研究论文

基于对抗神经网络的跨模态谣言检测*

孟佳娜　王晓培　李　婷　刘　爽　赵　迪
（大连民族大学计算机科学与工程学院 大连 116600）

摘要：【目的】通过结合文本和图像模数据，提出跨模态对抗神经网络模型，提高谣言检测对新数据的泛化能力。【方法】采用融合自注意力机制的双向长短时记忆网络模型表示文本特征，使用预训练的VGG19网络模型表示图像特征，通过对抗神经网络学习事件共同特征。【结果】所提模型在准确率、精确率、召回率和F1值得分等方面都优于对比模型，在微博、推特两个数据集上的准确率分别比基线模型的最优结果提高了3.6个百分点和3.5个百分点。【局限】不同模态信息下的特征关联分析不够，跨模态数据的语义鸿沟问题没有很好解决。【结论】所提模型能够比现有方法更好地学习特征表示，在谣言检测上取得了较好的结果。

关键词：谣言检测　对抗神经网络　双向长短时记忆网络　自注意力机制　VGG19
分类号：TP391
DOI：10.11925/infotech.2096-3467.2022.0064
引用本文：孟佳娜，王晓培，李婷等. 基于对抗神经网络的跨模态谣言检测[J]. 数据分析与知识发现，2022，6（12）：32-42. (Meng Jiana, Wang Xiaopei, Li Ting, et al. Cross-Modal Rumor Detection Based on Adversarial Neural Network[J]. Data Analysis and Knowledge Discovery, 2022, 6(12): 32-42.)

1　引　言

随着社交媒体的飞速发展，资讯的分享与获取达到了前所未有的便利，以微博、推特、微信为代表的社会媒体通过开放平台鼓励用户自己生产内容，并通过社交网络进行发布、分享、交流和传播[1]。这种基于社交媒体发布、分享多媒体内容的社交行为方式成为人们生活中不可或缺的一部分。然而，社交媒体平台对信息缺乏有效的监管也导致网络谣言的泛滥。根据新浪微博2020年发布的《微博辟谣2019年度报告》显示，三分之一的谣言始发于社交媒体。这些信息在未经处理的情况下可能被迅速地歪曲和放大，从而误导公众。

为有效地对谣言进行检测，人们提出了很多谣言检测方法，现有方法大多是基于特征的，主要检测语言是否具有煽动性，即通过检测是否含有大量标点、语气词、表情符号，是否具有诱导转发词、描述细节模糊等现象，而非谣言则是措辞中立、细节描写丰富、时间/机构等描述具体，从而对谣言进行判别。Castillo等[2]通过统计文本中的字数、标点符号、表情、超链接等出现频次，设计一种简单的推特信息真实性评估模型。之后，通过话题级别的特征反映出来的写作风格特征对谣言进行判别的方法也被相继提出[3]。Rashkin等[4]使用更复杂的语法信息与心理语言特征工具LIWC设计了多种语言特征，并结合长短时记忆网络（Long Short-Term Memory，LSTM）构建了虚假信息识别模型。视觉特征已经被证明是谣言检测的一个重要指标。从帖子所附图片中提取的特征已被证明有助于提供大量信息[5]。

随着谣言从单文本模态逐渐向图像和文本结合的形式转变，人们发现可以通过结合图像等多模态信息辅助判断，提升谣言检测性能。Guo等[6]观察到

通讯作者（Corresponding author）：刘爽（Liu Shuang），ORCID：0000-0002-0095-4328，E-mail：liushuang@dlnu.edu.cn。

*本文系国家自然科学基金项目（项目编号：61876031）和辽宁省自然科学基金计划项目（项目编号：2022-BS-104）的研究成果之一。
The work is supported by the National Natural Science Foundation of China (Grant No. 61876031), the Natural Science Foundation of Liaoning Province, China (Grant No. 2022-BS-104).

每篇推文都包含数据集中的视频/图像，而包含相同视频/图像的推文则是相对独立的，但彼此之间却具有很强的联系。换句话说，这些推文具有相同的主题，因此将这些推文聚集在一起并构建主题级别的分类器。也有谣言检测方法通过训练数据，尽可能获取谣言帖子中文本模态、视觉模态等多模态的特征表示[7-8]，再将这些特征迁移到新的帖子上进行谣言检测。通过以上分析发现，目前大多数谣言检测模型倾向于捕捉先验事件的特定特征，尽管多模态特征相比于单一模态特征谣言检测效果更好，但跨模态特征表示仍然高度依赖于数据集中的特定事件，导致这些特征难以进行迁移，并且会降低模型的泛化能力，导致无法识别新事件是否为谣言。

针对这些问题，本文提出了基于对抗神经网络的跨模态谣言检测模型，该模型由跨模态特征提取器、事件判别器和谣言检测器三部分组成。跨模态特征提取器中包括文本特征提取器和视觉特征提取器。在文本特征表示上，采用双向长短时记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）并融合自注意力机制模型。在图像特征表示上，选择经过 ImageNet 预训练的 VGG19 模型。事件判别器由两个具有相应激活功能的完全连接层组成，目的是将新出现的帖子正确分类为一个事件，其计算损失越大，表示不同事件分布越相似，学习的特征越具有事件不变性。谣言检测器使用 Softmax 函数，以预测帖子是否为谣言。模型利用对抗神经网络去除事件的特有特征，学习所有事件的共同特征，增强模型的泛化能力和特征的可迁移性，有利于预测新出现的事件是否为谣言，最终提高跨模态谣言检测任务的性能。

2　相关工作

关于谣言，个人可能有自己的直观定义，不同的定义可能会造成相互冲突或重叠。因此，本文遵循文献[9]对谣言的定义：谣言为真实价值未经证实或故意虚假的故事或陈述。根据网络谣言的传播内容划分，可将其归结为政治、经济、社会、灾难、文化五大类型。本文主要针对社交媒体上广泛传播或还未发生的所有谣言类型进行研究。

