计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20200413
57(11): 2328-2336, 2020

MSRD:多模态网络谣言检测方法

刘金硕[1] 冯 阔[1] Jeff Z. Pan[2] 邓 娟[1] 王丽娜[1]
[1](空天信息安全与可信计算教育部重点实验室,武汉大学国家网络安全学院 武汉 430072)
[2](阿伯丁大学 苏格兰阿伯丁 AB24 3FX)
(liujinshuo@whu.edu.cn)

MSRD: Multi-Modal Web Rumor Detection Method

Liu Jinshuo[1], Feng Kuo[1], Jeff Z. Pan[2], Deng Juan[1], and Wang Lina[1]
[1](*Key Laboratory of Aerospace Information Security and Trusted Computing*, *Ministry of Education*, *School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430072)
[2](*University of Aberdeen*, *Aberdeen*, *Scotland AB24 3FX*)

Abstract The multi-modal web rumors that combine images and texts are more confusing and inflammatory, so they are more harmful to national security and social stability. At present, the web rumor detection work fully considers the text content of the essay in the rumor, and ignores the image content and the embedded text in the image. Therefore, this paper proposes a multi-modal web rumors detection method MSRD for the image, embedded text in the image and the text of the essay based on deep neural networks. This method uses the VGG-19 network to extract image content features, DenseNet to extract embedded text content, and LSTM network to extract text content features. After concatenating with the image features, the mean and variance vectors of the image and text shared representations are obtained through the fully connected layer, and the random variables sampled from the Gaussian distribution are used to form a re-parameterized multi-modal feature and used as the input of the rumor detector. Experiments show that the method achieves 68.5% and 79.4% accuracy on the two data sets of Twitter and Weibo.

Key words multimodal; rumor detection; inline text in image; natural language processing; deep neural network

摘 要 图像和文本相结合的多模态网络谣言由于更具迷惑性和煽动性,对国家安全和社会稳定的危害性更严重.目前网络谣言检测工作充分考虑了谣言中配文的文本内容而忽略了图像内容以及图像中的内嵌文本内容,因此,提出了一种基于深度神经网络针对图像、图像内嵌文本以及配文文本内容的多模态网络谣言检测方法 MSRD.该方法使用 VGG-19 网络提取图像内容特征,使用 DenseNet 提取图像内嵌文本内容,使用 LSTM 网络提取文本内容特征,与图像特征串接后,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以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.实验表明:该方法在 Twitter 和微博两大数据集上达到了 68.5%和 79.4%的准确率.

关键词 多模态;谣言检测;图像内嵌文本;自然语言处理;深度神经网络

中图法分类号 TP391

收稿日期:2020-06-10;**修回日期**:2020-07-24

基金项目:国家自然科学基金项目(U1936107,6187613,61672393)
This work was supported by the National Natural Science Foundation of China (U1936107, 6187613, 61672393).

通信作者:邓娟(dengjuan@whu.edu.cn)

网络谣言因受众广大、易于传播等特点,能够轻易煽动群众焦虑、恐慌等情绪,引发各类群体性事件,严重危害社会治安.在互联网时代,迅速普及的各类社交媒体平台成为了谣言滋生和传播的温床.为了博取关注、引导转发,网络谣言逐步由单纯的文本向图像与文本信息联合等多模态谣言转型.由于图像比文本更具有欺骗性,且其来源和真实性难以考证,因此图文联合的网络谣言更具危害性.因此,准确及时地针对图文联合的多模态网络谣言进行检测有利于维护社交平台舆情稳定、捍卫国家网络话语权和保证社会秩序平稳发展,具有十分重要的现实意义.

目前针对网络谣言的图像内容进行理解从而判别谣言的方法较为少见,已有的工作集中于识别谣言中的文本内容.其中一部分工作[1-3]采用以文本内容和用户信息为主要特征的基于特征构造检测的方法,另一部分工作[4-5]以传播时间、传播结构、语言特征等方面因素作为考量,提出基于传播结构检测法以及时间序列检测法.仅有少部分工作[6-7]考虑了图像信息,利用深度神经网络提取图像特征,并联合文本特征进行谣言判别.然而这些工作忽略了图像内嵌文本的处理,如图 1 所示.该谣言信息包含消息文本、图像和图像内嵌文本 3 部分.图像中的文本信息往往由谣言传播者人为添加,用于增加谣言的可信度.目前的谣言检测工作中缺乏对图像内嵌文本信息的挖掘,该部分对谣言判别具有重要意义.然而如何有效定位图像内嵌文本区域,并对其文本内容进行识别成为主要挑战.另外,如何有效融合文本特征与图像特征进行谣言检测成为另一挑战.

Guys MH370 spotted already y'all can chill now

Fig. 1 Multi-modal Web rumor

图 1 多模态网络谣言

为了解决以上问题,本文提出了一种端到端的多模态融合网络谣言检测方法.该方法融合了消息文本特征、图像内嵌文本特征以及图像内容特征,通过谣言检测器进行谣言判别.

