计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200804
58(7): 1456-1465, 2021

语义增强的多模态虚假新闻检测

亓 鹏 曹 娟 盛 强
(中国科学院智能信息处理重点实验室(中国科学院计算技术研究所) 北京 100190)
(中国科学院计算技术研究所 北京 100190)
(中国科学院大学 北京 100049)
(qipeng@ict.ac.cn)

Semantics-Enhanced Multi-Modal Fake News Detection

Qi Peng, Cao Juan, and Sheng Qiang
(*Key Laboratory of Intelligent Information Processing of Chinese Academy of Science* (*Institute of Computing Technology, Chinese Academy of Sciences*), *Beijing* 100190)
(*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
(*University of Chinese Academy of Sciences, Beijing* 100049)

Abstract In recent years, social media has become the main access where people acquire the latest news. However, the convenience and openness of social media have also facilitated the proliferation of fake news. With the development of multimedia technology, fake news on social media has been evolving from text-only posts to multimedia posts containing images or videos. Therefore, multi-modal fake news detection is attracting more and more attention. Existing methods for multi-modal fake news detection mostly focus on capturing appearance-level features that are highly dependent on the dataset distribution but insufficiently exploit the semantics-level features. Thus, the methods often fail to understand the deep semantics of textual and visual entities in the fake news, which indeed limits the generalizability of models in real applications. To tackle this problem, this paper proposes a semantics-enhanced multi-modal model for fake news detection, which better models the underlying semantics of multi-modal news by implicitly utilizing the factual knowledge in the pre-trained language model and explicitly extracting the visual entities. Furthermore, the proposed method extracts visual features of different semantic levels and models the semantic interaction between the textual and visual features by the text-guided attention mechanism, which better fuses the multi-modal heterogeneous features. Extensive experiments on the Weibo dataset strongly evidence that our method outperforms the state of the art significantly.

Key words social media; fake news detection; multi-modal; knowledge fusion; attention mechanism

摘 要 近年来社交媒体逐渐成为人们获取新闻信息的主要渠道,但其在给人们带来方便的同时也促进了虚假新闻的传播.在社交媒体的富媒体化趋势下,虚假新闻逐渐由单一的文本形式向多模态形式转变,因此多模态虚假新闻检测正在受到越来越多的关注.现有的多模态虚假新闻检测方法大多依赖于和数据集高度相关的表现层面特征,对新闻的语义层面特征建模不足,难以理解文本和视觉实体的深层

收稿日期:2020-10-09;**修回日期**:2020-12-15
基金项目:国家自然科学基金重点项目(U1703261)
This work was supported by the Key Program of the National Natural Science Foundation of China (U1703261).
通信作者:曹娟(caojuan@ict.ac.cn)

语义,在新数据上的泛化能力受限.提出了一种语义增强的多模态虚假新闻检测方法,通过利用预训练语言模型中隐含的事实知识以及显式的视觉实体提取,更好地理解多模态新闻的深层语义.提取不同语义层次的视觉特征,在此基础上采用文本引导的注意力机制建模图文之间的语义交互,从而更好地融合多模态异构特征.在基于微博新闻的真实数据集上的实验结果表明:该方法能够有效提高多模态虚假新闻检测的性能.

关键词　社交媒体;虚假新闻检测;多模态;知识融合;注意力机制

中图法分类号　TP391

中国社会科学院2020年发布的《中国新媒体发展报告No.11》[1]显示,以微信、微博等为代表的社交媒体已经成为我国公众获取新闻信息的主要渠道.社交媒体的实时性、开放性、便捷性和双向性使得人们可以快速地获取并传播信息.但与此同时,社交媒体低门槛的特点也促进了虚假信息尤其是虚假新闻在网络空间的滋长蔓延.网络虚假新闻不仅使受众深受其害,冲击了主流媒体的权威性和公信力,还产生了经济、政治等多个方面的风险隐患①.近年来,在社交媒体的富媒体化趋势下,用户发布的内容由纯文本向图文并茂的多媒体形式转变.虚假新闻的发布者也开始利用一些极具误导性甚至经过篡改的图片来吸引读者的注意,进一步促进虚假新闻的传播[2].因此,基于社交媒体的多模态虚假新闻检测已经成为近年来的研究热点.