目前，谣言检测引起了国家、社会及科研机构等的高度重视。国内外研究者分别从谣言传播规律、内容特征等角度出发设计了大量谣言判别特征，同时设计谣言检测模型时，往往结合了机器学习分类方法及深度学习网络。

在基于传播规律的谣言检测上，研究者通过挖掘谣言传播与正常信息传播的异同，有效地辅助谣言检测。Gupta 等[5]构造了一个包含用户、微博消息和事件的可信度传播网络，将不同实体基于相似度连接在一起。基于半监督学习的模式采取了一种启发式的迭代算法对可信度的传播结果进行求解。宋之杰等[10]在谣言传播规则、人群节点划分以及节点感染概率等方面进行完善，构建突发事件谣言传播的 $SIHR_1R_2$ 模型。以上工作说明了基于传播规律的特征有助于提升检测的性能。

在基于内容特征的谣言检测上，以往基于单模态的谣言检测方法大部分是从文本角度上分析，在基于传统机器学习的方法中，Kwon 等[11]提出一种随机森林分类器，分别建立作用于时间、结构和语言特征的决策树、随机森林、支持向量机（Support Vector Machine，SVM）分类器。Tong 等[12]利用启发式规则和线性 SVM 在推特文本上进行谣言分类。由于传统基于特征的机器学习模型在学习深层潜在特征及各特征之间的相关性方面受到限制，所以研究者提出基于深度学习的谣言检测模型。Ma 等[13]通过抽取相关事件推文组成谣言事件，利用词嵌入和循环神经网络（Recurrent Neural Network，RNN）模型对谣言事件进行检测，证明了 RNN 模型在谣言检测中的有效性。刘政等[14]采用卷积神经网络（Convolutional Neural Network，CNN）检测微博谣言，模型包含一个卷积层和一个池化层，使用 Doc2Vec 训练向量矩阵，此模型与 RNN 模型进行比较，精确率提高了 10.2%，由此可见由于 CNN 模型通过发现微博谣言事件间的关系构造特征，比 RNN 模型更适合进行谣言检测。Chen 等[15]引入了注意力机制（Attentions Mechanism），选择性地学习帖子的暂时性隐藏表示以识别谣言。Popat 等[16]利用谣言言论语句和外部证据语句结合 BiLSTM 和注意力机制，设计了一种端到端的言论验证模型。Wang 等[17]提出了一种端到端框架，提取有利于检测突发事件假新闻的事件不变特征。Yang 等[18]提出了一种基

于图结构对抗学习的框架，利用文本和传播网络中的图结构从对抗的角度识别社交媒体上的谣言。Ni等[19]提出了一种分层对抗谣言检测模型，该模型从对抗的角度考虑了谣言的伪装和可变性，在后级和事件级嵌入向量上动态生成扰动，增强了模型的鲁棒性，并防止了伪装和重构谣言的欺骗。李奥等[20]提出一种改进的生成对抗神经网络模型用于谣言检测，通过对抗训练的方式强化谣言指示性特征的学习，有效提高了单文本模态谣言检测的效果。但是，仅仅从文本上获取谣言信息毕竟有限，通过结合图像、视频等多模态信息进行辅助判断，可以提升谣言检测性能。结果表明，跨模态谣言检测可以识别许多单一模态下无法判别的谣言。因此，受前人研究启发，本文提出了将BiLSTM、Attention、VGG19这些在单一模态表现优秀的方法结合，应用于跨模态谣言检测，弥补了使用单一模态对谣言检测的局限，同时采用对抗神经网络框架，去除特定事件的特征，进一步学习所有事件的共同特征，从而有效预测新出现的事件是否为谣言。

3 基于对抗神经网络的跨模态的谣言检测

3.1 跨模态谣言检测模型

本文提出基于对抗神经网络的跨模态的谣言检测模型，该模型由跨模态特征提取器、事件判别器和谣言检测器构成。模型首先分别对文本和图像进行预处理，然后分别提取特征表示。在文本特征提取方面，采用BiLSTM并融合自注意力机制模型。在图像特征表示上，利用预训练VGG19模型，文本特征提取和图像特征提取这里统称为跨模态特征抽取器，通过串联的方式将两类特征进行特征融合。事件判别器是一个神经网络，由两个具有相应激活功能的完全连接层组成，它通过使用单遍聚类的算法将新出现的媒体信息正确分类为某一类事件。事件的计算损失越大，表示不同事件分布越相似，学习的特征越是事件不变的特征。谣言检测器使用Softmax部署一个完全连接层，以预测媒体信息是否为谣言。对抗神经网络主要发生在跨模态特征提取器和事件判别器之间。在训练阶段，一方面，跨模态特征提取器试图欺骗事件判别器，使事件判别器损失最大化以获得事件共享特征表示；另一方面，事件判别器试图使自己损失最小化以识别每个事件，两者相互博弈，最终目的是获取事件的不变特征表示。同时，跨模态特征提取器需要与谣言检测器配合使检测损失达到最小化，三者通过一个损失函数共同达到一个权衡，最终提高谣言检测任务的性能。整体模型框架如图1所示。