概括来说,本文的主要贡献有 3 个方面:

1) 提出了一个联合消息文本信息以及图像内嵌文本信息与图像信息的多模态谣言检测模型 MSRD;

2) 提出了一种基于密集连接网络和空洞空间金字塔池化模型的图像文本定位方法;

3) 提出了一种融合文本特征与图像特征的共享特征表示方法.

1 相关工作

1.1 图像内嵌文本定位

目前国内外学者对图像内嵌文本定位工作的研究主要基于深度学习的方法.从技术路线角度主要分为 2 种:以连接文本提议网络(connectionist text proposal network,CTPN)[8]为代表的基于区域建议的文本检测方法,和以高效准确的场景文本检测器(efficient and accurate scene text detector,EAST)[9]为代表的基于图像分割的文本检测方法.CTPN 框架的缺点在于对非水平排列的文本区域定位效果不佳;EAST 检测器的缺点在于网络的感受野不够充足,导致对长文本区域定位效果不佳.以上工作为本文图像文本定位提供了新思路.

1.2 图像内嵌文本识别

不同于文本定位的粗略二分类任务,图像文本识别任务需要进行更精确的分类.2016 年牛津大学视觉几何组团队发表文献[10],使用卷积神经网络(convolutional neural network, CNN)率先在场景文本识别领域取得领先地位.2017 年文献[11]提出了基于卷积循环神经网络(convolutional recurrent neural network, CRNN)的文本检测框架,使用 CNN 提取图像特征,使用双向长短时记忆神经网络(bidirectional long short-term memory, BLSTM)提取序列信息,通过对特征编码输出预测结果,是 2017 年文本检测结果最好的端到端模型之一.2018 年众多研究者对已有模型的改进集中在应对任意方向排列的文本检测上,文献[12]提出了快速定向的文本识别系统(fast oriented text spotting, FOTS),文献[13]将"Textboxes"扩展成为"Textboxes++"以应对任意方向的文本区域.综上,目前基于深度学习的文本字符识别方法通用模式是使用 CNN 获取图像特征,使用 LSTM 获取序列特征,最后使用时序分类算法(connectionist temporal classification, CTC)输出识别结果.

1.3 谣言检测方法

已有的谣言检测工作从3个方面设计谣言检测模型:1)以文本内容和用户信息为主要特征的基于特征构造检测模型[1-3];2)以传播路径和传播节点为主要特征的基于传播结构检测模型;3)以文本信息随时间变化的统计特征为主要特征的基于时间序列检测模型[4-5].特征构造检测模型是谣言检测工作中应用最为广泛的典型模型.文献[1]提出了一种通过提取博文内容标签信息、外链信息和互相提及信息为主要特征的Twitter上误导性博文检测算法.文献[2]通过组合内容文本流行度、文本情感极性、用户影响力和博文转发率等特征构造微博谣言文本深层次特征来检测谣言.文献[3]基于博文文本内容的关键词汇来建立谣言检测模型.除特征构造检测模型之外,文献[4]认为谣言文本和非谣言文本在时间序列上变化的模式不同,并利用统计特征及特征在检测区间上的斜率变化等因素组成特征向量来检测谣言.文献[5]认为Twitter上谣言传播关键结构和语言差异的波动具有周期性,以传播时间、传播结构和语言特征3方面因素作为特征考量,建立周期时间序列模型用以识别谣言.

除了谣言文本信息,图像特征在谣言检测中起着非常重要的作用[14],文献[15]尝试提取图像的基本特征进行谣言检测,但是,手工提取的特征难以表示图像内信息的复杂分布.

与传统的特征工程相比,深度神经网络已被广泛应用于学习图像和文本的特征表示,并成功应用于各种工作,包括图像字幕[16-17]、多模态问题解答[18]和谣言检测[19-20]等.具体来说,卷积神经网络(CNN)广泛应用于图像的特征表示中[14-15],而递归神经网络(recurrent neural network, RNN)在编码文本信息中发挥强大的作用[21-22].文献[23]提出利用自编码器模型进行多模态网络谣言检测.文献[19]提出利用注意力模型来联合多模态特征进行谣言检测.然而,较少有工作考虑到图像中内嵌文本的特征提取问题,以及如何更加精确地联合文本与图像的多模态特征来进行谣言检测.

2 多模态网络谣言检测模型MSRD

多模态网络谣言检测模型MSRD的总体框架示意图如图2所示:

