。实验结果表明，本文模型性能超越了其他基线模型。

关键词　虚假新闻检测；多模态融合；Transformer；注意力机制

Research on Fake News Detection Based on Multimodal Transformer

Wang Zhenyu and Zhu Xuefang

(*School of Information Management, Nanjing University, Nanjing*　210023)

Abstract: Fake news detection has been an essential area in natural language processing to reduce the negative impact of misinformation on society. Most existing multimodal fake news detection methods use pre-trained models to act as feature extractors; however, these methods have the following shortcomings: (1) Pre-trained model parameters are typically frozen during model training. However, it is crucial to note that these pre-trained models are not flawless; (2) CNN-based image feature extractor structures are typically more complex than Transformer-based text feature extractor structures, and because image features are typically stored in advance, the shortcomings of these models are negligible. Therefore, this study proposes a multimodal end-to-end Transformer, unifies the feature extraction process for different modalities by extracting image features using a vision Transformer rather than a CNN, achieves cross-fusion of image features and text features using a co-attention module, and conducts comparative experiments on three public datasets. The experimental results show that the performance of the model proposed in this study outperforms other baseline models.

Keywords: fake news detection; multimodal fusion; Transformer; attention mechanism

0　引　言

近年来，移动智能设备的迅速发展改变了人们获取信息的方式，社交媒体已经成为人们浏览信息、表达和交流意见的主要渠道。随着微博、Twitter、抖音等应用程序的用户数量大幅上升，社交媒

收稿日期：2022-11-16；**修回日期**：2023-09-07

基金项目：国家社会科学基金项目“5G环境下中国智慧知识服务体系构建研究”（22BTQ017）。

作者简介：王震宇，男，1996年生，博士研究生，主要研究领域为自然语言处理；朱学芳，通信作者，男，1962年生，博士，教授，博士生导师，主要研究领域为数字信息资源管理及服务、多媒体信息处理、模式识别与人工智能，E-mail：xfzhu@nju.edu.cn。

体平台上存在的信息数据也呈现爆发式增长。虚假新闻[1]常被定义为“故意、可证实为虚假并可能误导读者的新闻文章”，由于用户不会验证其发布或转发信息的真实性，造成了大量虚假新闻的广泛传播。如果没有恰当的监管，那么这些虚假新闻很可能会误导更多的读者，甚至造成恶劣的社会影响，因此，针对社交媒体平台上的虚假新闻检测研究已经成为一个新的研究热点。

为了遏制虚假新闻的传播，社交媒体平台做了许多尝试。早期主要是通过邀请相关领域的专家或机构对用户发布的虚假新闻进行确认，这种方法费时、费力且无法应对指数式增长的信息数据，当前自动化虚假新闻检测方法受到了广泛关注。现有的自动化虚假新闻检测方法可以归纳为传统机器学习方法和深度学习方法两类。其中，传统机器学习方法包括支持向量机（support vector machine，SVM）[2]和决策树[3]等，这些方法依赖于从新闻相关信息中手工提取的特征。然而，虚假新闻的内容十分复杂，难以依靠手工提取到足够的有效特征，深度学习方法利用神经网络充当特征提取器，能够从原始数据中自动提取特征。例如，Ma 等[4]利用循环神经网络（recurrent neural network，RNN）从新闻的文本和社会背景中学习隐藏表示；Yu 等[5]利用卷积神经网络（convolutional neural network，CNN）从虚假新闻中提取关键特征和特征高阶交互关系。但是，上述方法通常只关注新闻中的文本信息，而忽略了其他模态的信息（如图像），这些信息对提升虚假新闻检测性能同样十分关键。图 1 是 Twitter 数据集中关于台风桑迪的虚假新闻示例，其中图片是经过处理的，结合图像信息和文本信息有助于对虚假新闻进行检测。

随着深度神经网络在各种非线性表示学习任务中不断取得突破，许多多模态表示学习任务也开始使用深度学习方法提取特征，其中包括多模态虚假新闻检测。Wang 等[6]提出了事件对抗神经网络（event adversarial neural network，EANN），该模型能够在提取虚假新闻中多模态特征的同时删除特定于某个事件的特征。Khattar 等[7]提出了多模态变分自编码器（multimodal variational auto encoder，MVAE）来提取新闻中的多模态特征，并将提取到的特征分别送入解码器和分类器中用于重建原始样本和虚假新闻检测。Singh 等[8]使用 NasNet Mobile 模型提取图像特征，同时使用 BERT（bidirectional encoder representations from transformers）和 ELECTRA（efficiently learning an encoder that classifies token replacements accurately）组合模型提取文本特征，大大减少了模型参数数量，提高了模型训练速度。虽然上述模型均在虚假新闻检测任务中表现出良好的性能，但是依然存在以下问题。

（1）现有的多模态虚假新闻检测方法主要使用预训练的深度卷积神经网络来提取图像特征，如 VGG16（visual geometry group 16）[9]、VGG19[6-7,10-11]、ResNet[12]。在实际训练过程中，充当图像特征提取器的预训练模型的参数会保持冻结，但是预训练模型并不完美，这会限制整个多模态模型的性能。为了减少特征提取时间，图像特征通常会被预先存储起来，往往会使得这些模型的缺点被忽略。

（2）由于不同模态数据之间可以相互补充，因此，处理好跨模态特征融合是多模态模型成功的关键。现有多模态虚假新闻检测方法使用的特征融合方式通常十分简单，例如，EANN[6]和 SpotFake[10]仅将图像特征和文本特征拼接在一起送入分类器中，没有充分考虑模态间的互补关系。