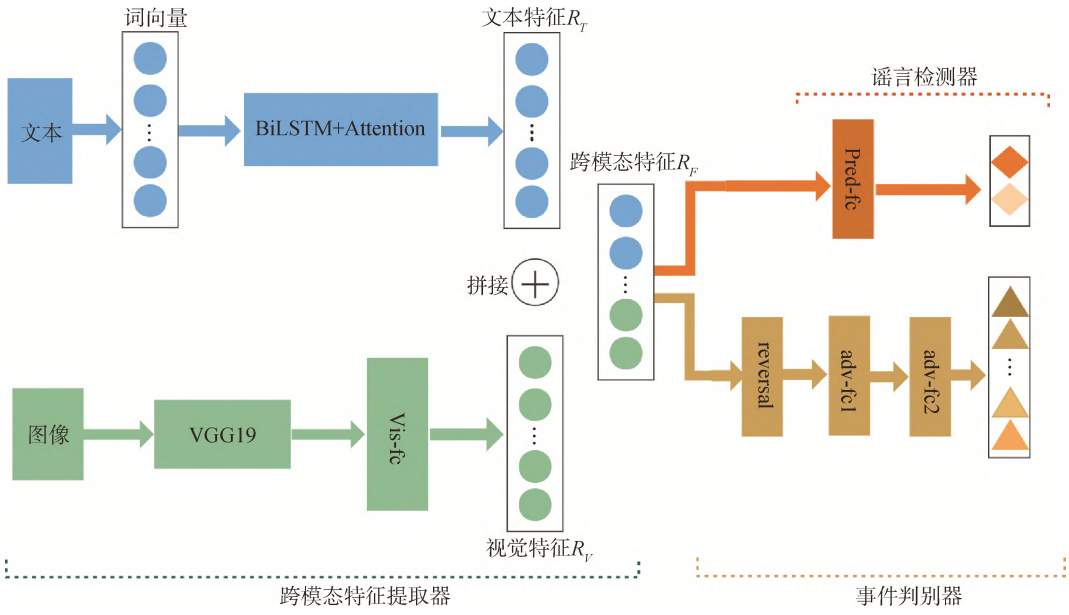

图1 整体模型框架

Fig.1 The Model Framework

3.2　对抗神经网络的模型集成方法

（1）跨模态特征提取器

跨模态特征提取器包括文本特征提取器和视觉特征提取器。其中，文本特征提取器采用融合注意力机制的 BiLSTM 模型，视觉特征提取器采用 ImageNet 预训练的 VGG19 模型提取视觉特征。

文本特征提取器中，即融合注意力机制的 BiLSTM 模型使用完全连接层确保最终的文本特征表示和视觉特征表示有相同的维度 p。全连接层后的结果如公式（1）所示。

$$R_T = \sigma_t(\boldsymbol{W}_{tf} \cdot R_t) \tag{1}$$

其中，R_t 表示融合注意力机制的 BiLSTM 模型提取到的特征表示；R_T 表示全连接层后的结果；$\boldsymbol{W}_{tf}$ 为文本特征提取器中完全连接层的权重矩阵。

帖子的附加图像是视觉特征提取器的输入，表示为 R_V。在 VGG19 模型的最后一层添加一个完全连接层调整最终视觉特征表示的尺寸。在与文本特征抽取器的联合训练过程中，预先训练的 VGG19 模型的参数保持不变，以避免过拟合。视觉特征表示为 $R_V \in R_P$，本文在文献[17]提出的公式上做出改进，特征提取器中最后一层的操作如公式（2）所示。

$$R_V = \sigma_v(\boldsymbol{W}_{vf} \cdot R_{VGG}) \tag{2}$$

其中，$\boldsymbol{W}_{vf}$ 为视觉特征提取器中完全连接层的权重矩阵；R_{VGG} 是从预训练 VGG19 获得的视觉特征表示。

将文本特征表示和视觉特征表示连接起来，形成跨模态特征表示，表示为 R_F，如公式（3）所示。

$$R_F = R_T \oplus R_V \tag{3}$$

R_F 是多模态特征提取器的输出。将多模态特征提取器表示为 $F(M;\ \theta_f)$，其中 M 通常是一组文本和视觉帖子，是多模态特征提取器的输入，θ_f 表示要学习的参数。

在 VGG19 模型的最后一层上，添加一个完全连接层调整最终视觉特征表示的尺寸。在与文本特征抽取器的联合训练过程中，预训练的 VGG19 模型的参数保持不变，以避免过拟合。

（2）谣言检测器

谣言检测器使用 Softmax 部署了一个完全连接层预测帖子是否为谣言。谣言检测器建立在跨模态谣言检测器特征提取器之上，获取跨模态特征表示 R_F 作为输入。将谣言检测器表示为 $D(;\ \theta_d)$，其中 θ_d 为所有包含的参数，第 i 个多媒体帖子的谣言检测器的输出为 m_i，表示该帖子是假帖子的概率。本文在文献[17]提出的公式上做出改进，如公式（4）所示。

$$p_\theta = D(F(m_i;\ \theta_f);\ \theta_d) \tag{4}$$

谣言检测器的目标是识别特定的帖子是否为谣言，使用 Y 表示标签集，并使用交叉熵计算检测损失，本文在文献[21]提出的公式上做出改进，如公式（5）所示。

$$L_d(\theta_f, \theta_d) = -E_{(y \in Y)}[\, y \log p_\theta + (1 - y)(\log(1 - p_\theta))\,] \tag{5}$$

通过寻求最佳参数 θ_f 和 θ_d 让损失函数 L_d 最小化检测，本文遵循文献[17]的优化方法，如公式（6）所示。

$$(\hat{\theta}_f, \hat{\theta}_d) = \arg\min_{\theta_f, \theta_d} L_d \tag{6}$$

（3）事件判别器

事件判别器是一个神经网络，由两个完全连接层组成，具有相应的激活函数。其目的是基于跨模态特征表示将帖子正确分类为 k 个事件之一，用 Y_e 表示事件标签集。将事件鉴别器表示为 $C(R_F;\ \theta_c)$，其中 θ_c 表示其参数。通过交叉熵定义事件鉴别器的损失，并使用它表示事件标签集，本文在文献[21]提出的公式上做出改进，如公式（7）所示。

$$L_c(\theta_f, \theta_c) = -E_{(m, y_e) \in (M, Y_e)}\left[\sum_{k=1}^{k} y_e \log(C(F(m;\ \theta_f));\ \theta_c)\right] \tag{7}$$

最小化损失 L_c 的事件鉴别器的参数遵循文献[17]的优化方法，如公式（8）所示。

$$\hat{\theta}_c = \arg\min_{\theta_c} L_c \tag{8}$$

上述损失 $L_c(\theta_f, \hat{\theta}_c)$ 可以用来估计不同事件分布的差异，为了消除每个事件的唯一性，需要通过寻找最佳参数 θ_f 最大化辨别损失。

该对抗神经网络促进了跨模态特征提取器和事件鉴别器之间的极小极大博弈。一方面，跨模态特征提取器试图欺骗事件鉴别器以使鉴别损失最大化；另一方面，事件鉴别器则企图通过特征表示中事件的特定信息识别事件。