现有研究表明:虚假新闻在表现层面上与真实新闻具有显著的差异性.虚假新闻往往呈现出更加强烈的情感煽动性、主观性[3-4],经常出现"紧急通知""快转"等高频短语;虚假新闻图片具有低质量、视觉冲击力强的特点[5-6].相比下,真实新闻往往更加客观严谨,配图质量更高.现有的多模态方法[7-9]一般采用通用的循环神经网络(recurrent neural network,RNN)和卷积神经网络(convolutional neural network,CNN)分别捕捉虚假新闻文本及视觉模态表现层面的特性.然而,虚假新闻表现层面的特性与数据集高度相关,这使得在特定数据集上性能不错的方法往往难以良好泛化到新数据集上,容易误判表现层特性不明显的假新闻.

事实上,对于虚假新闻检测任务而言,仅仅关注新闻是如何表述的,即新闻表现层面的特点是不够的,还应该关注新闻具体描述了什么内容,即新闻语义层面的特点.在语义层面上,虚假新闻往往会涉及一些极具争议性的话题,或者存在图文不符等现象.与表现层相比,虚假新闻语义层面的特点往往更难捕获.一方面,新闻作为一种特殊的叙事文体,往往包含人名、地名、机构名及其他专有名词等命名实体.理解这些实体对建模虚假新闻语义层面的特点起到重要的作用,但他们的含义难以简单地通过上下文理解,需要引入外部事实知识.另一方面,在多模态新闻的语义理解中,图片模态经常提供有利于模型预测的关键实体信息(名人、地标、旗帜标志等).例如我们可以通过核对图文中出现人物身份的一致性推断该新闻的可信度.然而,通用的视觉特征表示大多停留在感知层面,无法找到并充分建模这些视觉实体背后的深层语义.另外,通用的视觉语义特征和文本语义处于不同的特征空间,存在语义鸿沟和特征异构的问题.因此,如何充分建模图文之间的语义交互,也是我们需要着重考虑的问题.

为了解决上述挑战,我们提出了一种语义增强的多模态虚假新闻检测方法.首先,我们利用预训练语言模型中隐含的大量的事实知识,更好地理解多模态新闻中的实体概念;其次,在提取通用的视觉特征向量的基础上,利用外部模型显式提取新闻图片中的视觉实体及嵌入文字,得到不同语义层次的视觉特征;最后,我们采用文本引导的注意力机制建模文本与不同层次的视觉特征之间的语义交互,进而得到统一的多模态特征表达.

本文的主要贡献包括3个方面:

1) 提出了新颖的语义增强的多模态虚假新闻检测方法.通过融合外部知识以及显式的视觉实体提取,更好地理解多模态新闻中的实体语义,从而更充分地挖掘多模态虚假新闻的语义线索;

2) 采用文本引导的注意力机制建模文本与不同层次的视觉特征之间的语义交互,更好地融合多模态异构特征;

3) 在真实世界的微博数据集上对本文提出的方法进行验证.与当前较好方法相比,我们的模型能够大幅提高虚假新闻检测的准确率.

① http://www.cac.gov.cn/2020-01/23/c_1581318267502085.htm

1 相关工作

根据研究对象的不同,虚假新闻检测可以分为事件层面的检测和微博层面的检测.事件层面的检测利用同一事件下所有微博的信息联合判断该新闻事件的可信度.但是事件形成往往需要一定时间.一些重大的虚假新闻可能在事件形成前已经在社交媒体上广泛传播,在非常短的时间内产生较大的消极影响.微博层面的检测是指判断单条微博消息的可信度.与事件层面的检测相比,这种方法在实际应用中可以做到实时检测,因此得到了研究人员的广泛关注.本文的研究专注于微博层面的虚假新闻检测.

大多数现有的研究利用文本内容和传播过程中产生的社交上下文检测虚假新闻[10].基于文本内容的检测方法主要基于虚假新闻特定的语言风格建模,包括早期提取语言学特征、主题特征等手工特征的方法[11-13],以及近年来基于深度模型自动学习数据高层特征的方法[14].基于社交上下文的方法主要包括基于用户行为可信度的方法[15-17]以及基于传播网络的方法[18-21].

