计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400057

融合多模态信息的社交网络谣言检测方法

张少钦[1]　杜圣东[1,2,3]　张晓博[1,2,3]　李天瑞[1,2,3]
1 西南交通大学信息科学与技术学院　成都 611756
2 西南交通大学人工智能研究院　成都 611756
3 综合交通大数据应用技术国家工程实验室　成都 611756
(zsq1024@163.com)

摘　要　随着社交网络平台的发展,社交网络已经成为人们获取信息的重要来源。然而社交网络的便利性也导致了虚假谣言的快速传播。与纯文本的谣言相比,带有多媒体信息的网络谣言更容易误导用户以及被传播,因此对多模态的网络谣言检测在现实生活中有着重要意义。研究者们已提出若干多模态的网络谣言检测方法,但这些方法都没有充分挖掘出视觉特征和融合文本与视觉的联合表征特征。为弥补这些不足,提出了一个基于深度学习的端到端的多模态融合网络。该网络首先抽取出图片中各个兴趣区域的视觉特征,然后使用多头注意力机制将文本和视觉特征进行更新与融合,最后将这些特征进行基于注意力机制的拼接以用于社交网络多模态谣言检测。在推特和微博公开数据集上进行对比实验,结果表明,所提方法在推特数据集上F1值有13.4%的提升,在微博数据集上F1值有1.6%的提升。

关键词:多模态;谣言检测;深度学习;目标检测
中图法分类号　TP391

Social Rumor Detection Method Based on Multimodal Fusion

ZHANG Shao-qin[1],DU Sheng-dong[1,2,3],ZHANG Xiao-bo[1,2,3] and LI Tian-rui[1,2,3]
1 School of Information Science and Technology,Southwest Jiaotong University,Chengdu 611756,China
2 Institute of Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
3 National Engineering Laboratory of Integrated Transportation Big Data Application Technology,Chengdu 611756,China

Abstract　With the development of social networking platforms,social networks have become an important source of information for people. However,the convenience of social networks has also led to the rapid propagation of false rumors. Compared with textual rumors,social network rumors with multimedia content are more likely to mislead users and get dissemination,so the detection of multi-modal rumors is of great significance in real life. Several multi-modal rumor detection methods have been proposed,but the visual features and joint representation of text and visual features have not been fully explored in current approaches. To make up for these shortcomings,an end-to-end multi-modal fusion network based on deep learning is developed. Firstly,the visual features of each region of interest in the image are extracted. Then,the text and the visual features are updated and fused by using a multi-head attention mechanism. Finally,these features are concatenated based on the attention mechanism for the detection of multi-modal rumors in social networks. Comparative experiments on the public data sets of Twitter and Weibo are conducted and experimental results show that the proposed method has a 13.4% F1 value increase on Twitter data set and a 1.6% F1 value increase on Weibo data set.

Keywords　Multi-moda,Rumor detection,Deep learning,Object detection

1　引言

根据以往研究对网络谣言的定义,网络谣言是指在网络平台产生、传播、影响的某阶段或全过程起到过关键作用的,内容未经证实且造成了一定社会舆论影响的阐释[1]。谣言不一定是假的,谣言是人们感兴趣或觉得重要的、但未经证实的阐述。随着移动互联网的迅速发展,社交媒体平台诸如新浪微博、Twitter等已经成为人们获取信息的重要来源。然而社交媒体的便利性也导致了谣言可以被迅速传播,恶意的谣言还可能带来巨大的经济和社会影响。与文本相比,图像可以描述视觉内容,吸引更多注意力,因此谣言可以利用多媒体内容误导用户,从而获得快速的传播。由于虚假谣言在社交媒

到稿日期:2020-04-14　返修日期:2020-07-13　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2017YFB1401400)
This work was supported by the National Key R&D Program of China (2017YFB1401400).
通信作者:李天瑞(trli@swjtu.edu.cn)

体平台上的广泛传播，国内外采取了一定措施来识别谣言。例如国外的 Snopes.com 和国内的微博辟谣平台等网站向公众公布虚假信息。但是这些平台需要有相关知识的人来鉴别谣言的真假，这类人工检测的手段非常耗时耗力，因此亟需使用自动化的手段来检测社交媒体中的多模态虚假谣言。

多模态研究涉及领域很多，如图像字幕(Image Caption)[2]和视觉问答(Visual Question Answer，VQA)[3-4]。图像字幕是指从图像生成文本描述的过程。视觉问答是对视觉图像的自然语言表达，模型需要在理解图像的基础上根据问题做出回答。这两个领域的工作都有将图片的目标区域的特征提取出来并将特征和文本特征相融合的过程。