（4）模型集成

在训练阶段，跨模态特征提取器 $F(\cdot;\ \theta_f)$ 需要

与谣言检测器$D(\cdot;\theta_d)$配合，最大限度地减小检测损失L_d，从而提高谣言检测性能。同时，跨模态特征提取器$F(\cdot;\theta_f)$试图欺骗事件鉴别器$C(R_F;\theta_c)$通过最大化事件鉴别损失L_c实现事件不变表示。事件判别器$C(R_F;\theta_c)$试图通过最小化事件鉴别损失识别每个事件。受文献[17]的启发，本文将三方博弈的最终损失进行改进，如公式(9)所示。

$$L_{final}=L_d-\lambda L_c \tag{9}$$

其中，系数λ控制谣言检测和事件判别的目标函数之间的权衡。只需将λ设置为1，无需调整折衷参数。使用随机梯度下降解决上述问题，这里采用梯度反转层(Gradient Reversal Layer，GRL)。梯度反转层在前向阶段充当一个恒等式函数，它将梯度乘以-λ，并在反向阶段将结果传递给前一层。GRL可以很容易地添加到跨模态特征提取器和事件判别器之间，将其表示反转层，根据文献[21]提出的公式，按照本文方法进行调整后的参数优化过程如公式(10)-公式(12)所示，其中μ为可以随时间变化的学习率。

$$\theta_f\leftarrow\theta_f-\mu\left(\frac{\partial L_d}{\partial\theta_f}-\lambda\frac{\partial L_c}{\partial\theta_f}\right) \tag{10}$$

$$\theta_c\leftarrow\theta_c-\mu\frac{\partial L_c}{\partial\theta_c} \tag{11}$$

$$\theta_d\leftarrow\theta_d-\mu\frac{\partial L_d}{\partial\theta_d} \tag{12}$$

为了稳定训练过程，本文遵循文献[22]方法衰减学习速率η，如公式(13)所示。

$$\eta'=\frac{\eta}{(1+\alpha\cdot p)^{\beta}} \tag{13}$$

其中，$\alpha=10$，$\beta=0.75$，p在训练过程中在0~1线性变化。

4 实验结果与分析

4.1 数据来源

实验使用微博和推特数据集。微博数据集[9]中的真实谣言是从中国权威谣言来源(如新华社)收集的。伪造的谣言爬取自2012年5月到2016年1月，并通过官方的谣言揭穿系统进行验证，该系统也是收集谣言消息的权威来源。预处理包括删除重复图像和低质量图像，以确保整个数据集的均匀性。推特数据集来自MediaEval 2016[23]，包含文本内容、附加的图像/视频和其他社交环境信息。实验中删除了没有任何文本或图像的推文。这两个数据集的详细统计信息如表1所示。

表1 数据集统计信息

Table 1 Statistical Information of Data

数据集	标签	数量	总计
微博	真	4 749	9 528
	假	4 779	
推特	真	7 021	12 995
	假	5 974	

4.2 实验设置

文本模态上，采用融合注意力机制的BiLSTM模型在微博数据集和推特数据集上进行实验。模型参数设置：单词嵌入的维度k设置为32，文本提取器中完全连接层的隐藏大小为64，BiLSTM的单元状态维度为128，激活函数为ReLU，批处理(batch size)大小为100，学习率为0.001，Dropout参数为0.500。视觉特征提取器采用224×224×3大小的图像作为输入，为了有效地提取视觉特征，视觉特征提取器前面部分采用和VGG19模型相同的结构。在VGG19模型的最后一层，添加了一个完全连接层调整最终视觉特征表示的尺寸。为了节约时间和资源成本，提高训练的效率，本文采用ImageNet数据集预训练好的VGG19模型，在与文本特征提取器的联合训练过程中，预先训练的VGG19模型的参数保持不变，以避免过拟合。在训练阶段使用相同的128个实例的批量，并且训练轮数为100。

本文采用的评价指标有准确率、精确率、召回率和F1值。基于谣言分类器得到的最终结果计算模型的评价指标，预测结果有4种可能性：谣言被辨别为谣言、谣言被辨别为非谣言、事实被辨别为谣言、事实被辨别为非谣言。

4.3 基于单文本模态的实验结果

在文本模态的谣言检测方法上，采用融合自注意力机制的双向长短时记忆网络(BiLSTM)。利用BiLSTM提取文本上下文关系特征，并通过自注意力机制动态调整特征权重。该模型解决了复杂的维度信息爆炸和原始循环神经网络存在的特征梯度消失或梯度弥散的问题，也有效地避免了一般的传统

模型对重点词在上下文的语义和语法上提取信息的缺失问题。同时，与 TextCNN 模型在相同数据集上进行对比研究。在微博、推特两个数据集上的实验结果如表 2 所示。

表 2　单文本模态的谣言检测结果

Table 2　Single Text Mode of Rumour Detection Results

数据集	模型	准确率	F1 值	精确率	召回率
微博	TextCNN	0.764	0.748	0.827	0.683
	BiLSTM-Attention	0.787	0.763	0.851	0.692
推特	TextCNN	0.532	0.568	0.598	0.541
	BiLSTM-Attention	0.585	0.629	0.642	0.618

实验中对比的深度神经网络模型如下。

（1）Text-CNN[24]。Text-CNN 模型通过 CNN 获取文本的局部信息，通过多个不同窗口大小的一维卷积核获取 *n*-gram 的局部文本特征。最后将多个卷积特征进行拼接，送入多层感知机进行分类。

（2）BiLSTM-Attention。BiLSTM-Attention 模型是融合自注意力机制的双向长短时记忆网络，在 BiLSTM 模型的基础上进行改进，通过 BiLSTM 模型提取文本上下文关系特征，再通过自注意力机制动态调整特征权重，最后送入多层感知机进行分类。通过自注意力机制，可以让模型更加关注重要程度较高的语义特征。