为了解决上述问题，本文提出了基于端到端训练的多模态 Transformer 模型（multimodal end-to-end transformer，MEET），训练过程中模型所有参数不会冻结。MEET 主要由多模态特征提取器和特征融合模块构成。通过使用视觉 Transformer 代替 CNN 提取图像特征，将对图像输入的处理简化为与处理文本输入一致的无卷积方式，统一了不同模态的特征提取过程。特征融合模块使用共同注意力（co-attention）模块[13]，其中文本特征和图像特征被分别输入两个对称的 Transformer 中，并使用交叉注意力机制来实现多模态交叉融合。此外，本文研究了端到端预训练对模型性能的影响，预训练数据集均为多模态数据集。通过在 3 个公开虚假新闻数据集

Text：*wow! #hurricane #Sandy over lady liberty*

Text：*It's getting real out there. Stay safe everyone, including Lady Liberty! #Sandy*

图 1 Twitter 虚假新闻示例

上的实验，证明了本文模型性能优于当前最好的方法。

本文的主要贡献如下：

（1）提出了 MEET 模型，使用视觉 Transfomer 作为图像特征提取器，以相同的方式处理不同模态的输入，同时采用端到端的方式对模型进行了训练。

（2）首次在虚假新闻检测任务中使用共同注意力模块，该模块已经成功应用于多个视觉语言任务中[14]，如图像问答、图像文本检索等。本文通过实例分析证明了共同注意力模块在虚假新闻检测中的有效性。

（3）第一次在虚假新闻检测任务中引入端到端预训练，并在 Twitter 数据集上与没有经过预训练的 MEET 模型进行了对比分析，实验结果验证了端到端预训练方法的优越性。

1　相关研究

1.1　虚假新闻检测

现有的虚假新闻检测方法可以大致分为传统方法和深度学习方法。早期研究者[15-16]主要使用由专家从新闻相关信息中手工提取的特征训练虚假新闻分类器，如用户特征、主题特征、传播特征等。虽然这些手动选择的特征被证实是有效的，但提取这些特征通常需要复杂的特征工程。与传统方法相比，深度学习方法能够从原始数据中自动提取特征，目前用于虚假新闻检测任务的深度学习方法可以分为单模态虚假新闻检测方法和多模态虚假新闻检测方法。

现有的单模态虚假新闻检测方法主要是从新闻文本中提取文本特征或从新闻图片中提取图像特征。Ma 等[4]使用循环神经网络从新闻中学习隐藏特征。Yu 等[5]使用卷积神经网络获取新闻的关键特征和高阶交互关系。Bahad 等[17]进一步研究了 CNN 和 RNN 在虚假新闻检测中的表现，使用新闻文本特征评估了双向长短期记忆（long short-term memory，LSTM）网络、CNN、RNN 和单向长短期记忆网络的性能。此外，Qi 等[18]提取了新闻图像不同像素域的视觉信息，并将其送入多域视觉神经网络来检测虚假新闻。

随着深度学习技术在特征提取和特征融合中的广泛应用，多模态虚假新闻检测方法受到越来越多的关注。Jin 等[19]首次在虚假新闻检测领域使用多模态模型，通过注意力机制融合了提取到的新闻图像、文本和社会背景特征。Khattar 等[7]提出了能够学习两种模态共享表示的 MVAE，模型被训练从学习到的共享表示中重建两种模态，并将学习到的共享用于虚假新闻检测。Singhal 等[10]使用预训练 BERT 模型提取新闻文本特征，同时使用预训练 VGG19 模型提取新闻图像特征。Qian 等[12]使用预训练模型学习新闻文本和图像表示，并将学习到的文本和图像表示输入多模态上下文注意网络以融合不同模态特征。

虽然预训练模型已经成功用于提取新闻的多模态特征，但是不同模态的特征提取器通常会使用不同的模型结构，如提取文本特征的 Transformer 结构、提取图像特征的 CNN 结构，CNN 结构比 Transformer 结构更加复杂，特征提取时间也更长。为了节省训练时间，实际训练时往往会选择冻结预训练模型参数，只训练整个模型的头部，而无法达到端到端的训练效果。

1.2　视觉 Transformer

尽管 Transformer 已经成为自然语言处理的主流架构[20]，但是直至最近才被用于图像处理[21-22]。为了将图像变为符合 Transformer 输入要求的序列形式，视觉 Transformer（vision transformer，ViT）将图片切分为大小相同的 patch 后组合成序列输入，patch 机制的引入极大简化了将图像嵌入形式转变为文本嵌入形式的过程。视觉 Transformer 已经在许多计算机视觉任务中取得了最先进的成果，如物体检测[23]、图像补全[24]、自动驾驶[25]等。本文提出的 MEET 模型是以视觉 Transformer 作为图像特征提取器的、完全基于 Transformer 的多模态模型。

2　模型设计

本文提出的基于多模态 Transformer 的虚假新闻检测模型结构如图 2 所示。整个模型主要由 3 个部分构成，分别是多模态特征提取器、共同注意力模块和虚假新闻检测器。多模态特征提取器负责提取新闻的文本特征和图像特征，之后，文本特征和图像特征会被送入共同注意力模块进行多模态特征融合，最后，融合特征会作为虚假新闻器的输入以生成最终的分类结果。