近年来,一些工作开始关注视觉模态在虚假新闻检测中的作用[5-6,22-26].虚假新闻图片主要包括篡改图片和误用图片两大类[6].篡改图片指使用工具故意进行像素级改动或是算法自动生成的非真实图片,而误用图片一般指未经刻意修改,取自其他事件或是图片内容被错误解读的真实图片.现有基于视觉模态的研究主要利用图片的取证特征[23]、语义特征[6]、分布特征[22]以及上下文特征[24-25]等进行虚假新闻检测.

文本模态和视觉模态为虚假新闻检测提供了各有侧重、相互补充的信息.因此,结合多模态信息进行虚假新闻检测的方法也备受关注.文献[7]第1次通过深度神经网络的方法将多模态信息引入到虚假新闻检测中,他们提出了一种带注意力机制的循环神经网络融合文本、视觉及社交上下文的信息.为提高模型在新的虚假新闻事件上的泛化性能;文献[8]利用对抗学习的方法,引入事件分类这一辅助任务,引导模型学习到更具泛化性能的与事件无关的多模态特征;文献[9]利用"编码器-解码器"结构来构建多模态新闻的特征表达.上述方法在多模态虚假新闻检测上具有一定的有效性,但是由于缺乏足够的事实知识,不能充分理解多模态新闻事件的深层语义.针对这一问题,文献[27]从外部知识图谱中提取文本实体对应的概念知识融入多模态的表达中,从而获得更好的语义理解能力;文献[28]提出利用图神经网络建模文本、知识以及图片中的物体之间的交互.上述方法通过引入外部知识图谱的方式增强对新闻文本语义的理解,但是在对图片语义信息建模以及多模态异构特征融合上仍存在欠缺之处.

因此,针对已有工作的不足,我们提出了一种语义增强的多模态虚假新闻检测方法,不仅能够利用外部知识深入理解文本及图片的语义信息,也能充分融合不同模态的异构特征.

2 语义增强的虚假新闻检测方法

我们的任务是判断给定的单条多模态新闻为真新闻或假新闻.图1展示了我们提出的语义增强的多模态虚假新闻检测模型,主要由文本语义编码器、视觉语义编码器、多模态特征融合以及分类4部分组成.

2.1 文本语义编码器

文本作为新闻事件的叙述主体,包含了丰富的信息,为新闻可信度的判定提供了不同层次的线索.现有方法大多利用循环神经网络等对输入文本的上下文信息进行建模,捕捉文本表现层的模式[7,9,14,27].然而,由于特征提取过程缺少相应事实知识的参与,这类方法对新闻文本中命名实体的理解能力有限,进而难以充分捕捉虚假新闻语义层面的线索.

近期一些工作[29-30]表明,以BERT[31](bidirectional encoder representations from transformers)为代表的预训练语言模型不仅具有强大的建模能力,通过在大规模预训练语料上的学习,其内部已经学习到了某些句法知识和常识知识.在BERT的基础上,百度提出了一种知识增强的语义表示模型ERNIE(enhanced representation from knowledge integration)[32].ERNIE的结构与BERT类似,都是利用多层的Transformer[33]作为基本的编码器,通过self-attention机制实现对上下文信息的建模.与BERT不同的是,ERNIE对词、实体等语义单元进行掩码,并扩展了一些知识类的中文语料进行预训练,能够更好地建模实体概念等先验语义知识,从而进一步提升模型的语义表示能力.ERNIE不仅能够作为上下文编码器产生句子的表达,还可以作为知识存储器,在产生句子表达的时候隐式地利用模型中存储的大量事实知识.因此,我们使用ERNIE作为文本模态的特征提取器,同时建模文本在表现层及语义层的特点.