从数据集上看，由于推特数据集中不同事件的推文数量不平衡，超过 70% 的推文与单个事件相关，导致所学的文本特征主要集中在一些特定的事件上。这意味着文本模态包含更明显的事件特定特征，严重阻碍了文本模型提取不同事件之间的共享特征，因此文本模态的准确性整体上偏低。本文提出的 BiLSTM-Attention 模型和 TextCNN 模型相比，在微博数据集和推特数据集上的 F1 值分别提高 1.5 个百分点和 6.1 个百分点，准确率分别提高 2.3 个百分点和 5.3 个百分点。TextCNN 模型虽然可通过卷积核提取文本的局部特征，却没有区分重要信息和一般信息，而利用 Attention 机制能够捕获文本中重要信息，因此 BiLSTM-Attention 模型实验效果要好于 TextCNN 模型。

综上所述，数据集中不同事件越多，精确率越高；不同事件越少，精确率越低，这也说明了本文模型非常依赖数据多样性的特点。同时也可以看出 BiLSTM-Attention 模型效果优于 TextCNN 模型。

4.4　基于单视觉模态的实验结果

在视觉模态上，采用 VGG19 模型作为视觉特征提取器，利用 ImageNet 预训练的 VGG19 模型获取图像的全局特征，并在本文数据集上进行微调，然后将特征送入多层感知机进行分类。为了更好地体现 VGG19 模型在谣言检测方面的有效性，采用与 ImageNet 预训练的 ResNet152 模型和 RCNN[25]（Region-CNN）模型进行对比实验。

（1）ResNet152。利用 ImageNet 预训练的 ResNet152 模型获取图像的全局特征，ResNet152 模型具有较深的网络深度，其深度达到 152 层。获取特征后，需要在本文数据集上进行微调，然后将特征送入多层感知机进行分类。

（2）RCNN。RCNN 是指将 CNN 方法应用到目标检测中，即采用 Selective Search 方法选取候选框，并将每个候选框送入 CNN 以抽取特征，再将候选框特征进行平均池化，最后送入多层感知机进行分类。

实验结果如表 3 所示。可以看到，在微博数据集上实验得到的准确率、F1 值都远高于推特数据集，同样是因为推特数据集不同事件数量不平衡导致的结果。在推特数据集中，和基于单文本模态方法相比，基于单视觉模态的性能稍微高一些，是因为图像特征相比文本特征更易传递，从而降低了不同事件数量不平衡帖子的影响。在微博数据集中，因为数据足够多，导致可以提取有用的文本特征进行谣言检测，而对于图像，微博图像比推特图像更加复杂，导致视觉模态特征学习相比文本模态特征学习性能较差。从模型上看，VGG19 模型在两个数据集上的 F1 值比 ResNet152 模型分别提高 2.3 个百分点和 3.5 个百分点，证明了在视觉特征提取上 VGG19 模型效果更好，经过反复实验也表明 VGG19 模型在谣言分类任务的图片数据集上比 ResNet 模型的表现更为稳定。同时，VGG19 模型在两个数据集上的 F1 值比 RCNN 模型分别提高 0.8 个百分点和 2.3 个百分点，也证明了 VGG19 模型的实验效果优于 RCNN 模型。综上所述，VGG19 模型的实验效果优于 ResNet152 模型和 RCNN 模型。

4.5　跨模态谣言检测实验结果

本文提出的跨模态谣言检测模型采用融合注意

表3　单视觉模态的谣言检测结果

Table 3　Rumour Detection Results of Single Visual Mode

数据集	模型	准确率	F1值	精确率	召回率
微博	ResNet152	0.690	0.675	0.705	0.647
	RCNN	0.689	0.690	0.690	0.700
	VGG19	0.730	0.698	0.789	0.626
推特	ResNet152	0.591	0.558	0.731	0.452
	RCNN	0.590	0.570	0.720	0.470
	VGG19	0.596	0.593	0.695	0.518

力机制的BiLSTM模型提取文本特征，采用VGG19模型提取视觉特征，同时提出了一种事件对抗的神经网络框架，可以提取事件不变特征，从而有利于对新事件的谣言进行检测。同时，本文提出的谣言检测器，完全连接层的隐藏大小为64。事件判别器由两个完全连接层组成：第一层的隐藏大小为64，第二层的隐藏大小为32。为了验证跨模态谣言检测模型的有效性，实现了多种跨模态模型进行性能对比，主要对比的模型如下。

（1）VQA[26]。VQA模型旨在根据给定的图像回答问题，同时原任务为多分类任务。为了做对比实验，将VQA模型改为二分类任务。此外，设计使用一层LSTM，LSTM层的隐藏大小为32。

（2）Att-RNN[9]。Att-RNN模型使用注意力机制组合文本、视觉和社交上下文特征。该模型中使用LSTM将文本和社交环境信息进行联合表示，再使用注意力模型融合视觉特征。为了更好地进行对比，实验中删除了处理社交环境信息的部分。

（3）MSRD[27]。刘金硕等[27]提出了一种融合文本信息、图像信息以及图像中的文本信息的谣言检测模型。该模型采用VGG19模型提取图像内容特征，采用DenseNet模型提取图像内嵌文本内容，使用LSTM提取文本内容特征，将其与图像特征连接后，通过完全连接层获取图像与文本共享表示的均值与方差向量，借助从高斯分布中采样的随机变量形成重新参数化的多模态特征并作为谣言检测器的输入进行谣言检测。

（4）DCNN[28]。陈志毅等[28]提出了一种针对文本、图像以及用户属性信息的多模态谣言检测方法DCNN。DCNN模型由多模态特征提取器和谣言检测器组成，多模态特征提取器分为三部分，即基于TextCNN的文本特征提取器、基于VGG19模型的图片特征提取器和基于DeepFM算法的用户社会特征提取器，分别用于学习微博不同模态上的特征表示，以形成重新参数化的多模态特征，将融合后的多模态特征作为谣言检测器的输入进行分类检测。