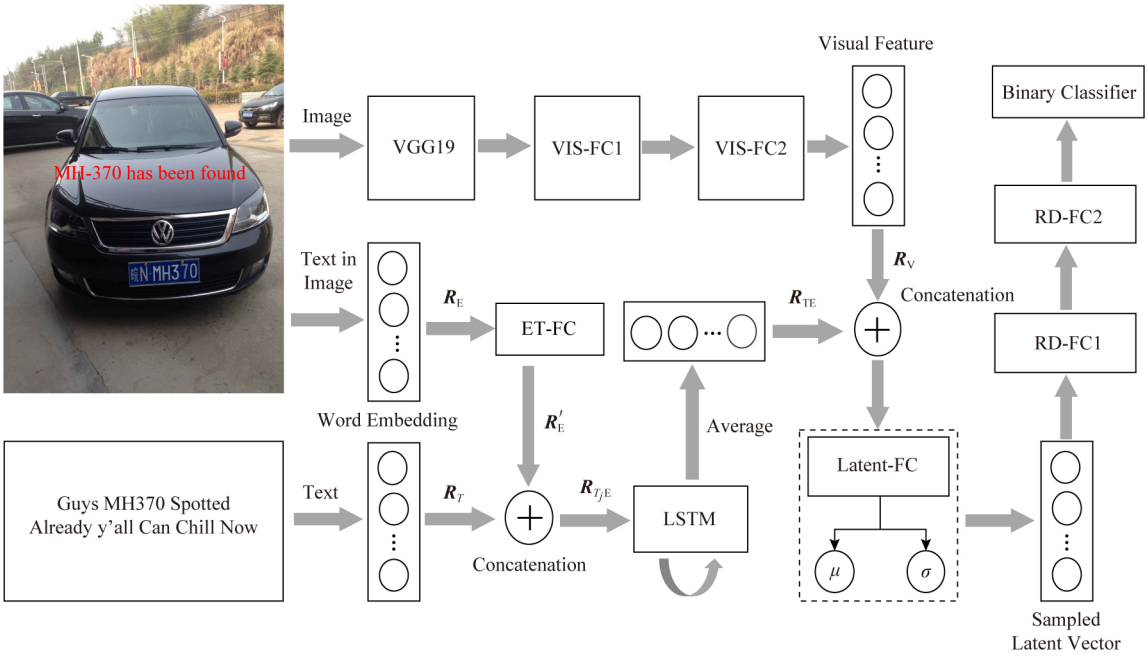

Fig. 2 Overall framework of the multimodal web rumor detection method

图2 多模态网络谣言检测方法总体框架

图2所示MSRD模型通过VGG19网络并添加2个全连接层VIS-FC1与VIS-FC2提取图像特征$\boldsymbol{R}_{\mathrm{V}}$,将向量化的内嵌文本信息$\boldsymbol{R}_{\mathrm{E}}$经过ET-FC全连接层转化为与配文文本向量维度相同的向量$\boldsymbol{R}'_{\mathrm{E}}$,与配文向量$\boldsymbol{R}_T$进行串接后通过LSTM网络以形成文本的联合表示$\boldsymbol{R}_{\mathrm{TE}}$,再与图像特征$\boldsymbol{R}_{\mathrm{V}}$串接后,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以形成

重新参数化的多模态特征,通过 2 个全连接层 RD-FC1 与 RD-FC2 后作为谣言检测器的输入进行谣言检测.

2.1 图像文本定位检测

谣言消息中的图像文本定位检测与自然场景图像中的文本定位检测工作是类似的,本文在文献[9]工作的基础上,提出了一种基于图像语义分割思想的文本定位检测方法,通过密集连接卷积网络(densely connected convolutional networks, DenseNet)DenseNet-121 提取特征,在不同的层级上抽取特征图(feature map),并且在特定尺度上引入空洞空间金字塔池化层(atrous spatial pyramid pooling, ASPP)扩大模型的感受野,然后从网络的顶部向下合并特征图,最终在输出层输出当前像素对应于原图中像素为文字的概率值.如果当前像素属于文字区域,输出该像素相对文本框 4 个顶点的坐标偏移值,最后通过非极大值抑制(non-maximum suppression)算法得到最终的文本框.图 3 表示了本文所用的图像文本定位检测算法结构及流程图.

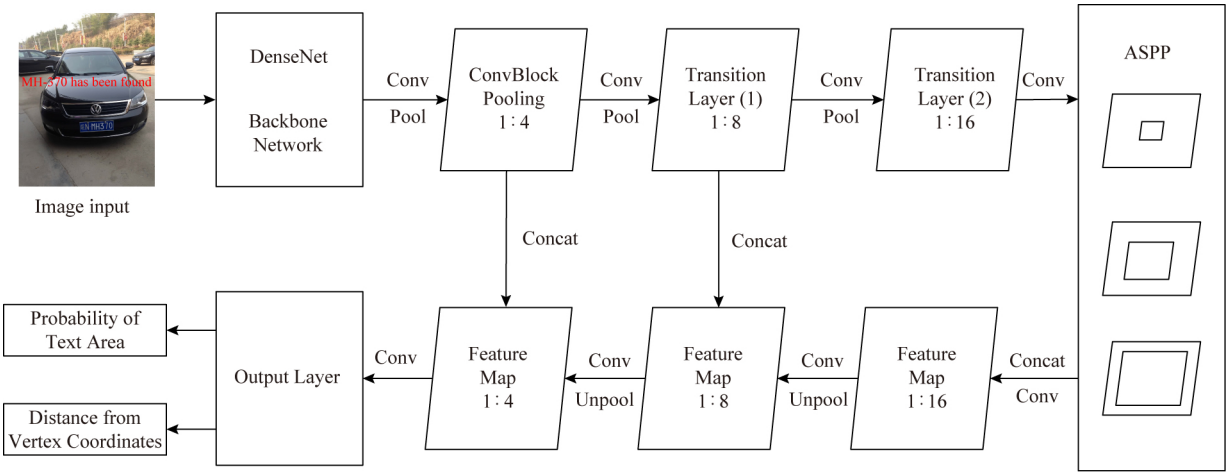

Fig. 3 Image text positioning algorithm structure and flow

图 3 图像文字定位算法结构及流程

2.1.1 空洞空间金字塔池化层的实现

ASPP 通过并行采用卷积核皆为 3×3 大小的同尺度、不同间距的多个滤波器来感知多尺度的视野,从而提取 Feature Map 上的多尺度特征.然后,将这些并行提取的特征通过使用 concat 操作串联起来,再通过 1×1 的卷积操作融合成指定深度 Feature Map 输出,这样就实现了在不改变输入数据体的尺寸规模的前提下,扩大了该网络层的感受野.