2.1　文本特征提取

文本特征提取器采用 Transformer 结构，Trans-

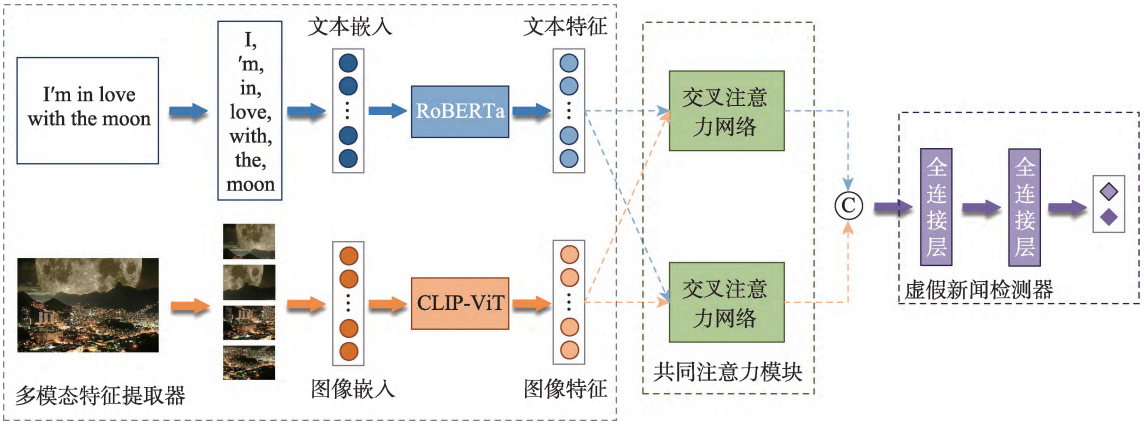

图2 基于多模态Transformer的虚假新闻检测模型

former在问答任务、命名实体识别、文本分类等多个领域均被证明是有效的[26-28]。为了提升模型的泛化性能，本文使用了在大规模语料上预训练的语言模型。目前常用的预训练语言模型有BERT[26]和RoBERTa[27]，两者均使用Transformer编码器作为网络主体。RoBERTa相较于BERT的主要改进在于修改了BERT原有的预训练策略，包括使用更大的文本嵌入词汇表、预训练任务中去除预测下一个句子和使用动态掩码策略等。RoBERTa已经在多个自然语言处理任务上表现出超越BERT的性能[27]。本文在第3节中对这两种文本编码器进行了对比分析。此外，为了证明预训练语言模型的必要性，本文还测试了只使用BERT的嵌入层作为文本编码器的情况。

令 $T=\{t_{[\mathrm{CLS}]},t_1,\cdots,t_m,t_{[\mathrm{SEP}]}\}$，其中 T 表示输入的文本嵌入；m 表示新闻文本中的单词数；$t_{[\mathrm{CLS}]}$ 为分类标记嵌入，表示该位置的特征向量用于分类任务；$t_{[\mathrm{SEP}]}$ 为分句标记嵌入，用于句子结尾。提取到的文本特征表示为 $L=\{l_{[\mathrm{CLS}]},l_1,\cdots,l_m,l_{[\mathrm{SEP}]}\}$，其中 l_i 对应于 t_i 转换后的特征；$l_{[\mathrm{CLS}]}$ 为分类标记的特征向量，代表文本的语义特征。L 的计算公式为

$$L=\{l_{[\mathrm{CLS}]},l_1,\cdots,l_m,l_{[\mathrm{SEP}]}\}=\mathrm{TextEncoder}(T) \quad (1)$$

其中，$l\in\mathbb{R}^{d_t}$ 为对应位置的输出层隐藏状态；d_t 为文本嵌入维数。

2.2 图像特征提取

为了使图像输入的三维矩阵结构变为符合Transformer输入要求的序列结构，首先要对图像进行序列化预处理，整个处理过程如图3所示。假设图像输入矩阵尺寸为224 × 224 × 3，使用卷积层将图像切分为14 × 14个patch，之后将所有patch展平成长度为196的序列，在序列前拼接分类标记嵌入

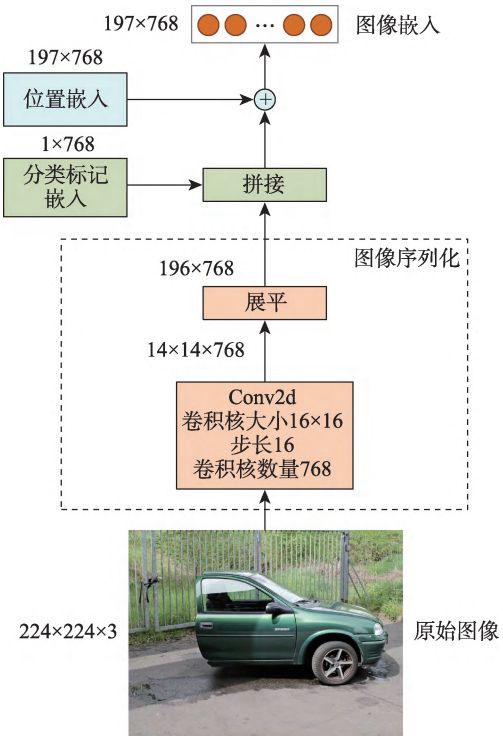

图3 图像预处理过程

再加上位置嵌入，就得到了完整的图像嵌入矩阵。本文图像特征提取器采用基于对比语言图像预训练（contrastive language-image pre-training，CLIP）的视觉Transformer模型[29]，以下简称CLIP-ViT。CLIP-ViT与其他预训练视觉Transfomer的区别在于其预训练数据是多模态的，是在从互联网上抓取的4亿个图像文本对上训练的。此外，CLIP-ViT在ImageNet分类等基准数据集上展现出强大的零样本学习能力。本文在第3节中通过对比实验深入研究了CLIP-ViT的作用。