多模态融合将多个模态信息进行整合以得到一个模型的输出，多模态信息融合能够获得更全面的特征，更能挖掘出数据的深层次信息。多模态融合有多种不同的方法，这些方法可以被分为早期融合、中间融合和晚期融合[5]。早期融合是在模型输入阶段就对不同模态的特征进行某种方式的拼接，然后再进行分类训练。中间融合则是在模型的中间训练过程中对多模态特征进行对齐和融合。晚期融合是将每种模态信息单独进行分类训练，最后将结果进行投票决策。

现有的谣言检测工作大多是基于单模态的，基于多模态的研究还处于比较初级的阶段。多模态谣言检测主要需要解决两个问题：一是如何从图像中提取出特征；二是如何将图片特征和文本特征相融合。为解决这两个问题，本文提出了一个基于多模态融合网络(Multimodal Fusion Network，MFN)的谣言检测模型，该模型先利用深度学习提取文本特征与视觉特征，再通过注意力机制来融合视觉和文本的特征。实验结果表明，相比目前最新的多模态谣言检测方法，该模型在性能上有所提升。

2 相关工作

由于虚假谣言检测的重要性，国内外许多学者都对此领域进行了深入的研究。早期的研究主要集中于提取手工设置的有效特征[6-9]，并通过训练分类器(如决策树)来进行虚假谣言检测。有效特征主要分为两大类：一是消息本身的内容的特征；二是传播过程中的上下文的特征。其中，消息本身的特征包括语言学特征和多媒体特征等；上下文特征包括用户特征和评论特征等。针对多模态数据，Jin 等[10]根据视觉特征和统计特征来识别虚假信息。其中，视觉特征是指图像内容特征，包括图像清晰度评分、图像多样性评分和图像聚类评分等；统计特征是指图像数量的统计特征，包括图像计数、多图像比和长图像比等。该方法是首次尝试将图像特征用于虚假谣言的验证任务。但是该工作主要依据人工设置的特征，因此需要复杂的特征工程。随着深度学习的发展，深度学习在文本和图像中的应用越来越广泛，与手工设置的特征相比，端到端的网络更能抽取到深层次的特征。

近年来，研究者们开始利用深度学习来研究虚假谣言检测问题。例如，在基于谣言的文本内容上，Liu 等[11]提出了基于卷积神经网络(Convolutional Neural Network，CNN)的谣言检测模型。该模型将微博中的谣言事件向量化，通过 CNN 隐含层的学习训练来挖掘表示文本深层的特征，避免了特征构建的问题，但是该方法没有考虑各个微博的时序关系。在考虑谣言的时序传播上，Liao 等[12]提出了一种基于分层注意力网络的社交网络谣言检测方法，该方法通过采用两层带有注意力机制的双向门控循环单元(Gated Recurrent Unit，GRU)网络来分别获取微博和时间段序列的隐层表示，从而在事件的特征表示中融入时间段内各微博间的时序信息。针对传播的拓扑结构，Ma 等[13]将深度学习模型用于谣言的传播结构上，提出了两种基于自下而上和自顶向下的树结构的神经网络的谣言表征学习和分类递归模型。实验表明该方法在早期检测阶段有着很好的性能。Ruchansky[14]通过矩阵分解得到文章的文本、收到的用户响应以及推广它的源用户这 3 个特征的交互特征，最后将其送入长短期记忆网络(Long Short-Term Memory，LSTM)[15]得到用户得分和分类结果，该方法避免了传播树的构建，同时获得了传播时序和传播结构的特征。Zhang 等[16]基于新闻文章、创作者和新闻主体之间的联系，提出了一种将网络结构信息纳入模型学习的深度扩散网络模型。Shu 等[17]对出版商、新闻和用户之间的交互关系进行建模，提出了三关系嵌入框架 TriFN，该框架可同时对发布者-新闻关系和用户-新闻交互进行建模，以进行假新闻分类。Liu 等[18]将循环神经网络(Recurrent Neural Network，RNN)[19]和 CNN 用于谣言的传播结构，建立了一个时间序列分类器，分别捕捉用户特征在传播路径上的全局和局部变化以检测假新闻。实验表明，该模型在虚假新闻开始传播后 5 min 能检测出假新闻。

以上工作多是基于单模态的文本特征。最近越来越多的学者开始关注多模态特征融合，Yang 等[20]将文本和图像信息与相应的显式特征和潜在特征相结合以用于谣言检测。显式特征是手工设计的特征，潜在特征则是用 CNN 进行抽取的特征，实验证明该方法在解决假新闻检测问题方面具有有效性。Jin 等[21]将多模态微博数据用于网络谣言检测，提出了一种具有注意机制的递归神经网络(att-RNN)来融合多模态特征。该端到端网络中，图像特征被结合到文本和上下文的联合特征中，这些联合特征通过 LSTM 获得，以产生可靠的融合分类。实验表明，该方法在检测多模态谣言上具有有效性。Wang[22]基于前者的工作增加了对抗网络，提出了事件对抗神经网络(Event Adversarial Neural Network，EANN)，该网络中多模态特征提取器被强制学习事件不变表示以欺骗鉴别器。通过这种方式，它消除了对收集到的数据集中的特定事件的紧密依赖，并且获得了对未知事件更好的泛化能力。