2.1.2 特征图合并策略

谣言消息中图像的文本区域大小不固定,需要同时兼顾神经网络模型的高层特征和低层特征,才能同时对不同尺寸的文本区域实现定位,因此需要将不同层级的 Feature Map 进行融合.

由图 3 可知,经过 ASPP 层输出的 Feature Map 尺寸比为 1∶16,由于在 ASPP 层已经经过 concat 操作,所以先使用 1×1 的卷积将 Feature Map 深度降维至 128,再使用 3×3 的卷积融合该层 Feature Map 中的特征.

卷积操作完成后使用反池化操作,使该层 Feature Map 的宽度和高度值与 Transition Layer (1)的输出数据体相匹配,并对这二者使用 concat 操作得到尺寸比为 1∶8的 Feature Map.

在尺寸比为 1∶8的 Feature Map 中,先使用 1×1 的卷积将深度降维至 64,再使用 3×3 的卷积融合特征后,经过反池化操作并与主干网络第一个 Pooling 层的输出 concat 串联起来得到 1∶4 的 Feature Map.

最后将 1∶4的 Feature Map 使用 1×1 的卷积将深度降维至 32,并使用 3×3 的卷积得到深度为 32 的最终输出层.最终输出层的尺寸与原图像的尺寸比也为 1∶4.

2.2 文本信息识别

本节进行的文本识别任务在 2.1 节图像文字定位的基础上完成,即通过上节对图像中文字区域的检测,截取原图像中含有文本的区域,输入给本节提出的文本识别模型.与其他的基于卷积循环神经网络的文本识别模型不同,由于已经完成了文本区域的定位检测工作,因此只需关注识别任务.由于截取

的文本区域中文本信息为主体部分,本文认为密集连接网络由于各隐藏层密集互联的特性,能够较好地完成全局特征的感知,不需要额外使用循环神经网络以兼顾文字序列前后信息.因此设计了本节基于密集连接网络的文本识别模型.

模型如图 4 所示.本节使用精简的密集连接网络,在 DenseNet-121 的基础上,去掉了第 4 个稠密块(dense block)以及最后的全局平均池化层,目的是尽可能保留中文汉字在垂直方向的笔画信息.由全连接层输出预测结果,将文本的识别任务视作分类任务,其类别数目等于预先构建的字典中字符数目.字典中字符经过排列预先编号,全连接层输出文字所属各类别的概率,通过 softmax 激活函数归一化后输出概率最大的类别编号,查阅字典即可得到预测的文本信息.由于 CTC 算法引入了空白标签,可以解决没有事先对齐的序列化数据训练问题,因此可加在密集连接网络后,对不定长的文本区域进行预测识别.

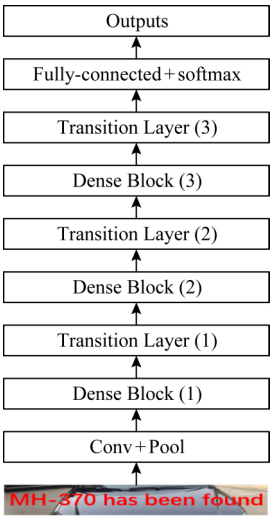

Fig. 4 Text recognition model

图 4 文字识别模型

2.3 LSTM 网络

本文使用带有长短期记忆(long short-term memory, LSTM)单元的 RNN 来学习提出的模型中消息文本和图像中文本的联合表示.RNN 是一种前馈神经网络,可用于对长度可变的顺序信息进行建模.给定输入序列$(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_M)$,基本的 RNN 模型更新隐藏状态$(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_M)$并生成输出向量$(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_M)$.$M$ 取决于输入的长度.当前的隐藏状态是使用循环单元计算的,循环单元采用最后的隐藏状态和当前的输入以产生当前的隐藏状态.