对于给定图像嵌入 R，通过CLIP-ViT提取到的图像特征可以表示为

$$V = \{v_{\text{Class}}, v_1, \cdots, v_n\} = \text{CLIP-ViT}(R) \tag{2}$$

其中，$v \in \mathbb{R}^{d_r}$；v_{Class} 为分类标记的特征；d_r 为图像嵌入维数。

2.3　共同注意力模块

为了使模型能够学习到图像和文本之间的语义对应关系，本文使用共同注意力模块对图像特征和文本特征进行交叉融合。如图 4 所示，共同注意力模块由两个交叉注意力网络构成，每个交叉注意力网络都是一个 N 层的 Transformer 结构，与一般 Transformer 相比，每层多了一个交叉注意力块。通过在两个网络对应层的交叉注意力块之间交换键矩阵 K 和值矩阵 V，使得图像对应的文本特征能够被纳入网络输出的图像表示中，同样地，文本对应的图像特征也会被纳入网络输出的文本表示中。共同注意力模块已经被用于视觉语言模型中，并且在图像问答、图像标注等任务上证明了其有效性[13-14]。

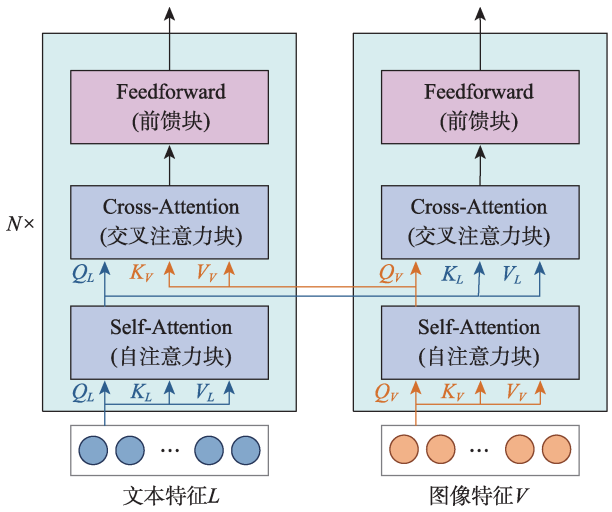

图 4　共同注意力模块

2.4　多模态融合及分类

通过共同注意力模块的交叉注意力机制，本文得到了更新后的图像特征 $W = \{w_0, w_1, \cdots, w_n\}$ 和文本特征 $S = \{s_0, s_1, \cdots, s_m\}$，其中 w_0 和 s_0 分别表示图像和文本的分类特征。将图像分类特征与文本语义分类进行拼接，得到多模态融合特征 C。

虚假新闻检测器以多模态融合特征 C 作为输入，利用两层全连接层来预测新闻是真假新闻的概率，计算公式为

$$H = \sigma_1(W_1 C + b_1) \tag{3}$$

$$P = \sigma_2(W_2 H + b_2) \tag{4}$$

其中，σ_1 为 gelu 激活函数；σ_2 为 softmax 激活函数；H 为第一层全连接层的输出；P 为最终输出的分类预测概率，模型损失函数为 P 与新闻标签真实值的交叉熵。

3　实证研究

3.1　数据集及评价指标

本文将提出的 MEET 模型与其他基线模型在 3 个公开的虚假新闻数据集上进行了比较，包括英文 Twitter 数据集[30]、中文 Weibo 数据集[19]以及中文 Bi-en 数据集①。

Twitter 数据集是在 MediaEval 研讨会上发布的虚假新闻检测数据集 MediaEval2015[30]，该数据集由 17000 条来自 Twitter 平台的推文文本及其相关图像组成，是多模态虚假新闻检测任务中最常用的数据集之一。遵照已有研究成果[7]，本文以没有重叠事件的方式将数据集划分为训练集（15000 条）和测试集（2000 条）。

Weibo 数据集由经过微博官方辟谣平台验证的虚假新闻和经新华社核实的真实新闻组成，这些新闻同样包含文本和图像[19]。使用不同语言的数据集能够更好地评估模型的泛用性和鲁棒性。按照已有方法[7]将该数据集划分为训练集（80%）和测试集（20%）。

Bien 数据集来自人工智能竞赛平台 BienData 举办的互联网虚假新闻检测挑战赛①，原始数据集分为两个部分：带标签的训练集（38471 条）和不带标签的测试集（4000 条）。与已有研究[31]一致，本文将原始训练集按照 4∶1 划分为训练集和测试集。

为提高数据质量，本文首先对 3 个数据集进行简单的预处理，筛选出既包含文本又包含图像的新闻，其中图像仅限静态图片，不包括动态图像和视频。经过筛选和处理后的数据集统计信息如表 1 所示。

表 1　3 个数据集的统计信息

类型	Twitter	Weibo	Bien
虚假新闻	6305	4583	10604
真实新闻	5598	4707	11008
图片	455	9288	18568

本文使用准确率（accuracy）作为模型主要评价指标，这是分类任务中的常用指标。此外，实验

① https://www.biendata.xyz/competition/falsenews/

中统计了模型的精确率（precision）、召回率（recall）和F1分数（F1-score）作为补充评价指标，可以减少类型不平衡时准确率指标可靠性下降的问题。