针对多种模态的谣言检测结果如表4所示。可以看出，本文模型性能优于其他对比模型。尽管单模态模型对谣言检测是有效的，但是仍比跨模态的性能差，证明了跨模态谣言检测方法是具有优越性。在跨模态谣言检测模型中，Att-RNN模型的性能优于VQA模型，表明应用注意力机制有助于改善模型性能；MSRD模型的性能也优于VQA模型，表明图像内的文本信息有助于谣言检测。DCNN模型在对比模型中表现最佳，表明在跨模态谣言检测中加入用户社会特征可以提高谣言检测性能。本文模型在准确率、精确率、召回率和F1值等方面都优于对比模型，并且准确率在两个数据集上分别比最优结果提高了3.6个百分点和3.5个百分点。与其他对比模型中最佳实验结果相比，本文模型的F1值在两个数据集上分别从0.799提高到了0.835（微博数据集），从0.678提高到了0.725（推特数据集），这表明本文模型在准确率和F1值的总体表现也比较好。其原因在于，对于文本，BiLSTM模型通过增加控制门解决了梯度消失和梯度爆炸的问题，解决了信息长期依赖问题，Attention机制能够捕获文本中重要信息，减少重要特征丢失，两种模型结合增强了文本特征提取能力。另外，VGG19模型的结构非常简洁有效，和ResNet模型比较，具有更好的性能。最后，利用端到端的事件对抗性神经网络框架，谣言检测的主要挑战来自新出现的事件，而现有方法在这些事件上未表现出令人满意的性能。为了解决这个问题，本文提出事件对抗神经网络框架，该框架可以提取事件不变特征，能够对新出现的事件进行有效、合理的检测，因此效果更优异。

4.6　消融实验分析

为了验证本文模型各组成部分的重要性，设计了模型的变体，分别是去掉对抗神经网络（Ours w/o adv）、去掉VGG19模型（Ours w/o VGG19）、去掉融合自注意力机制的BiLSTM模型（Ours w/o BiLSTM-Attention），主要包括三种变体形式。

表4　跨模态谣言检测性能对比

Table 4　Cross-Modal Rumor Detection Results

数据集	模型	准确率	F1值	精确率	召回率
微博	VQA	0.736	0.706	0.797	0.634
	Att-RNN	0.772	0.789	0.778	0.799
	MSRD	0.794	0.779	0.854	0.716
	DCNN	0.803	0.799	0.801	0.809
	本文	0.839	0.835	0.853	0.818
推特	VQA	0.631	0.611	0.765	0.509
	Att-RNN	0.664	0.676	0.749	0.615
	MSRD	0.685	0.678	0.725	0.636
	DCNN	–	–	–	–
	本文	0.720	0.725	0.832	0.643

表5　消融实验结果对比

Table 5　Ablation Results

数据集	模型	准确率	F1值	精确率	召回率
微博	Ours	0.839	0.835	0.853	0.818
	Ours w/o adv	0.806	0.812	0.816	0.809
	Ours w/o VGG19	0.774	0.775	0.775	0.774
	Ours w/o BiLSTM-Attention	0.757	0.746	0.756	0.738
推特	Ours	0.720	0.725	0.832	0.643
	Ours w/o adv	0.656	0.629	0.824	0.509
	Ours w/o VGG19	0.620	0.604	0.615	0.606
	Ours w/o BiLSTM-Attention	0.592	0.597	0.604	0.590

（1）去掉对抗神经网络。对于文本模态，采用融合自注意力机制的BiLSTM模型提取文特征；对于视觉模态，采用预训练VGG19模型提取图像特征，之后将两种模态特征向量进行拼接并作为输入，送入带有Softmax的全连接层进行分类。

（2）去掉VGG19模型。移除微博或推特帖子的图片部分，此时输入的跨模态表示由原来的文本模态和视觉模态的联合表示变为仅是文本模态的特征表示。同时存在事件判别器和谣言检测器。

（3）去掉融合自注意力机制的BiLSTM模型。移除微博或推特帖子的文本部分，此时输入的跨模态表示由原来的文本模态和视觉模态的联合表示变为仅是图像模态的特征表示。同时存在事件判别器和谣言检测器。

消融实验结果如表5所示，可以得到两个结论。首先，移除模型的任何部分，模型的分类准确率都会出现一定程度的下降，这说明了各模型组件对于实验的有效性；其次，按照移除后模型分类准确率的下降程度，可以将各模型组件的重要性进行排序：文本模态>视觉模态>对抗神经网络。这说明对于谣言检测任务，文本比图像发挥的作用更重要，对抗神经网络的作用相比两种模态较弱，但是将这三种元素组合在微博和推特数据集上进行实验，实验结果要优于消融实验的结果，再次验证了将BiLSTM-Attention、VGG19以及对抗神经网络三者结合运用于跨模态谣言检测任务的优越性。

4.7　收敛分析

为验证模型的收敛性，对于公式（13）中的参数，实验中取α=10，β=0.75，观察损失函数的变化，实验结果如图2和图3所示。在训练过程中，开始阶段数据集上训练集损失、测试集损失和辨别损失（对抗性损失）都有所减少，随后辨别损失增加并稳定在一定水平。在开始时减少的辨别损失表示跨模态特征提取器的特征表示中的事件特定信息。当判别器和特征提取器之间的极小极大博弈继续进行时，事件特定信息被逐渐删除，特征表示慢慢趋向于事件不变，并且辨别损失随着时间的推移而增加。在训练过程中，三种损失平稳收敛，这意味着达到了一定程度的均衡。随着训练损失的稳步下降，测试损失也在稳步下降，并且呈现出非常相似的趋势，对抗性损失在前期呈现轻微的先下降后上升的状态，随着训练的继续进行逐渐趋于平稳，这一结果也证明了本文使用对抗神经网络学习事件通用特征的意义。如果对抗性损失被优化到0，则意味着模型学习到的是事件的特定特征，这一特征将很难用于迁移学习以提升模型的泛化能力。综上所述，通过本文模型可以有效捕获所有事件中的不变信息，并且可以将这种特征迁移到新发生的谣言事件检测中，对于提升模型的泛化能力很有帮助。