基于以上相关研究工作，我们发现现有的方法还没有充分挖掘出视觉特征和融合文本特征与视觉特征的联合表征，新的多模态融合谣言检测模型与方法亟待发展。

3 预备知识

本节介绍所提方法所需要的预备知识。

(1)RNN[19]。RNN 是一种具有记忆功能的神经网络，适合用于序列数据的建模，是除 CNN 外深度学习中最常用的一种网络结构。但是 RNN 模型在训练过程中存在梯度消失和梯度爆炸的问题。

(2)LSTM[15]。LSTM 模型是在 RNN 的基础上通过加入门控机制来使其具有遗忘功能，这解决了由传统的 RNN 循环单元的状态在每个时间步会被覆盖造成的无法处理长期依赖的问题。LSTM 单元主要由细胞状态和门结构组成，细胞状态负责对历史信息进行存储，门结构负责保护和控制细胞状态。一个记忆单元具有 3 个门结构，分别是输入门、输出门、遗忘门。其中，遗忘门决定了从细胞状态中舍弃的信息，从而达到对历史信息进行过滤的效果，解决了梯度消失的问题。在 t 时刻，对于给定的输入 $\boldsymbol{x}_t$，LSTM 的计算公式如下：

$$\begin{aligned}
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1}) \\
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1}) \\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1}) \\
\tilde{\boldsymbol{c}}_t &= \tanh(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1}) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{1}$$

其中，$\boldsymbol{W}_i$，$\boldsymbol{U}_i$，$\boldsymbol{W}_f$，$\boldsymbol{U}_f$，$\boldsymbol{W}_o$，$\boldsymbol{U}_o$，$\boldsymbol{W}_c$，$\boldsymbol{U}_c$ 为权重矩阵，$\boldsymbol{h}_t$ 为隐藏层向量，σ 为 sigmoid 函数，$\odot$ 表示点积。

但是，LSTM 只能记住当前时间步之前的信息，而无法获知下文的信息。

(3) BLSTM[23]。双向长短期记忆网络(Bidirectional Long Short-Term Memory，BLSTM)通过两个 LSTM，即分别记忆前向序列和后向序列，将两个方向的隐藏状态在相应时间步上进行拼接，扩展了 LSTM 对上下文的学习能力。对于 t 时刻，BLSTM 的输出如式(2)所示：

$$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}] \tag{2}$$

(4)Text-CNN[24]。对于社交网络上的短文本来说，RNN 可能不能充分挖掘出文本的语义信息，因此这时提取文本的局部信息，即 n-gram 特征，可能会更有效。Text-CNN 通过 CNN 来获取文本的局部信息。Text-CNN 通过多个不同窗口大小的一维卷积核来获取 n-gram 的局部文本特征。具体来说，假设 n 个单词的特征表示为 $\boldsymbol{X} \in \mathbf{R}^{n*k}$，则 h-gram 的卷积核的大小为 $f \in \mathbf{R}^{h*k}$，每次卷积的操作可以表示为 $t_i = \sigma(\boldsymbol{W}_c \boldsymbol{X}_{i,i+h-1})$，其中，$\boldsymbol{X}_{i,i+h-1}$ 表示从第 i 个单词开始的 h 个单词特征，$\boldsymbol{W}_c$ 是卷积核的参数。最后将多个卷积特征进行拼接，然后将其送入全连接层中。

(5)RCNN[25]。RCNN(Regions with CNN features)是指将 CNN 方法应用到目标检测中，即采用 Selective Search 方法来选取候选框，并将每个候选框送入 CNN 网络以抽取特征。

(6)Faster RCNN[26]。与 RCNN 不同，Faster RCNN 首先通过区域生成网络(Region Proposal Network，RPN)来选取候选框，然后根据候选框来裁剪多个特征图的特征，最后将特征进行训练和分类。其特点在于候选框的生成和最后特征的抽取都在同一个特征图上进行训练。因此，相比 RCNN，Faster RCNN 的检测速度更快。同时，Faster RCNN 提出了 RoI Pooling 层，将多个特征图进行池化，而不是只使用最后一层卷积的特征，这使得抽取的图片特征更加具有鲁棒性。

4 一种新的多模态融合谣言检测方法

本文提出了一个基于 MFN 的谣言检测新模型，其包含 4 个模块，分别是图像特征提取模块、文本特征提取模块、多模态融合模块和分类模块。模型结构如图 1 所示，主要有以下特点：