3.2 端到端预训练设置

本文对MEET模型进行了端到端的视觉语言预训练（vision-and-language pre-training，VLP）。预训练任务包括掩码语言建模（masked language modeling，MLM）和图像文本匹配（image-text matching，ITM）。在MLM任务中，将15%的输入文本替换为掩码标记（[MASK]），并让模型学习输出被替换的原始文本。在ITM任务中，按相同概率采样匹配和不匹配的图像标题对送入模型，模型需要输出输入的图像标题对是否匹配。

本文遵循已有研究[32]，在4个公开数据集上对模型进行端到端预训练，包括COCO数据集[33]、Conceptual Captions数据集[34]、SBU Captions数据集[35]和Visual Genome数据集[36]。为了使预训练数据集与微调数据集中的文本语言保持一致，本文只在英文Twitter数据集上对端到端预训练效果进行了验证。

3.3 实验设置及模型超参数

根据对3个公开数据集中文本长度的统计，将Twitter数据集文本序列最大长度设置为50，Weibo数据集和Bien数据集文本序列最大长度设置为200，超出部分截断，不足部分补零。对于图片，所有图片输入大小均被调整为224 × 224 × 3，训练过程中对图片应用随机的数据增强[37]以加强模型泛化性能，验证和测试过程中不使用数据增强。

本文所有实验均在内存为32G，显卡为NVIDIA RTX 3090的服务器上完成。本文使用的编程语言为python 3.8，使用的深度学习框架为pytorch-lightning 1.3.2、pytorch 1.7.1和transformers 4.6.0。

MEET模型的图像特征提取器和文本特征提取器均是12层Transformer结构，图像嵌入和文本嵌入维数均为768。共同注意模块中两个交叉注意力网络均为6层Transformer结构。虚假新闻检测器中两层全连接层的神经元个数分别为1536和2，激活函数分别为gelu和softmax，损失函数为交叉熵损失函数。

本文使用AdamW优化器，训练批次大小为256，为了减缓模型过拟合同时加速模型收敛，学习率在训练总步数的前10%中会从0线性递增到设置的学习率，之后再线性衰减到0。

3.4 文本和视觉编码器的对比分析

由于完全训练一个MEET模型耗时较长，本文先在较少训练轮数下探究了不同文本和视觉编码器的表现。实验分为两个阶段：首先，评估了在缺少视觉或文本模态时各种编码器的性能；其次，通过研究不同文本编码器与视觉编码器的组合，深入分析了视觉编码器的作用。为了保证实验的可靠性，每个实验在不同的随机数种子上执行5次，并采用测试集的平均准确率作为评价指标。实验中所有模型的训练轮数设定为10，底层和顶层学习率分别设定为1e-5和1e-4，底层包括文本编码器和视觉编码器，顶层包括共同注意力模块和虚假新闻检测器。

3.4.1 文本编码器对比

如表2所示，BERT和RoBERTa在各数据集上表现存在显著差异。在中文Weibo数据集和Bien数据集中，两者表现相当；但在英文Twitter数据集上，BERT明显优于RoBERTa。值得注意的是，仅基于文本的BERT和RoBERTa在两个中文数据集的测试集上的准确率已超过部分使用非Transformer结构文本编码器的多模态模型。此外，采用BERT嵌入层的模型在3个数据集的测试集上仅持续预测同一类别，这表明预训练文本编码器在仅基于文本的虚假新闻检测任务中是必要的。

表2 无视觉模态时的文本编码器对比

文本编码器	Twitter	Weibo	Bien
Emb.	—	—	—
BERT	**62.85**	**86.56**	95.81
RoBERTa	57.72	86.26	**96.09**

注：粗体表示该行模型在该列数据集或评估指标上取得了最佳性能。

3.4.2 视觉编码器对比

如表3所示，本文在图像方面比较了CLIP-ViT-16和CLIP-ViT-32两种模型，其中16和32表示模型的patch大小。实验结果表明，CLIP-ViT-16在所有数据集上性能更佳。在Twitter数据集上，视觉编码

表3 无文本模态时的视觉编码器对比

视觉编码器	Twitter	Weibo	Bien
CLIP-16	**73.37**	**80.11**	**93.97**
CLIP-32	70.35	79.46	92.39

注：粗体表示该行模型在该列数据集或评估指标上取得了最佳性能。

器明显优于文本编码器，这与各数据集中新闻图片数量有关。Twitter 数据集的图片数量远少于 Weibo 数据集和 Bien 数据集，因此，视觉编码器在 Twitter 数据集上需要学习的虚假新闻图像特征也较少。

3.4.3　文本和视觉编码器组合对比

如表 4 所示，所有多模态组合模型相较于单一模态模型的性能都有明显提升，这包括仅使用 BERT 嵌入层作为文本编码器的模型。在引入视觉模态后，各文本编码器之间的性能差距显著减小，但使用一个预训练的文本编码器仍具有重要意义。在视觉编码器方面，CLiP-ViT-16 和 CLiP-ViT-32 均表现出良好性能，尤其是 CLiP-ViT-16 在 Weibo 数据集和 Bien 数据集上分别达到了 89.70% 和 97.15% 的平均准确率，超越了现有最佳模型的表现。

表 4　文本和视觉编码器组合对比

视觉编码器	文本编码器	Twitter	Weibo	Bien
CLIP-16	Emb.	73.79	87.76	96.14
	BERT	**75.32**	89.65	96.83
	RoBERTa	74.94	**89.70**	**97.15**
CLIP-32	BERT	74.43	88.33	96.67
	RoBERTa	74.81	88.44	96.69