为了在学习长距离时间相关性时处理梯度的消失或爆炸问题[24-25],LSTM 通过将信息长时间存储在精心设计的存储单元中来扩展基本 RNN.具体而言,LSTM 中的读写存储单元 $\boldsymbol{c}$ 由一组 sigmoid 门控制:输入门 $\boldsymbol{i}$、输出门 $\boldsymbol{o}$ 和遗忘门 $\boldsymbol{f}$.对于每个时间步长 m,LSTM 单元从当前输入 $\boldsymbol{x}_m$,先前的隐藏状态 $\boldsymbol{h}_m$ 和先前的存储单元 $\boldsymbol{c}_m$ 接收输入.这些门的更新为[26-27]:

$$\boldsymbol{i}_m=\sigma(\boldsymbol{W}_{x_i}\boldsymbol{x}_m+\boldsymbol{W}_{h_i}\boldsymbol{h}_{m-1}+\boldsymbol{b}_i),\tag{1}$$

$$\boldsymbol{f}_m=\sigma(\boldsymbol{W}_{x_f}\boldsymbol{x}_m+\boldsymbol{W}_{h_f}\boldsymbol{h}_{m-1}+\boldsymbol{b}_f),\tag{2}$$

$$\boldsymbol{o}_m=\sigma(\boldsymbol{W}_{x_o}\boldsymbol{x}_m+\boldsymbol{W}_{h_o}\boldsymbol{h}_{m-1}+\boldsymbol{b}_o),\tag{3}$$

$$\boldsymbol{g}_m=\phi(\boldsymbol{W}_{x_c}\boldsymbol{x}_m+\boldsymbol{W}_{h_c}\boldsymbol{h}_{m-1}+\boldsymbol{b}_c),\tag{4}$$

$$\boldsymbol{c}_m=\boldsymbol{f}_m\odot\boldsymbol{c}_{m-1}+\boldsymbol{i}_m\odot\boldsymbol{g}_m,\tag{5}$$

$$\boldsymbol{h}_m=\boldsymbol{o}_m\odot\phi(\boldsymbol{c}_t),\tag{6}$$

其中,$\boldsymbol{W}^*$ 是对应门的权重矩阵,$\boldsymbol{b}^*$ 是偏差项,可以从网络中获知.σ 是 sigmoid 激活函数,$\sigma(x)=1/(1+\exp(-x))$;ϕ 是双曲正切函数,$\phi(x)=(\exp(x)-\exp(-x))/(\exp(x)+\exp(-x))$.$\odot$ 表示 2 个向量之间的逐元素乘法. 输入门 $\boldsymbol{i}$ 决定将新存储器添加到存储单元的程度.遗忘门 $\boldsymbol{f}$ 确定遗忘现有存储器的程度.通过忘记一部分现有存储器并添加新的存储器 $\boldsymbol{g}$ 来更新存储器单元 $\boldsymbol{c}$.

2.4 消息文本和图像中文本的联合表示

文本内容是消息中包含单词的集合:$T=\{T_1,T_2,\cdots,T_n\}$(n 是文本中单词的数量).文本中的每个单词 $T_j\in T$ 都表示为单词嵌入向量.每个单词的嵌入向量都是通过 Word Embedding 获得的,该网络在给定的数据集上进行了无监督预训练.

传统的多模态谣言检测往往只单独提取文本与图像特征,忽略了图像中的内嵌文本,本文认为该文本对谣言检测有一定的作用,因此本文通过 2.1~2.2 节中提到的图像文本定位与识别方法将图像中内嵌文本提取出来,同样采用 Word Embedding 形成初始的图像文本表示 $\boldsymbol{R}_{\mathrm{E}}=(p_1,p_2,\cdots,p_k)^{\mathrm{T}}$($k$ 是图像文本中单词的数量).图像文本特征 $\boldsymbol{R}_{\mathrm{E}}$ 通过图 2 所示的 ET-FC 全连接层转换为与配文文本向量相同维度的表示形式:

$$\boldsymbol{R}'_{\mathrm{E}}=\boldsymbol{W}_{\mathrm{ef}}\boldsymbol{R}_{\mathrm{E}},\tag{7}$$

其中,$\boldsymbol{W}_{\mathrm{ef}}$是维度转换全连接层的权重,在每个时间步,LSTM 取 $\boldsymbol{R}_{T_j\mathrm{E}}=(\boldsymbol{R}_{T_j},\boldsymbol{R}'_{\mathrm{E}})$作为输入,即第 j 个单词的嵌入 $\boldsymbol{R}_{T_j}$ 和转换后的图像文本特征 $\boldsymbol{R}'_{\mathrm{E}}$ 的串联.对 LSTM 输出的每个单词的表示取平均,以形成文本和图像文本 $\boldsymbol{R}_{\mathrm{TE}}$的联合表示.

2.5 图像特征表示

视觉子网络(图 2 的上部分支)采用图像作为输入,并构造视觉神经元提取图像特征.该网络前面部分采用和 VGG-19 网络相同的结构,并添加 2 个全连接层,将每个图像表示为 $\boldsymbol{R}_{\mathrm{V}}=(v_1,v_2,\cdots,v_{32})^{\mathrm{T}}$,共同训练整个谣言检测网络,挖掘消息图像中的视觉信息.其中视觉子网络 VGG-19 可以首先通过 ImageNet 数据集进行预训练,但是,在与 LSTM 子网络的联合训练过程中,冻结 VGG 网络的参数,只有最后的 2 个完全连接层的参数会更新,以提高训练效率,防止参数爆炸.

$$\boldsymbol{R}_{\mathrm{V}}=\boldsymbol{W}_{\mathrm{vf2}}\psi(\boldsymbol{W}_{\mathrm{vf1}}\boldsymbol{R}_{\mathrm{Vp}}),\tag{8}$$

其中,$\boldsymbol{R}_{\mathrm{Vp}}$是从预训练的 VGG 网络得到的图像特征表示,$\boldsymbol{W}_{\mathrm{vf1}}$是具有 ReLU 激活函数的第 1 个完全连接层中的权重,$\boldsymbol{W}_{\mathrm{vf2}}$是具有 softmax 功能的第 2 个完全连接层中的权重,ψ 表示 ReLU 激活函数.