为了进一步分析事件判别器的有效性，将本文模型和去除对抗神经网络的模型在微博数据集上学习的文本特征表示进行可视化。去除对抗神经网络的文本特征表示的可视化和带有对抗神经网络的文本特征表示的可视化分别如图4和图5所示。其中，红点表示为谣言的标签特征，蓝点表示为非谣言的标签特征。由于微博测试数据集数量比推特测试数

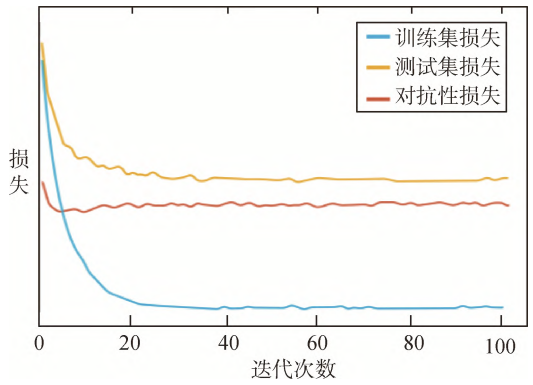

图2　微博数据集中训练集损失、测试集损失以及对抗性损失对比

Fig.2　Training Set Loss, Test Set Loss and Adversarial Loss in Microblog Dataset

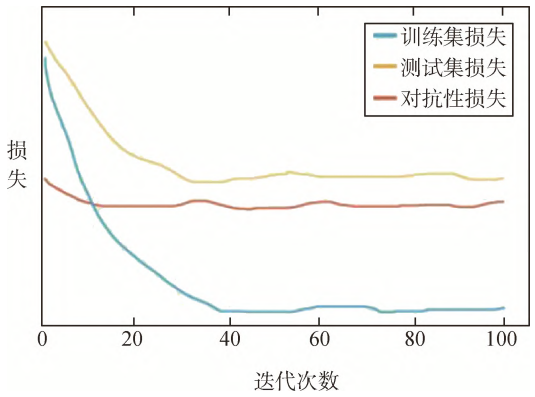

图3　推特数据集中训练集损失、测试集损失以及对抗性损失对比

Fig.3　Training Set Loss, Test Set Loss and Adversarial Loss in Twitter Dataset

据集数量少，微博测试数据集上在移除和未移除对抗神经网络点的分布都多于推特测试数据集在移除和未移除对抗神经网络点的分布。根据特征分布可以观察到，移除对抗神经网络的模型虽然可以学习可辨别的特征，但学习的特征仍然缠绕在一起，相比之下，本文模型学习到的特征表示更容易辨别，并且存在更大的隔离区域图。这也证明在训练阶段，事件判别器尝试移除特征表示和特定事件之间的依赖关系。在极大极小博弈的帮助下，跨模态特征抽取器可以学习不同事件的不变特征表示，从而在应对未知突发事件时，可以利用学习到的通用特征进行迁移学习以辨别突发谣言的真伪，使模型迁移能力更强，提升了对新事件的泛化能力，从而提升了谣言检测的性能。在微博和推特数据集上对本文模型进行消融实验分析可以说明文本模态对谣言检测的影响最大，说明文本特征相比图像特征发挥的作用更大，视觉模态对谣言的检测也有很重要的影响，最后对抗神经网络基于两种模态在谣言检测中也发挥了明显作用，提升了对新事件的泛化能力，在谣言传播的早期能及时、有效地检测出谣言。

(a) 移除对抗神经网络　(b) 有对抗神经网络

图4　微博测试集上学习的文本特征表示的可视化

Fig.4　Visual Comparison of Text Feature Representations Learned on the Microblog Test Set

(a) 移除对抗神经网络　(b) 有对抗神经网络

图5　推特测试集上学习的文本特征表示的可视化

Fig.5　Visual Comparison of Text Feature Representations Learned on the Twitter Test Set

5　结　语

本文提出了对抗神经网络的思想，利用对抗神经网络去除对特定事件的特征，学习所有事件的共同特征，进一步预测新出现的事件是否为谣言。整体模型框架由三部分组成，分别是跨模态特征提取器、事件判别器、谣言检测器。通过跨模态特征提取器与事件判别器相互博弈，去除事件特定特征学习事件不变表示，可以消除对谣言中的特定事件的紧

密依赖关系，同时获取谣言中事件的共享特征，使模型特征迁移能力更强，提升了对新出现事件的谣言识别的泛化能力。最后，跨模态提取器与谣言检测器合作识别谣言，在常用的社交媒体平台微博和推特上收集数据进行大量实验，表明了本文模型的有效性。

参考文献：

[1] 谭荧，张进，夏立新．社交媒体情境下的情感分析研究综述[J]. 数据分析与知识发现，2020，4（1）：1-11.(Tan Ying, Zhang Jin, Xia Lixin. A Survey of Sentiment Analysis on Social Media[J]. Data Analysis and Knowledge Discovery, 2020, 4(1): 1-11.)

[2] Castillo C, Mendoza M, Poblete B. Information Credibility on Twitter [C]//Proceedings of the 20th International Conference on World Wide Web. 2011: 675-684.

[3] Wu K, Yang S, Zhu K Q. False Rumors Detection on Sina Weibo by Propagation Structures[C]//Proceedings of the 31st International Conference on Data Engineering. 2015: 651-662.

[4] 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. 2017: 2931-2937.

[5] Gupta A, Lamba H, Kumaraguru P, et al. Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy [C]//Proceedings of the 22nd International Conference on World Wide Web. 2013: 729-736.

[6] Guo H, Cao J, Zhang Y Z, et al. Rumor Detection with Hierarchical Social Attention Network[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 943-951.

[7] 蒋雨肖，丁晟春，吴鹏．基于BiLSTM-VGG16的多模态信息特征分类研究[J]. 情报理论与实践，2021，44(11)：180-186，179.(Jiang Yuxiao, Ding Shengchun, Wu Peng. A Study on the Classification of Features of Multi-Modal Information Based on BiLSTM-VGG16[J]. Information Studies: Theory & Application, 2021, 44(11): 180-186, 179.)