注：粗体表示该行模型在该列数据集或评估指标上取得了最佳性能。

3.4.4　学习率设置

本文深入探讨了不同学习率对模型性能的影响，并选用 CLiP-ViT-16 和 RoBERTa 作为默认编码器。表 5 展示了在 4 种学习率设置下训练的模型表现。对于 Weibo 数据集和 Bien 数据集，对模型顶层参数采用更高的学习率有助于取得更好的结果。然而，在 Twitter 数据集上模型性能却呈现下降趋势。这是由于 Twitter 数据集中视觉模态相对文本模态更为重要，多模态融合过程中视觉模态起主导作用，较高的学习率容易导致过拟合现象。相反地，在 Weibo 数据集和 Bien 数据集中文本模态和视觉模态的重要性相当，较高的学习率有助于学习到更复杂的多模态融合策略。

表 5　不同学习率设置对比

底层学习率	顶层学习率	Twitter	Weibo	Bien
1e-6	1e-6	76.46	86.23	92.09
5e-6	5e-6	**79.64**	88.18	96.99
1e-5	1e-5	76.59	88.60	95.93
5e-6	2.5e-5	76.72	**89.23**	**97.04**

注：粗体表示该行模型在该列数据集或评估指标上取得了最佳性能。

3.5　基线模型

3.5.1　单模态模型

（1）Textual：该模型仅使用新闻文本作为模型输入。使用预训练的词嵌入模型生成文本嵌入，将文本嵌入输入双向 LSTM 模型以提取新闻文本特征，最后使用全连接层输出分类结果。

（2）Visual：该模型仅使用新闻图片作为模型输入。先将图片输入预训练 VGG19 模型提取图像特征，再将图像特征输入全连接层进行虚假新闻检测。

3.5.2　多模态模型

（1）EANN[6]：EANN 主要由 3 个部分组成，即多模态特征提取器、虚假新闻检测器和事件鉴别器。在多模态特征提取器中，分别使用 TextCNN 模型和预训练 VGG19 模型提取文本特征和图像特征，将提取到的文本特征和图像特征拼接后输入虚假新闻检测器中。为了保证实验公平，本文使用的是不包含事件鉴别器的简化版 EANN。

（2）MVAE[7]：MVAE 使用双模态变分自编码器和二值分类器进行虚假新闻检测。其中，双模态变分自编码器使用双向 LSTM 模型和预训练 VGG19 模型作为文本编码器和图像编码器，利用全连接层进行特征融合。

（3）SpotFake[10]：SpotFake 使用预训练语言模型（BERT）提取文本特征，并使用预训练 VGG19 模型提取图像特征，没有使用特征融合方法。

（4）HMCAN（hierarchical multi-modal contextual attention network）[12]：HMCAN 使用预训练 BERT 模型和 ResNet 模型提取新闻文本特征和图像特征，并将提取到的特征输入多模态上下文注意网络进行特征融合，此外模型还使用层次编码网络捕捉输入文本的层次语义特征。

（5）CEMM（correlation extension multimodal）[31]：CEMM 先利用光学字符识别（optical character recognition，OCR）技术从附加图像中识别文本信息，再使用 BERT 和双向 LSTM 从新闻文章及其 OCR 文本中提取文本特征，并计算两者的相似性得分。最后将这些特征与图像直方图特征拼接后输入分类器以进行虚假新闻检测。

3.6　实验结果及分析

根据 3.4 节的实验结果，本文选择 RoBERTa 作

为MEET模型的文本编码器，CLIP-ViT-16作为视觉编码器。表6展示了MEET模型与其他基线模型在3个公开数据集上的性能对比。为了公平对比，本文在这3个数据集上训练了HMCAN模型，学习率设定为1e-3，保持其他训练设置与MEET模型一致。

如表6所示，在Twitter数据集上HMCAN模型的复现结果与原文存在较大差异，这可能是因为在划分Twitter数据集时，HMCAN没有确保训练集与测试集的事件无重叠。此外，从HMCAN的公开源码可知该模型并非端到端训练，这也可能对其性能产生影响。同时，HMCAN在处理BERT提取的文本特征时将其平均分成3段与图像特征进行分层特征融合，但在这一过程中模型没有充分考虑短文本的情况，导致后两段文本中存在大量无效的填充标记（padding token）。然而，源码中模型并未对这些填充标记进行掩码处理。对于Twitter数据集，本文仅关注HMCAN模型的复现结果。

表6 不同方法在3个数据集上的实验结果

数据集	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
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
	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
	SpotFake	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	HMCAN	0.897	0.971	0.801	0.878	0.853	0.979	0.912
	HMCAN*	0.774	0.772	0.863	0.815	0.777	0.651	0.708
	MEET（VLP）	**0.818**	0.786	**0.969**	**0.860**	**0.936**	0.611	0.740
	MEET	0.804	0.791	0.886	0.846	0.813	0.681	**0.741**
Weibo	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
	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
	SpotFake	0.892	0.902	**0.964**	**0.932**	0.847	0.656	0.739
	HMCAN	0.885	**0.920**	0.845	0.881	0.856	**0.926**	0.890
	HMCAN*	0.876	0.868	0.877	0.872	0.884	0.875	0.880
	MEET	**0.904**	0.914	0.886	0.899	**0.896**	0.922	**0.909**
Bien	Textual	0.874	0.900	0.843	0.871	0.850	0.905	0.877
	Visual	0.838	0.848	0.821	0.835	0.828	0.854	0.841
	EANN	0.950	0.967	0.935	0.949	0.935	0.968	0.952
	MVAE	0.881	0.880	0.885	0.882	0.882	0.877	0.880
	HMCAN*	0.964	0.965	0.964	0.965	0.961	0.963	0.962
	CEMM	0.964	0.974	0.954	0.964	0.955	0.974	0.964
	MEET	**0.983**	**0.986**	**0.981**	**0.983**	**0.980**	**0.985**	**0.982**