在模型中直接使用图像特征与文本-图像文本特征进行训练的问题是:其中一种特征表示可能会压制另一种特征表示,这样便无法发挥多模态的融合作用,因此,在下面的部分,介绍一种特征的共享表示,以更好地融合文本与图像特征的联合表示.

2.6 共享表示

将联合文本特征表示 $\boldsymbol{R}_{TE}$和图像特征表示 $\boldsymbol{R}_{\mathrm{V}}$ 串接起来,通过图 2 中的完全连接层 Latent-FC,从中获得 2 个向量 $\boldsymbol{\mu}$ 和 $\boldsymbol{\sigma}$,它们可以分别视为共享表示的分布的均值和方差,从高斯分布中采样随机变量 ε.用 $\boldsymbol{R}_m$ 表示最终重新参数化的多模态特征:

$$\boldsymbol{R}_m=\boldsymbol{\mu}+\boldsymbol{\sigma}\circ\varepsilon.\tag{9}$$

本文将整个特征提取网络表示为 $G_{\mathrm{fr}}(m,\theta_{\mathrm{fr}})$,其中 θ_{fr}表示特征提取网络中所有学习的参数,用 m 表示消息集合中一条待判别的消息,则:

$$\boldsymbol{R}_m=G_{\mathrm{fr}}(m,\theta_{\mathrm{fr}}).\tag{10}$$

2.7 谣言检测器

谣言检测器将图像与文本特征的共享表示特征 $\boldsymbol{R}_m$ 作为输入,用来判别消息是否为谣言.它由具有激活函数的多个完全连接层组成,我们将谣言检测器表示为 $G_{\mathrm{rd}}(\boldsymbol{R}_m,\theta_{\mathrm{rd}})$,其中 θ_{rd}表示谣言检测器中所有的参数,谣言检测器的输出是该消息是谣言的概率.

$$\bar{y}_m=G_{\mathrm{rd}}(\boldsymbol{R}_m,\theta_{\mathrm{rd}}).\tag{11}$$

将 $\bar{y}_m$ 的值视为标签 1 表示消息 m 是假的,否则为 0.为了将输出值限制在 0 到 1 之间,我们使用 Sigmoid 逻辑函数.因此,为了计算分类损失,我们采用交叉熵:

$$L_{\mathrm{rd}}(\theta_{\mathrm{fr}},\theta_{\mathrm{rd}})=-E_{(m,y)\sim(M,Y)}[y\log(\bar{y}_m)+(1-y)\log(1-\bar{y}_m)],\tag{12}$$

其中,M 表示消息集合,Y 表示对应的真实标签,我们通过寻找最优参数 θ_{fr}^* 和 θ_{rd}^* 来最小化分类损失,表示如下:

$$(\theta_{\mathrm{fr}}^*,\theta_{\mathrm{rd}}^*)=\arg\min L_{\mathrm{rd}}.\tag{13}$$

3 实 验

本节,首先介绍了实验中所用到的 2 种数据集;然后介绍了实验中网络模型的参数设置;最后,为了验证 MSRD 方法可行性和先进性,设计的实验包括:单文本检测方法 Textual、单图像检测方法 Visual、图像加图像中文字检测方法 TVisual 和当前的谣言检测方法对比的实验.

3.1 数据集

鉴于结构化多媒体数据的稀疏可用性,本文利用 2 个标准数据集来评估我们用于谣言检测的网络模型 MSRD.这 2 个数据集包括从 Twitter 和微博收集的真实社交媒体信息.这些是仅有的具有成对的图像和文本信息的数据集.

3.1.1 Twitter 数据集

作为 MediaEval[21] 的一部分,该数据集分为 2 部分:开发集(9 000 条谣言 tweet,6 000 条真实消息 tweet)和测试集(2 000 条 tweet).考虑到本文只关心图像和文本信息,本文会过滤掉所有带有视频的推文.本文将开发集用于训练,将测试集用于测试,以保持与基准相同的数据拆分方案.

3.1.2 微博数据集

在文献[19]中用于谣言检测的微博数据集包括从中国权威新闻来源新华社和中国网站微博收集的数据.数据集的使用及预处理类似于文献[19]的方法进行.初步步骤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.然后,将数据集分成训练集和测试集,如 Jin 等人所述,数据比例约为 4∶1[19].

3.2 实验设置

对于文本特征,采用基于神经网络的分布式表示[29].对于这 2 个数据集,在标准文本预处理之后,使用默认参数设置,以无监督的方式对整个数据集进行 Word2Vec 模型的预训练.将数据集中的每个单词转化为 32 维嵌入向量.选择词嵌入表示而不是独热编码表示法的原因是,当独热编码表示法中的词汇量太大时,文本不足会导致文本特征差.