[8] 李莎，张怀文，钱胜胜，等．多模态多层次事件网络的谣言检测[J]. 中国图象图形学报，2021，26(7)：1648-1657.(Li Sha, Zhang Huaiwen, Qian Shengsheng, et al. Multi-Modal Multi-Level Event Network for Rumor Detection[J]. Journal of Image and Graphics, 2021, 26(7): 1648-1657.)

[9] 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. 2017: 795-816.

[10] 宋之杰，王建，贾杰．基于 $SIHR_1R_2$ 的突发事件谣言传播研究[J]. 情报杂志，2016，35（3）：118-124，136.(Song Zhijie, Wang Jian, Jia Jie. Research on Rumor Spreading in Emergency Based on $SIHR_1R_2$ Model[J]. Journal of Intelligence, 2016, 35(3): 118-124, 136.)

[11] Kwon S, Cha M, Jung K, et al. Prominent Features of Rumor Propagation in Online Social Media[C]//Proceedings of the 13th International Conference on Data Mining. 2013: 1103-1108.

[12] Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification[J]. Journal of Machine Learning Research. 2002, 2：45-66.

[13] 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. 2016: 3818-3824.

[14] 刘政，卫志华，张韧弦．基于卷积神经网络的谣言检测[J]. 计算机应用，2017，37(11)：3053-3056. (Liu Zheng, Wei Zhihua, Zhang Renxian. Rumor Detection Based on Convolutional Neural Network[J]. Journal of Computer Applications, 2017, 37 (11): 3053-3056.)

[15] Chen T, Li X, Yin H Z, et al. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection[C]//Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2018: 40-52.

[16] Popat K, Mukherjee S, Yates A, et al. DeClarE: Debunking Fake News and False Claims Using Evidence-Aware Deep Learning [OL]. arXiv Preprint, arXiv: 1809.06416.

[17] 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. 2018: 849-857.

[18] Yang X Y, Lyu Y F, Tian T, et al. Rumor Detection on Social Media with Graph Structured Adversarial Learning[C]// Proceedings of the 29th International Conference on Artificial Intelligence. 2021: 1417-1423.

[19] Ni S W, Li J W, Kao H Y. Rumor Detection on Social Media with Hierarchical Adversarial Training[OL]. arXiv Preprint, arXiv: 2110.00425.

[20] 李奥，但志平，董方敏，等．基于改进生成对抗网络的谣言检测方法[J]. 中文信息学报，2020，34(9)：78-88.(Li Ao, Dan Zhiping, Dong Fangmin, et al. An Improved Generative Adversarial Network for Rumor Detection[J]. Journal of Chinese Information Processing, 2020, 34(9): 78-88.)

[21] Ganin Y, Lempitsky V. Unsupervised Domain Adaptation by Backpropagation[C]//Proceedings of the 32nd International Conference on Machine Learning. 2015: 1180-1189.

[22] Ganin Y, Ustinova E, Ajakan H, et al. Domain-Adversarial Training of Neural Networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.

[23] Maigrot C, Claveau V, Kijak E, et al. MediaEval 2016: A Multimodal System for the Verifying Multimedia Use Task[C]// Proceedings of the 2016 MediaEval Workshop. 2016.

[24] Chen Y H. Convolutional Neural Network for Sentence Classification[D]. Waterloo, ON: University of Waterloo, 2015.

[25] Schuster M, Paliwal K K. Bidirectional Recurrent Neural Networks[J]. IEEE Transactions on Signal Processing, 1997, 45 (11): 2673-2681.

[26] Antol S, Agrawal A, Lu J S, et al. VQA: Visual Question Answering[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. 2015: 2425-2433.

[27] 刘金硕, 冯阔, Jeff Z. Pan , 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 Jinshuo, Feng Kuo, Pan J Z, et al. MSRD: Multi-Modal Web Rumor Detection Method[J]. Journal of Computer Research and Development, 2020, 57(11): 2328-2336.)

[28] 陈志毅, 隋杰 . 基于 DeepFM 和卷积神经网络的集成式多模态谣言检测方法 [J]. 计算机科学 , 2022, 49(1): 101-107. (Chen Zhiyi, Sui Jie. DeepFM and Convolutional Neural Networks Ensembles for Multimodal Rumor Detection[J]. Computer Science, 2022, 49(1): 101-107.)

作者贡献声明：

孟佳娜：提出研究思路；
王晓培：设计研究方案，论文修订；
李婷：起草论文，设计实验；
刘爽：论文修订；
赵迪：数据处理。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

[1] 刘爽 . 实验数据集 . DOI:10.57760/sciencedb.j00133.00016.
[2] 刘爽 . 预处理后数据 . DOI:10.57760/sciencedb.j00133.00017.
[3] 刘爽 . 实验参数 . DOI:10.57760/sciencedb.j00133.00018.

收稿日期：2022-01-21
收修改稿日期：2022-05-29

Cross-Modal Rumor Detection Based on Adversarial Neural Network

Meng Jiana　Wang Xiaopei　Li Ting　Liu Shuang　Zhao Di
(School of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China)

Abstract: [**Objective**] This paper proposes an adversarial neural network model combining the text and image data, aiming to improve the effectiveness of rumor detection. [**Methods**] First, we integrated the self-attention mechanism with the Bi-directional Long Short-Term Memory network (BiLSTM) model to represent the text features. Then, we used the pre-trained VGG19 network model to represent the image features. Finally, we used the adversarial neural network to study the events' common features. [**Results**] It is superior to the existing baseline models in terms of accuracy, precision, recall and F1 scores. The accuracy on Weibo and Twitter data sets is 3.6% and 3.5%, higher than the best result compared with the baseline models respectively. [**Limitations**] More research is needed to examine the feature association between the modal information, and bridge the semantic gap of cross-modal data. [**Conclusions**] The proposed model could more effectively learn feature representation and detect rumors.

Keywords: Rumor Detection　Adversarial Neural Network　Bi-directional Long Short-Term Memory　Self-Attentional Mechanism　VGG19