注：粗体表示该行模型在该列数据集或评估指标上取得了最佳性能，MEET(VLP)表示端到端预训练的MEET模型，*表示本文复现的结果。

实验结果显示，多模态模型相较于单模态模型具有显著优势。除了本文方法外，SpotFake模型和HMCAN模型的表现同样出色，这表明预训练的BERT模型能够更有效地从新闻文本中提取特征。本文提出的MEET模型在3个数据集上的准确率均超过其他基线模型，并在其他评价指标上也能取得最佳或次佳的成绩。MEET(VLP)模型在Twitter数据集上的表现尤为突出，进一步证实了端到端预训练能提升模型性能。图5是MEET(VLP)模型和MEET模型在训练过程中的损失曲线，可以看出，MEET(VLP)模型在前200步的训练损失下降速度更快，这表明端到端预训练不仅能提升模型性能，还可以加速模型收敛。

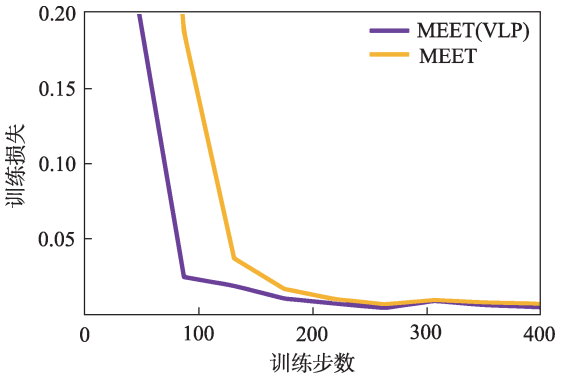

图5 不同预训练设置下的训练损失曲线

为了展示MEET模型在多模态融合上的优越性，本文对两个虚假新闻实例进行了注意力可视化分析。如图6所示，虽然从文本内容上看，这两则新闻似乎无法判断真伪，但图6a中窗外的海底景色和图6b中墙上的合影照片均显得异常可疑。在共同注意力模块的第一层注意力图中，模型最初将注意力分散在图片的各个区域。经过一层交叉注意力网络后，第二层注意力图显示模型能够在图像中检测到文本的部分语义对象，如“views”“bedroom”“toilets”。在多次交叉融合后，最终层注意力图体现出模型将注意力集中在两幅图中最不合理的部分，并成功判断出这些新闻为虚假信息。以上可视化分析结果可以证实，本文模型能够有效地利用多模态信息对虚假新闻进行检测。

4 总结与展望

针对现有多模态虚假新闻检测方法的不足，本

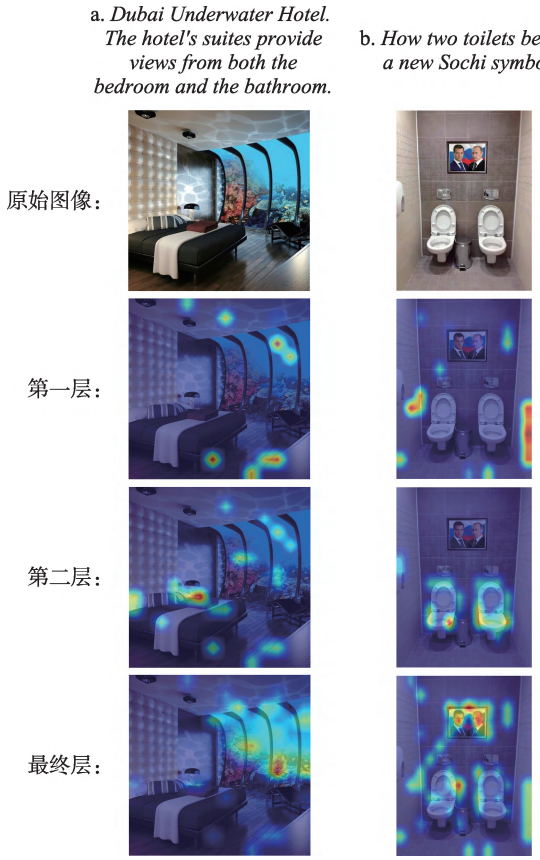

图6　MEET模型多模态融合注意力可视化

文提出了一种基于多模态 Transformer 的虚假新闻检测模型。首先，该模型将图像输入序列化成文本输入的形式；其次，利用预训练 Transformer 以相同的方式提取文本特征和图像特征；再其次，通过共同注意力模块实现不同模态间的交叉融合；最后，将融合后的图像特征和文本特征拼接起来送入全连接层生成检测结果。本文在 3 个公开数据集上进行了对比实验和实例分析，实验结果证明了该模型的优势和模型中每个模块的有效性。本文部分内容已用于申请发明专利[38]。

同时，本文尚存在以下不足。由于受到端到端预训练数据集的限制，本文只在 Twitter 数据集上验证了端到端预训练的效果，后续可以尝试在预训练数据集中添加中文数据集或多语言数据集，从而能够在更多不同语言的虚假新闻数据集上进行端到端预训练的实验。此外，本文模型只考虑了新闻的文本信息和图像信息，未来可以考虑引入更多模态以提升模型检测性能。

参　考　文　献

[1] Allcott H, Gentzkow M. Social media and fake news in the 2016 election[J]. Journal of Economic Perspectives, 2017, 31(2): 211-236.