对于图像特征,使用在 ImageNet 集上预训练的 19 层 VGGNet 的第 2 层到最后一层的输出[30].从 VGG-19 获得的特征维度为 4 096.冻结 VGG 网络的权重,不再进行调整.

文本特征提取子网络由 LSTM 组成,隐藏层的尺寸大小为 32,使用的完全连接层的大小为 32.图像特征提取子网络在 VGG 网络后由大小为 1 024 和 32 的 2 个完全连接层组成.最终的谣言检测器有 2 个完全连接层,大小分别为 64 和 32.

在整个网络的训练中,我们使用 128 个实例的批处理大小.该模型训练了 300 个 epoch,模型学习率为 10^{-5},并使用了早停法.为了防止过度拟合,本文在模型的权重上使用 L2 正则.为了模型寻找最优参数,使用 Adam[31]作为优化器.

3.3 有效性和先进性实验

为了验证本文所提出的多模态谣言检测模型 MSRD 的性能,本文在比对实验中选择了单文本检测方法 Textual、单图像检测方法 Visual、图像加图像中文字检测方法 TVisual 和 3 种较新的谣言检测方法.其中,Textual,Visual 和 TVisual 这 3 种方法分别为本文所提出网络模型 MSRD 的子网络,结构不再赘述.下面介绍 3 种较新的用于对比实验的谣言检测方法.

1) VQA[18].VQA 模型旨在回答有关给定图像的问题.本文将最初为多分类任务设计的 Visual QA 模型适应了我们的二分类任务.本文通过用二分类层替换最终的多分类层来完成,使用一层 LSTM,隐藏单元数设置为 32.

2) att-RNN[19].att-RNN 使用注意力机制来组合文本、视觉和社交上下文特征.在此端到端网络中,图像特征被合并到使用 LSTM 网络获得的文本和社交环境的联合表示中.LSTM 网络输出后,注意力模型是融合视觉特征必不可少的部分.为了公平比较,在本文的实验中删除了处理社交环境信息的部分.

3) EANN[20].事件对抗神经网络(EANN)由 3 个主要组件组成:多模式特征提取器、假新闻检测器和事件鉴别器.多模式特征提取器从帖子中提取文本和视觉特征.它与假新闻检测器一起使用,事件鉴别器负责删除任何特定于事件的功能.还可以仅使用 2 个组件(多模式特征提取器和假新闻检测器)来检测假新闻.因此,为了公平比较,在实验中使用 EANN 的变体,其中不包含事件鉴别器.

4 实验结果与分析

表 1 汇总了 2 个数据集的基线以及 MSRD 方法的结果.可以清楚地看到,MSRD 的性能要优于基线方法.

Table1 Comparison of Experimental Results Between MSRD Model and Other Methods

表 1 MSRD 模型与其他方法对比实验结果

Dataset	Method	Accuracy	Precision	Recall	$F1$
Twitter	Textual	0.504	0.572	0.546	0.559
	Visual	0.575	0.604	0.554	0.578
	TVisual	0.592	0.611	0.572	0.591
	VQA	0.631	0.765	0.509	0.611
	att-RNN	0.664	0.749	0.615	0.676
	EANN	0.648	**0.810**	0.498	0.617
	MSRD	**0.685**	0.725	**0.636**	**0.678**
Weibo	Textual	0.612	0.624	0.563	0.592
	Visual	0.622	0.578	0.591	0.585
	TVisual	0.635	0.632	0.608	0.620
	VQA	0.736	0.797	0.634	0.706
	att-RNN	0.772	0.854	0.656	0.742
	EANN	0.782	0.827	0.697	0.756
	MSRD	**0.794**	**0.854**	**0.716**	**0.779**

Note: The bold numbers are the best performance in experiments.

在 Twitter 数据集上,单独检测图像判断谣言效果要优于单独检测文本,而检测图像及图像中的文本的效果略高于单独检测图像,这说明挖掘图像中文本信息是具有一定意义的,这可能会给整体的谣言检测器的准确率带来提升.在已有多模态谣言检测模型中,att-RNN 模型优于 EANN,其表明注意力机制可以通过考虑与文本相关的图像部分来帮助改善模型的性能.本文提出的 MSRD 模型更好地融合了文本与图像特征,使谣言检测结果优于基线模型,将准确性从 66.4%提高到 68.5%,并将 $F1$ 分数从 67.6%增加到 67.8%.

在微博数据集上,由表 1 的结果中可以看到相似的趋势.多模态模型 EANN 和 att-RNN 的性能要优于单模态模型和 VQA.MSRD 模型的性能要优于所有基准,并且将准确性从 78.2%提高到79.4%,并且与以前的最佳基准相比,$F1$ 分数从 75.6%提高到 77.9%.这验证了 MSRD 方法在检测社交媒体上的多模态网络谣言方面的有效性、先进性和鲁棒性.

5 总结与展望

本文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 MSRD,该模型采用密集连接网络和空洞空间金字塔池化方法对图像文本进行定位,采用共享特征方法将文本特征与图像特征进行了较好地融合表示用于谣言检测.在 Twitter 和微博两大数据集上进行了实验验证,实验结果表明:挖掘图像中的文本信息对谣言检测具有一定的作用,MSRD 模型要优于基线模型.在未来的研究中,我们应考虑谣言信息传播过程中的机制以及用户特征信息等问题.