(1)引入预训练的目标检测网络 Faster RCNN，以抽取图片中多个目标兴趣区域的特征；

(2)基于多头注意力机制的方法，利用文本特征和视觉特征来增强彼此的表示；

(3)多个模态的特征进行基于注意力权重的拼接，以增强某个模态的特征贡献。以下分别介绍 MFN 模型中的各个模块。

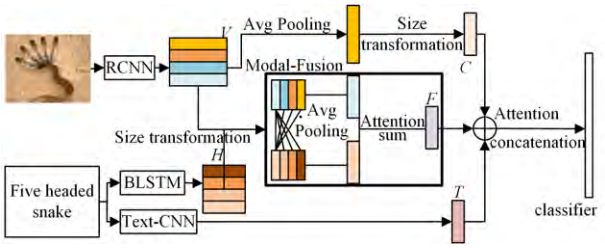

图 1 MFN 的结构图

Fig. 1 Structure diagram of MFN

4.1 文本特征抽取

由于社交网络的推文多为短文本，MFN 模型采用 BLSTM 和 Text-CNN 来抽取文本特征，以提取文本的时序语义特征和 n-gram 局部特征。本文以预训练的词向量作为模型的输入，将其分别输入到 BLSTM 和 Text-CNN 中。通过 BLSTM 获取文本特征，隐藏层表示向量为 $\boldsymbol{H} \in \mathbf{R}^{n*d_H}$，其中，$d_H$ 是特征维度，n 是单词的个数。通过 Text-CNN 模块来获取文本的局部特征，得到的特征记为 $\boldsymbol{T} \in \mathbf{R}^{d}$。

4.2 视觉特征抽取

为抽取视觉特征，MFN 模型使用至下而上和至上而下模型(Bottom-up & top-down attention model)[3]中的 Faster RCNN 模块来抽取视觉特征，即通过在 Visual Genome 数据集上以 ResNet-101 为骨干网络的预训练的 Faster RCNN 来抽取候选框的视觉特征。每张图片抽取固定数量的候选框。

通过以上方法我们获得图片候选框特征 $\boldsymbol{V} \in \mathbf{R}^{m*d_v}$，其中，$m$ 是候选框数量，d_v 是每个候选框的视觉特征维度。

为了不丢失图像的原始特征，MFN 模型也将 Fater RCNN 抽取的若干个候选框的视觉特征直接进行平均池化，并送入激活函数为 RELU 的全连接层中，得到的视觉特征记为 $\boldsymbol{C} \in \mathbf{R}^{d}$。

4.3 多模态融合

与 Gao 等[4]的工作类似，MFN 模型采用多头放缩点积注意力[27]来进行文本和视觉特征的融合。其特点在于考虑文本中的每个单词特征与每个候选框特征的潜在联系，从而对两者进行更新与融合。具体处理流程如下。

(1)把 LSTM 生成的文本特征转换为 key 和 $value$，即 $\boldsymbol{H}_K$ 和 $\boldsymbol{H}_V$。计算过程如式(3)所示：

$$\boldsymbol{H}_K=\text{Linear}(\boldsymbol{H};\boldsymbol{\theta}_{HK})$$
$$\boldsymbol{H}_V=\text{Linear}(\boldsymbol{H};\boldsymbol{\theta}_{HV}) \tag{3}$$

其中，$\boldsymbol{\theta}_{HK}$，$\boldsymbol{\theta}_{HV}$是训练参数，Linear 表示不带激活函数的全连接层，$\boldsymbol{H}_K$，$\boldsymbol{H}_V\in\mathbf{R}^{n*\dim}$，dim 是转换后的特征维度。

(2)通过全连接层将视觉特征转换为和文本特征 $\boldsymbol{H}$ 相同的特征维度 $\boldsymbol{M}\in\mathbf{R}^{m*d_H}$，然后将转换后的图像特征转换为 $query$，$\boldsymbol{M}_Q$ 的计算过程如式(4)所示：

$$\boldsymbol{M}_Q=\text{Linear}(\boldsymbol{M};\boldsymbol{\theta}_{MQ}) \tag{4}$$

其中，$\boldsymbol{\theta}_{MQ}$是训练参数，$\boldsymbol{M}_Q\in\mathbf{R}^{m*\dim}$，$m$ 表示候选框的个数。

(3)采用多头注意力机制以得到多维度的注意力权重。多头注意力就是将注意力计算过程进行多次重复。每一次通过放缩点积来计算注意力权重，计算过程如式(5)所示：

$$\boldsymbol{atten}_i^{MH}=\left(\text{softmax}\frac{\boldsymbol{\theta}_Q\boldsymbol{H}_K^{\mathrm{T}}}{\sqrt{\dim}}\right) \tag{5}$$

其中，softmax 表示激活函数，$\boldsymbol{atten}_i^{MH}\in\mathbf{R}^{n*1}$，将放缩点积重复 s 次，得到最后的注意力权重为 $\boldsymbol{Atten}^{MH}\in\mathbf{R}^{s*n}$。