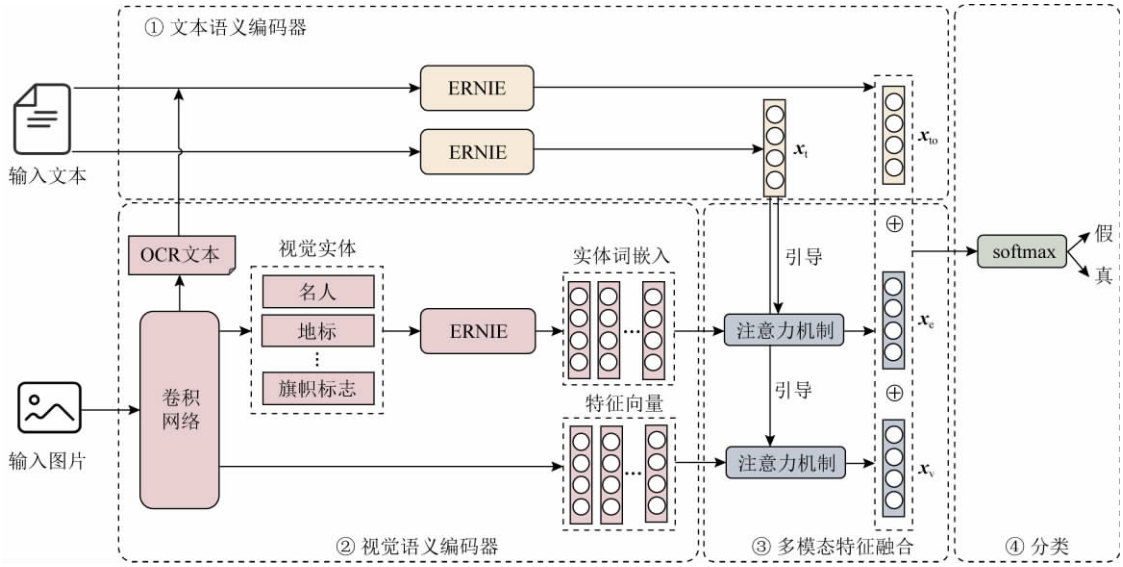

Fig. 1　Framework of our semantics-enhanced multi-modal fake news detection model

图 1　语义增强的多模态虚假新闻检测模型结构图

具体地,我们首先在虚假新闻分类任务的数据集上对 ERNIE 进行微调.对于输入句子 $T=[w_1, w_2, \cdots, w_n]$,其中 w_i 代表句子中的第 i 个词,ERNIE 会先对其进行编码,添加[MASK],[SEP],[CLS]等标记,然后进行训练.我们提取[CLS]对应的 768 维的特征向量作为输入句子的最终语义表示如式(1):

$$\boldsymbol{x}_t=\mathrm{ERNIE}(T),\ \boldsymbol{x}_t\in\mathbb{R}^{768}. \tag{1}$$

另外,社交媒体上存在很多以文字型图片为主体的新闻,即新闻的主要文本用图片的形式表示.我们使用百度预训练的 OCR 文字检测模型① 提取图片中的文本信息.经过数据预处理后,可以将图片中识别到的文本表示为词序列 $O=[w_1, w_2, \cdots, w_n]$,其中 w_i 表示句子的第 i 个词.为充分建模输入文本 T 与图片文本 O 的语义交互,我们将两者拼接成一个序列,用[SEP]进行分隔,输入到 ERNIE 网络中,得到对应的语义表示:

$$\boldsymbol{x}_{to}=\mathrm{ERNIE}(T[\mathrm{SEP}]O). \tag{2}$$

2.2　视觉语义编码器

与真实新闻的配图相比,虚假新闻图片往往具有更低的图片质量,更具视觉冲击和情感煽动的图片风格[6].因此,现有方法大多通过卷积神经网络提取颜色、边缘、纹理等层次化的视觉特征来建模图片的质量及风格特性.然而,由于缺乏外部知识的引入,这类通用的视觉特征表示大多停留在感知层面,无法充分建模新闻图片的深层语义.

事实上,新闻图片往往包含一些极具新闻性的视觉实体,包括名人、地标、旗帜标志以及一些敏感目标等.准确识别这些实体有助于我们更加充分地理解多模态新闻的语义,从而更好地捕捉虚假新闻的线索.例如,通过对图片进行名人及地标识别,可以发现图片中所展示的人物及地点与新闻文本描述不符;通过识别图片中的敏感标志及物体,可以强调文本中的相关实体,从而更好地理解多模态新闻的争议点.因此,为充分建模虚假新闻图片的语义特性,我们一方面提取图片的视觉特征向量建模其质量及风格特性,另一方面引入外部模型显式提取图片中的视觉实体建模其深层语义.