参 考 文 献

[1] Ratkiewicz J, Conover M D, Meiss M, et al. Detecting and tracking political abuse in social media [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2011: 1-8

[2] Zhang Qiao, Zhang Shuiyuan, Dong Jian, et al. Automatic detection of rumor on social network [C] //Proc of the 4th CCF Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2015: 113-122

[3] Zhao Zhe, Resnick P, Mei Qiaozhu. Enquiring minds: Early detection of rumors in social media from enquiry posts [C] //Proc of the 24th Int Conf on World Wide Web. New York: ACM, 2015: 1395-1405

[4] Ma Jing, Gao Wei, Wei Zhongyu, et al. Detect rumors using time series of social context information on microblogging websites [C] //Proc of the 24th ACM Int on Conf on Information and Knowledge Management. New York: ACM, 2015: 1751-1754

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

[6] Sun Shengyun, Liu Hongyan, He Jun, et al. Detecting event rumors on sina weibo automatically [C] //Proc of Asia-Pacific Web Conf. Berlin: Springer, 2013: 120-131

[7] Jin Zhiwei, Cao Juan, Zhang Yongdong, et al. Novel visual and statistical image features for microblogs news verification [J]. IEEE Transactions on Multimedia, 2017, 19(3): 598-608

[8] Tian Zhi, Huang Weilin, He Tong, et al. Detecting text in natural image with connectionist text proposal network [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2016: 56-72

[9] Zhou Xinyu, Yao Cong, Wen He, et al. EAST: An efficient and accurate scene text detector [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5551-5560

[10] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading text in the wild with convolutional neural networks [J]. International Journal of Computer Vision, 2016, 116(1): 1-20

[11] Shi Baoguang, Bai Xiang, Yao Cong. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2298-2304

[12] Liu Xuebo, Liang Ding, Yan Shi, et al. Fots: Fast oriented text spotting with a unified network [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5676-5685

[13] Liao Minghui, Shi Baoguang, Bai Xiang. Textboxes++: A single-shot oriented scene text detector [J]. IEEE Transactions on Image Processing, 2018, 27(8): 3676-3690

[14] Wu Ke, Yang Song, Zhu K Q. False rumors detection on sina weibo by propagation structures [C] //Proc of the 31st Int IEEE Conf on Data Engineering. Piscataway, NJ: IEEE, 2015: 651-662

[15] Tian Dongping. A review on image feature extraction and representation techniques [J]. Journal of Multimedia and Ubiquitous Engineering, 2013, 8(4): 385-396

[16] Karpathy A, Li Feifei. Deep visual-semantic alignments for generating image descriptions [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3128-3137

[17] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3156-3164

[18] Agrawal A, Lu Jiasen, Antol S, et al. VQA: Visual question answering [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 2425-2433

[19] Jin Zhiwei, Cao Juan, Guo Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs [C] //Proc of the 2017 ACM on Multimedia Conf. New York: ACM, 2017: 795-816

[20] Wang Yaqi, Ma Fenglong, Jin Zhiwei, et al. EANN: Event adversarial neural networks for multiModal fake news detection [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 849-857

[21] Boididou C, Papadopoulos S, Dang-Nguyen D, et al. Verifying multimedia use at mediaEval 2016 [C] //Proc of the MediaEval 2016 Workshop. [2020-03-01]. http://ceur-ws.org/Vol-1739/MediaEval_2016_paper_3.pdf

[22] Shu Kai, Sliva A, Wang Suhang, et al. Fake news detection on social media: A data mining perspective [C] //Proc of ACM SIGKDD Explorations Newsletter. New York: ACM, 2017: 22-36

[23] Dhruv K, JaiPal S, Manish G, et al. MVAE: Multimodal variational autoencoder for fake news detection [C] //Proc of the 2019 World Wide Web Conf. New York: ACM, 2019: 2915-2921

[24] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166

[25] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks [C] //Proc of the Int Conf on Machine Learning. New York: ACM, 2013: 1310-1318

[26] Gers F, Schraudolph N, Schmidhuber J. Learning precise timing with lstm recurrent networks [J]. Journal of Machine Learning Research, 2002, 3(2): 115-143

[27] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780

[28] Ma Jing, Gao Wei, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence. New York: ACM, 2016: 3818-3824

[29] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [C] //Proc of the 26th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 3111-3119

[30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C] //Proc of the 3rd Int Conf on Learning Representations. New York: ACM, 2015: 358-406

[31] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint, arXiv: 1412.6980, 2014

Liu Jinshuo, born in 1974. PhD, associate professor. Member of CCF. Her main research interests include cyber security and data mining.

Feng Kuo, born in 1996. Master candidate. His main research interests include public opinion analysis and rumor detection.

Jeff Z. Pan, born in 1974. PhD, professor. His main research interests include knowledge representation, artificial intelligence and data science.

Deng Juan, born in 1976. PhD, associate professor. Member of CCF. Her main research interests include artificial intelligence, machine learning and high performance computing.

Wang Lina, born in 1964. PhD, professor. Member of CCF. Her main research interests include system security and steganalysis.