(4)首先利用文本信息来更新视觉特征，然后将更新后的视觉特征与原有的视觉特征进行拼接，最后通过不带激活函数的全连接层来将其转换为原来的维度。计算过程如式(6)所示：

$$\boldsymbol{M}_{update}=\boldsymbol{Atten}^{MH}\times\boldsymbol{H}_V$$
$$\boldsymbol{M}=\text{Linear}([\boldsymbol{M},\boldsymbol{M}_{update}];\boldsymbol{\theta}_M) \tag{6}$$

其中，$\boldsymbol{\theta}_M$ 是训练参数。最后将视觉特征 $\boldsymbol{M}$ 进行平均池化，得到最终的视觉特征 $\boldsymbol{M}'\in\mathbf{R}^{d_H}$。

为了使 $\boldsymbol{M}_{update}$和 $\boldsymbol{M}$ 维度一致，我们取 dim 的值为 d_H/s。

(5)通过视觉特征引导文本特征的更新，与文本特征引导视觉特征更新的流程类似，将视觉特征转换为 key 和 $value$，将文本特征转换为 $query$，通过多头注意力机制将文本特征更新为 $\boldsymbol{H}'\in\mathbf{R}^{d_H}$。

需要说明的是，本文的 MFN 方法不是简单地将不同模态特征拼接起来，而是通过两层前馈神经网络来计算每次更新后的模态的注意力权值，并将更新后的视觉特征和文本特征通过注意力权重进行融合，计算过程如式(7)所示：

$$\boldsymbol{Atten}^{HM}=\text{softmax}(\boldsymbol{W}_2^{HM}\tanh(\boldsymbol{W}_1^{HM}[\boldsymbol{H}',\boldsymbol{M}']+\boldsymbol{b}_1^{HM})+\boldsymbol{b}_2^{HM})$$
$$\boldsymbol{F}=\text{ReLU}(\text{Linear}(\boldsymbol{Atten}^{HM}[\boldsymbol{H}',\boldsymbol{M}']^{\mathrm{T}})) \tag{7}$$

其中，$\boldsymbol{W}^{HM}$是权重矩阵，$\boldsymbol{b}^{HM}$是偏差，$\boldsymbol{F}\in\mathbf{R}^d$。

4.4 分类器

考虑到不是所有的模态都对最后的分类有相同的贡献。参考 Gu 等[28]的工作，本文通过注意力权重来增加某类模型的表示。本文的 MFN 方法采用两层前馈神经网络来计算注意力权重。为了增强某类模态的表示的同时不损失模态的特征表示，将注意力权重加一。具体流程如式(8)所示：

$$\boldsymbol{\alpha}_F=\text{softmax}[\boldsymbol{W}_2^F\tanh(\boldsymbol{W}_1^F[\boldsymbol{T},\boldsymbol{F},\boldsymbol{C}]+\boldsymbol{b}_1^F)+\boldsymbol{b}_2^F]$$
$$\boldsymbol{Fusion}=(\boldsymbol{\alpha}_F^{\mathrm{T}}+1)[\boldsymbol{T},\boldsymbol{F},\boldsymbol{C}] \tag{8}$$

将增强后的模态特征进行拼接，最后送入多层感知机中进行分类。

4.5 损失函数和优化器

在模型训练过程中，本文的 MFN 方法选取了交叉熵函数作为损失函数，公式如式(9)所示：

$$L(\boldsymbol{y},f(\boldsymbol{x}))=-\frac{1}{n}\sum_x[\boldsymbol{y}\ln(f(\boldsymbol{x}))+(1-\boldsymbol{y})\ln(1-f(\boldsymbol{x}))] \tag{9}$$

其中，$\boldsymbol{y}$ 是真实标签，$f(\boldsymbol{x})$是模型预测的概率，n 是模型训练样本的总数。

本文的 MFN 方法中优化器采用 Adam，对学习率添加动态约束，使其在一定范围内波动。Adam 具体优化的计算过程如式(10)所示：

$$m_t=\mu m_{t-1}+(1-\mu)g_t$$
$$n_t=\upsilon n_{t-1}+(1-\upsilon)g_t^2$$
$$\hat{m}_t=\frac{m_t}{1-\mu^t}$$
$$\hat{n}_t=\frac{n_t}{1-\upsilon^t}$$
$$\theta=\theta-lr\frac{\hat{m}_t}{\sqrt{\hat{n}_t}+\varepsilon} \tag{10}$$

其中，g_t 为 t 时间步的梯度，m_t 和 n_t 分别是对梯度的一阶矩估计和二阶矩估计，$\hat{m}_t$ 和$\hat{n}_t$ 是对 m_t 和 n_t 的校正，μ，υ，ε 和 lr 为超参数。