具体地,为了捕捉图片质量及风格上的特性,我们参照前人的工作[7-9,27],采用 VGG19[34] 网络来提取图片的视觉特征向量.一些前置实验表明,VGG19 在虚假新闻分类任务的图片数据集上比 ResNet[35],Inception[36] 等模型的表现更为稳定.考虑到图片中不同空间区域的信息密度以及重要程度不一致,我们对输入图片进行分块地特征提取.我们首先在虚假新闻分类任务的数据集上对基于 ImageNet[37] 预训练的 VGG19 进行微调.对于输入图片 I,我们可以从 VGG19 网络的最后一个卷积层得到大小为 $7\times7\times512$ 的特征图,并将其进一步表示为特征向量序列 $\boldsymbol{V}=[\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n]$,$\boldsymbol{v}_i\in\mathbb{R}^{512}$,$n=49$,其中 $\boldsymbol{v}_i$ 表示第 i 个图片块对应的视觉特征向量.

① https://ai.baidu.com/tech/ocr

为了准确识别图片中出现的名人、地标等命名实体,可以先在相应任务的标注数据集上预训练目标检测模型,再对本任务中的新闻图片进行检测.考虑到目前和上述任务相关的可用的大规模中文标注数据集比较少,我们采用百度 AI 平台提供的接口①行视觉实体识别.其中,名人检测模型可以识别中外著名政治及公众人物;地标检测模型可以识别中外著名地标;旗帜标志检测模型可以识别国旗国徽、党旗党徽、警徽、民族服饰以及各类反动组织的标志等;敏感目标检测模型可以识别枪械、军事武器、血腥、疾病表征、色情、暴恐、爆炸火灾、车祸等敏感的视觉目标.通过对输入图片 I 进行实体识别,我们可以得到对应的实体列表.为结合外部知识充分理解实体背后潜在的语义信息,我们将实体名称列表输入到 ERNIE 网络中,得到对应的实体表达序列 $\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n]$,$\boldsymbol{e}_i\in\mathbb{R}^{768}$,其中 $\boldsymbol{e}_i$ 表示图片中识别到的第 i 个视觉实体对应的语义表示.

2.3 多模态特征融合

至此,我们得到了文本的表达 $\boldsymbol{x}_{\mathrm{t}}$,文本及图片文本的联合表达 $\boldsymbol{x}_{\mathrm{to}}$、视觉实体序列的表达 $\boldsymbol{E}$ 以及视觉特征向量序列的表达 $\boldsymbol{V}$,本节将介绍如何融合上述多种异构特征得到一个统一的多模态表达.

图片中可能存在多个视觉实体,但并非所有检测到的实体都对虚假新闻分类的任务有帮助,融合所有的实体信息可能会导致信息冗余甚至引入噪声.经过观察,我们发现能够与文本对应的视觉实体往往更加重要.因此,我们对图片中识别到的多个视觉实体 $\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n]$ 进行基于文本引导的注意力机制的融合.我们首先根据文本特征 $\boldsymbol{x}_{\mathrm{t}}$,计算每个视觉实体 $\boldsymbol{e}_i$ 的重要性:

$$\mathcal{F}(\boldsymbol{x}_{\mathrm{t}},\boldsymbol{e}_i)=f(\boldsymbol{x}_{\mathrm{t}}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{e}_i),i\in[1,n],\tag{3}$$

其中,$\boldsymbol{W}$ 为随机初始化并在训练过程中联合优化的参数矩阵,$f(\cdot)$为激活函数.我们对权值进行归一化:

$$\alpha_{\boldsymbol{e}_i}=\frac{\exp(\mathcal{F}(\boldsymbol{x}_{\mathrm{t}},\boldsymbol{e}_i))}{\sum_{i=1}^{n}\exp(\mathcal{F}(\boldsymbol{x}_{\mathrm{t}},\boldsymbol{e}_i))},\tag{4}$$

并根据得到的权重对不同的视觉实体表示进行加权求和,得到最终的视觉实体表示:

$$\boldsymbol{x}_{\mathrm{e}}=\sum_{i=1}^{n}\alpha_{\boldsymbol{e}_i}\boldsymbol{e}_i.\tag{5}$$