[2] Castillo C, Mendoza M, Poblete B. Information credibility on twitter[C]// Proceedings of the 20th International Conference on World Wide Web. New York: ACM Press, 2011: 675-684.

[3] Kwon S, Cha M, Jung K, et al. Prominent features of rumor propagation in online social media[C]// Proceedings of the 2013 IEEE 13th International Conference on Data Mining. Piscataway: IEEE, 2013: 1103-1108.

[4] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 3818-3824.

[5] Yu F, Liu Q A, Wu S, et al. A convolutional approach for misinformation identification[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 3901-3907.

[6] 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 Press, 2018: 849-857.

[7] Khattar D, Goud J S, Gupta M, et al. MVAE: multimodal variational autoencoder for fake news detection[C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM Press, 2019: 2915-2921.

[8] Singh P, Srivastava R, Rana K P S, et al. SEMI-FND: stacked ensemble based multimodal inference for faster fake news detection [OL]. (2022-05-17) [2022-11-10]. https://arxiv. org/ftp/arxiv/papers/2205/2205.08159.pdf.

[9] 张国标, 李洁, 胡潇戈 . 基于多模态特征融合的社交媒体虚假新闻检测[J]. 情报科学, 2021, 39(10): 126-132.

[10] Singhal S, Shah R R, Chakraborty T, et al. SpotFake: a multimodal framework for fake news detection[C]// Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data. Piscataway: IEEE, 2019: 39-47.

[11] 王婕, 刘芸, 纪淑娟 . 基于矩阵分解双线性池化的多模态融合虚假新闻检测[J]. 计算机应用研究, 2022, 39(10): 2968-2973, 2978.

[12] Qian S S, Wang J G, Hu J, 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. New York: ACM Press, 2021: 153-162.

[13] Lu J S, Batra D, Parikh D, et al. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2019: 13-23.

[14] Hendricks L A, Mellor J, Schneider R, et al. Decoupling the role of data, attention, and losses in multimodal transformers[J].

Transactions of the Association for Computational Linguistics, 2021, 9: 570-585.

[15] Rashkin H, Choi E, Jang J Y, et al. Truth of varying shades: analyzing language in fake news and political fact-checking[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017: 2931-2937.

[16] Ma J, Gao W, Wei Z Y, 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. New York: ACM Press, 2015: 1751-1754.

[17] Bahad P, Saxena P, Kamal R. Fake news detection using bi-directional LSTM-recurrent neural network[J]. Procedia Computer Science, 2019, 165: 74-82.

[18] Qi P, Cao J, Yang T Y, et al. Exploiting multi-domain visual information for fake news detection[C]// Proceedings of the 2019 IEEE International Conference on Data Mining. Piscataway: IEEE, 2019: 518-527.

[19] 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 Press, 2017: 795-816.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2017: 6000-6010.

[21] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[OL]. (2020-06-03) [2022-11-10]. https://arxiv.org/pdf/2010.11929.pdf.

[22] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 9992-10002.

[23] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]// Proceedings of the 16th European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[24] Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C]// Proceedings of the 37th International Conference on Machine Learning. Cambridge: MIT Press, 2020: 1691-1703.

[25] Liu R J, Yuan Z J, Liu T, et al. End-to-end lane shape prediction with transformers[C]// Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2021: 3693-3701.

[26] Kenton J D M W C, Toutanova L K. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[27] Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach[OL]. (2019-07-26) [2022-11-10]. https://arxiv.org/pdf/1907.11692.pdf.

[28] Sun C, Qiu X P, Xu Y G, et al. How to fine-tune BERT for text classification?[C]// Proceedings of the 18th China National Conference on Chinese Computational Linguistics. Cham: Springer, 2019: 194-206.

[29] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning. Cambridge: MIT Press, 2021: 8748-8763.

[30] Boididou C, Andreadou K, Papadopoulos S, et al. Verifying multimedia use at MediaEval 2015[C]// Proceedings of the MediaEval 2015 Workshop. CEUR-WS.org, 2015: Paper 4.

[31] Li Y Q, Ji K, Ma K, et al. Fake news detection based on the correlation extension of multimodal information[C]// Proceedings of the 6th Aisa-Pacific Web and Web-Age Information Management Joint International Conference on Web and Big Data. Cham: Springer, 2023: 443-450.

[32] Kim W, Son B, Kim I. Vilt: vision-and-language transformer without convolution or region supervision[C]// Proceedings of the 38th International Conference on Machine Learning. Cambridge: MIT Press, 2021: 5583-5594.

[33] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]// Proceedings of the 13th European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[34] Sharma P, Ding N, Goodman S, et al. Conceptual Captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 2556-2565.

[35] Ordonez V, Kulkarni G, Berg T L. Im2Text: describing images using 1 million captioned photographs[C]// Proceedings of the 24th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2011: 1143-1151.

[36] Krishna R, Zhu Y K, Groth O, et al. Visual Genome: connecting language and vision using crowdsourced dense image annotations [J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[37] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: practical automated data augmentation with a reduced search space[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2020: 3008-3017.

[38] 朱学芳, 王震宇 . 基于多模态 Transformer 的虚假新闻检测方法: CN115982350A[P]. 2023-04-18.

（责任编辑 王克平）