5 实验验证

5.1 实验设置

本文使用两个标准数据集来进行对比实验。这两个数据集分别来自推特和微博的真实数据。数据集中有推文和对应图像，并且测试集和训练集所包含的事件并不重复。

(1)推特数据集是 MediaEval2016[29]的数据集，主要用来验证多媒体信息任务，即验证社交媒体中虚假的图像，这个数据集首先采集相关事件中的虚假图像和真实图像。然后通过虚假的图像和真实的图像来寻找推特上相关的推文，因此该数据集只包含 500 多条视觉信息，并通过这些图片找到了大约 17000 条相关推文。这个数据集已经被划分为训练集和测试集，且两个数据集覆盖了不同的事件。每个推文都有相关的文本信息和视觉信息。因为本文主要关注视觉特征和文本特征，所以将包含视频的推文去掉。为了与基线模型进行对比，将数据集中的训练集用于训练模型，测试集用于测试模型性能。

(2)微博数据集是 Jin 等[21]的 att-RNN 中所使用的数据集。该数据集是对微博官方辟谣系统中从 2012 年 5 月到 2016 年 1 月的所有经过验证的虚假谣言帖子进行抓取所构成的。数据集的统计信息如表 1 所列。

表 1 数据集统计信息

Table 1 Dataset statistics

(单位：条)

数据集	训练集	测试集
微博	7532	1996
推特	12933	991

5.2 超参数设置

本文对优化器中超参数的设置如表 2 所列。

表 2 优化器超参数设置

Table 2 Setting optimizer super parameters

超参数	数值
μ	0.9
v	0.999
ε	10^{-8}
lr	0.001

为了避免过拟合,本文所有的全连接层的后面都添加了一个 dropout 层。其他超参数选择如表 3 所列。

表 3 网络超参数设置

Table 3 Setting network super parameters

超参数	数值
m(候选框个数)	36
d_v(候选框特征个数)	2048
d_H(LSTM 文本特征维度)	512
s(注意力头数)	8
d(隐藏层维度)	128
dropout	0.1
Text-CNN 卷积核大小	1×20,2×20,3×20,4×20

5.3 评价指标

在分类问题中,通常使用准确度(accuracy)、精确率(precision)、召回率(recall)和 F1 值(F1-score)作为评价指标。首先介绍混淆矩阵的概念,如表 4 所列,TP,TN,FP 和 FN 的含义如下:

(1)True Positive(TP):将正类正确预测为正类的数量;

(2)True Negative(TN):将负类正确预测为负类的数量;

(3)False Positive(FP):将正类错误预测为负类的数量;

(4)False Negative(FN):将负类错误预测为正类的数量。

表 4 混淆矩阵

Table 4 Confusion matrix

	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

精确率表示所有预测为正类样本被成功预测的比例,计算公式如式(11)所示:

$$P=\frac{TP}{TP+FP} \tag{11}$$

召回率表示所有正类样本中被成功预测的比例,计算公式如式(12)所示:

$$R=\frac{TP}{TP+FN} \tag{12}$$

由于精确率和召回率有时会出现矛盾的情况,为了综合考虑这两个指标,通常使用的方法是 F1 值,这相当于是对精确率和召回率做了加权调和平均,计算公式如式(13)所示:

$$F1=\frac{2*P*R}{P+R} \tag{13}$$

为了证明 MFN 能够提取多模态信息,并对虚假谣言进行有效的检测,本文只与一些基于端到端的深度学习的多模态谣言方法进行对比。

5.4 基线模型

本文选取的基线模型不包含处理额外信息的模型。额外信息包括用户信息、评论信息等。以下介绍与本文模型进行对比的基线模型。

(1)Text-CNN[24]。利用 CNN 网络来提取文本 n-gram 局部特征。将卷积后的特征进行拼接后送入多层感知机进行分类。

(2)BLSTM[23]。为了验证文本语义特征的有效性,首先利用 BLSTM 来抽取文本特征,然后将多个时间步的特征进行池化,最后送入多层感知机进行分类。

(3)RCNN。利用 Faster RCNN[26]来获取若干个候选框的特征,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。

(4)VGG19[30]。利用预训练的 VGG19 来获取图像的全局特征,然后将特征送入多层感知机进行分类。

(5)EANN[22]。EANN 首先通过预训练的 VGG19 来抽取视觉特征,然后利用 Text-CNN 模块来抽取文本特征,最后将两种特征进行拼接后送入分类器中。本文只关注多模态融合,因此将 EANN 中的事件分类器去掉,只保留谣言检测器。

(6)att-RNN[21]。att-RNN 利用注意力机制来融合文本、视觉和上下文特征。其中,将 LSTM 的每一时间步的特征作为引导特征生成 512 维的注意力特征,将注意力特征与视觉特征相融合。将融合后的视觉特征与文本特征相拼接并送入分类器中。本文实验删去了上下文特征的部分。

(7)MFN-。在 MFN 模型的基础上去掉模态融合的模块。通过 Fater RCNN 抽取视觉特征,将多个区域特征进行池化;通过 Text-CNN 抽取文本特征,将视觉特征和文本特征进行拼接;将拼接后的特征进行分类。

5.5 结果分析

在推特和微博数据集上的实验结果分别如表 5 和表 6 所列。

表 5 推特实验的结果

Table 5 Results on Twitter dataset of different methods

Model	Accuracy	Precision	Recall	F1
Text-CNN	0.532	0.598	0.541	0.568
BiLSTM	0.567	0.629	0.602	0.615
RCNN	0.592	0.726	0.468	0.568
VGG19	0.596	0.695	0.518	0.593
att-RNN	0.664	0.749	0.615	0.676
EANN	0.648	**0.810**	0.498	0.617
MFN -	0.644	0.744	0.582	0.653
MFN	**0.764**	0.754	**0.876**	**0.810**

表 6 微博实验的结果

Table 6 Results on Weibo dataset of different methods

Model	Accuracy	Precision	Recall	F1
Text-CNN	0.763	0.827	0.683	0.748
BiLSTM	0.752	0.813	0.665	0.731
RCNN	0.699	0.701	0.711	0.706
VGG19	0.615	0.615	0.677	0.645
att-RNN	0.779	0.778	0.799	0.789
EANN	0.795	0.806	0.795	0.800
MFN -	0.790	**0.858**	0.703	0.772
MFN	**0.810**	0.804	**0.828**	**0.816**

推特数据集中只有 500 多张图片,图片的特征不够丰富,因此 VGG19 的实验效果比 RCNN 的实验效果好。但是两者都不足以分辨谣言的真实性。在推特数据集实验中,文本的准确性是所有方法中最低的,原因在于推特数据集覆盖多个

事件，文本的词语差异比较大，因此不能共享文本的局部特征。同时采用 BLSTM 的模型更能抽取复杂文本的语义信息，因此 BLSTM 的实验效果比 Text-CNN 的实验效果好。推特数据集中的图片不丰富，多个推文有相似的图像，因此基于视觉模态的检测效果会比基于文本模态的效果更好。MFN-比 EANN 效果好，这说明图像的局部区域特征对谣言检测是有效的。基于注意力机制的 att-RNN 和 MFN 都比 EANN 的效果好，这说明应用注意力机制对模态进行融合可以帮助改善预测模型的性能。MFN 的 F1 值比 att-RNN 的 F1 值高出 13.4%，这说明与图像全局特征和单词特征的联系相比，图像区域特征和单词特征的潜在联系更有助于改善模型的性能。

在微博数据集上，实验结果与推特数据集类似但有些许不同。微博数据集中，文本模态的检测效果比视觉模态的检测效果好，原因是微博数据集采集过程与推特数据集采集过程有所不同。微博利用局部感知哈希算法把重复的图片去除，因此微博数据集的图像信息比较丰富，从而使得每条推文之间的图像不能够共享一些特征。同时数据集通过单通道聚类来标注不同的微博事件，可能导致事件之间的差异不太明显，这样就使得测试集中的文本与训练集中的文本存在相似的单词特征。因此与 BLSTM 抽取的文本语义相比，Text-CNN 抽取的文本的局部特征更能表征谣言的特征，从 EANN 与 att-RNN 的实验结果中也能得出这样的结论，EANN 的效果比 att-RNN 的效果好，这说明了 Text-CNN 在抽取局部特征上的有效性。微博数据集中的图像在语义上比推特数据集中的图像复杂得多，因此 RCNN 更能抽取到图像的复杂特征。MFN 的 F1 值比 EANN 的 F1 值高出 1.6%，这说明将文本特征与视觉特征相融合是有效的。

在两个数据集中，MFN 模型融合多个模态的特征有助于改善谣言检测模型。两种数据集的采集方法不同，推特数据是通过虚假图像来找到虚假谣言数据，即图片是虚假的而文本内容是真实的或者虚假的；微博数据集来自真实的数据集，数据分布更为复杂，如真实的图像附带虚假的文本内容。因此，MFN 在不同数据集上 F1 值的提升程度是不同的。其在推特数据集上 F1 值最高提升了 13.4%，在微博数据集上 F1 值最高提升了 1.6%。

结束语　如何充分挖掘谣言中的视觉特征并有效融合文本特征和视觉特征是目前多模态的网络谣言检测方法面临的关键问题。本文针对这两个问题提出了基于深度学习的端到端的多模态融合网络——MFN 模型。该模型首先采用预训练目标检测网络 Faster CNN 来挖掘图像的细粒度区域特征，然后通过基于多头注意力的模态融合方法，将多模态特征在多个表示子空间内进行融合，最后在两个公开的数据集上进行实验，实验结果表明所构建的 MFN 模型在两个数据集上的性能都得到了提升。

实验结果表明，随着数据分布的不同，MFN 模型中各个模块的效果也不尽相同。因此，在未来的工作中可以考虑建立一个更加全面的数据集，以保证文本信息和图片信息不重复，从而使得在图像特征和文本特征上的分布更贴近现实生活。同时，随着短视频平台的发展，虚假的谣言视频也在网络上被广泛传播。如何充分挖掘和融合语音特征、文本特征以及视觉特征，从而判断视频的真假也是未来的研究方向之一。

参 考 文 献

[1] CHEN Y F, LI Z Y, LIANG X, et al. Review on Rumor Detection of Online Social Networks[J]. Chinese Journal of Computers, 2018, 41(7): 1648-1677.

[2] KARPATHY A, LI F F. Deep Visual-Semantic Alignments for Generating Image Descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664-676.

[3] ANDERSON P, HE X, BUEHLER C, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. ACM, 2018: 6077-6086.

[4] GAO P, JIANG Z, YOU H, et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering [C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. ACM, 2019: 6632-6641.

[5] BALTRUSAITIS T, AHUJA C, MORENCY L P. Multimodal Machine Learning: A Survey and Taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.

[6] CASTILLO C, MENDOZA M, POBLETE B. Predicting information credibility in time-sensitive social media[J]. Internet Research, 2013, 23(5): 560-588.

[7] ZHAO Z, RESNICK P, MEI Q. Enquiring minds: Early detection of rumors in social media from enquiry posts[C]// Proceedings of the 24th International Conference on World Wide Web. 2015: 1395-1405.

[8] TSCHIATSCHEK S, SINGLA A, GOMEZ R M, et al. Fake News Detection in Social Networks via Crowd Signals[C]// Companion of The Web Conference 2018 on The Web Conference 2018. 2018: 517-524.

[9] 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. ACM, 2013: 729-736.

[10] JIN Z, CAO J, ZHANG Y, et al. Novel visual and statistical image features for microblogs news verification[J]. IEEE Transactions on Multimedia, 2017, 19(3): 598-608.

[11] LIU Z, WEI Z H, ZHANG R X. Rumor detection based on convolutional neural network[J]. Journal of Computer Applications, 2017(11): 21-24, 68.

[12] LIAO X W, HUANG Z, YANG D D, et al. Rumor detection in social media based on a hierarchical attention network[J]. Scientia Sinica Informationis, 2018, 48(11): 1558-1574.

[13] MA J, GAO W, WONG K. Rumor detection on twitter with tree-structured recursive neural networks[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 1980-1989.

[14] RUCHANSKY N, SEO S, LIU Y. CSI: A hybrid deep model for

fake news detection[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017:797-806.

[15] HOCHREITER S,SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation,1997,9(8):1735-1780.

[16] ZHANG J,DONG B,YU P S. Deep Diffusive Neural Network based Fake News Detection from Heterogeneous Social Networks[C]// Proceedings of IEEE International Conference on Big Data. 2019:1259-1266.

[17] SHU K,WANG S,LIU H. Beyond news contents: The role of social context for fake news detection[C]// Proceedings of the 12th ACM International Conference on Web Search and Data Mining. 2019:312-320.

[18] LIU Y,WU Y. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]// Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence. 2018:354-361.

[19] MIKOLOV T,KARAFIAT M,BURGET L,et al. Recurrent neural network based language model[C]// Proceedings of 11th Annual Conference of the International Speech Communication Association. 2010:1045-1048.

[20] YANG Y,ZHENG L,ZHANG J,et al. TI-CNN:Convolutional Neural Networks for Fake News Detection[J]. arXiv: 1806.00749,2018.

[21] JIN Z,CAO J,HAN G,et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs[C]// Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017:795-816.

[22] 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 and Data Mining. ACM,2018:849-857.

[23] SCHUSTER M,PALIWAL K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11):2673-2681.

[24] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACM, 2014:1746-1751.

[25] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. ACM 2014: 580-587.

[26] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017(6):1137-1149.

[27] VASWANI A,SHAZEER N,PARMAR N,et al. Attention Is All You Need[J]. Advances in Neural Information Processing Systems,2017(12):5999-6009.

[28] GU Y,YANG K,FU S,et al. Hybrid Attention based Multimodal Network for Spoken Language Classification[C]// Proceedings of the 27th International Conference on Computational Linguistics. ACM,2018:2379-2390.

[29] 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.

[30] SIMONYAN K,ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]// Proceedings of International Conference on Learning Representations. ICLR, 2015:1-14.

ZHANG Shao-qin, born in 1994, postgraduate. Her main research interests include data mining and natural language processing.

LI Tian-rui, born in 1969, Ph.D, professor, Ph.D supervisor, is a distinguished member of China Computer Federation. His main research interests include big data intelligence, rough sets and granular computing.