同样地,图片的不同区域对于语义理解也具有不同的重要性.因此,我们对图片不同区域的特征向量 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n]$ 进行基于文本引导的注意力机制的融合,得到最终的视觉特征向量表示:

$$\mathcal{F}(\boldsymbol{x}_{\mathrm{t}},\boldsymbol{v}_i)=f(\boldsymbol{x}_{\mathrm{t}}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{v}_i),i\in[1,n],\tag{6}$$

$$\alpha_{\boldsymbol{v}_i}=\frac{\exp(\mathcal{F}(\boldsymbol{x}_{\mathrm{t}},\boldsymbol{v}_i))}{\sum_{i=1}^{n}\exp(\mathcal{F}(\boldsymbol{x}_{\mathrm{t}},\boldsymbol{v}_i))},\tag{7}$$

$$\boldsymbol{x}_{\mathrm{v}}=\sum_{i=1}^{n}\alpha_{\boldsymbol{v}_i}\boldsymbol{v}_i,\tag{8}$$

经过上述操作,我们得到了原始文本以及图片文字的联合表示 $\boldsymbol{x}_{\mathrm{to}}$,图片的视觉实体表示 $\boldsymbol{x}_{\mathrm{e}}$ 以及图片的视觉特征向量表示 $\boldsymbol{x}_{\mathrm{v}}$.这些特征从不同角度建模了输入的多模态新闻不同层次的语义信息,具有一定的互补性.我们将这些特征拼接在一起,得到该条新闻最终的多模态表示:

$$\boldsymbol{x}=\boldsymbol{x}_{\mathrm{to}}\oplus\boldsymbol{x}_{\mathrm{e}}\oplus\boldsymbol{x}_{\mathrm{v}},\tag{9}$$

其中,$\oplus$是拼接操作.

2.4 分 类

在得到输入新闻的多模态表示 $\boldsymbol{x}$ 之后,我们将其输入全连接层,并将全连接层的输出通过 softmax 层产生分类标签的分布:

$$\boldsymbol{p}=\mathrm{softmax}(\boldsymbol{W}_C\boldsymbol{x}+\boldsymbol{b}_C),\tag{10}$$

其中,$\boldsymbol{W}_C$ 和 $\boldsymbol{b}_C$ 是模型的参数.我们采用交叉熵作为模型的损失函数:

$$L=-\sum[y^f\log p^f+(1-y^f)\log(1-p^f)],\tag{11}$$

其中,y^f 是样本的真实标签,1 表示该样本为假新闻,0 表示该样本为真新闻;p^f 表示该样本被预测为假新闻的概率.

3 实验与分析

3.1 数据集

在目前的虚假新闻研究中,公开的多模态数据集比较少,故在本文的后续实验中主要讨论在中文微博数据集上的性能,但是本文提出的模型同样也适用于英文多模态虚假新闻数据集.这是因为本文提出的模型主要关注文本及图片深层语义的提取和交互,与文本语言的表现形式关系不大.语言形式对模型的影响将在今后进一步的工作中进行验证.

本文采用 Jin 等人[7]基于中文新浪微博平台构建的虚假新闻数据集.该数据集包含微博官方谣言举报平台上从 2012-05—2016-01 所有官方认证为

① https://ai.baidu.com/tech/imagecensoring

假的新闻消息,以及从新华社的热点新闻发现系统采集的同时期的真实新闻的微博消息.由于社交媒体平台上的消息存在一定噪声和冗余,为保证数据集的质量,Jin 等人去除了重复图像、过小的图像以及垃圾图像等.为更好地验证模型在新的新闻事件上的泛化能力,在划分训练数据、验证数据及测试数据时,本文先将所有数据进行聚类,得到不同的事件.在此基础上对所有数据进行事件级别的划分,从而保证训练数据、验证数据以及测试数据不会包含同一事件的新闻.由于整体数据量比较小,本文按照3:1:1的比例划分最终的训练集、验证集和测试集,相关数据指标如表 1 所示.

Table 1　Statistics of the Dataset

表 1　数据集统计指标

数据集划分	训练集	验证集	测试集	总计
假新闻	2 849	950	950	4 749
真新闻	2 879	950	950	4 779
总计	5 728	1 900	1 900	9 528

3.2　实验设置

本文使用准确率(accuracy)和假新闻类别上的 F1 值、精确率(precision)及召回率(recall)作为评估指标.在模型的实现上,预训练的 ERNIE 模型来自 GitHub 上的开源项目 Transformers[38].在对 VGG19 进行微调时,采用了图片翻转等数据增强的策略提升模型的泛化性能.句子的最大长度设置为 128,batch size 设置为 64.使用 ReLU 作为非线性激活函数,使用 Adam 方法[39]优化损失函数.

3.3　实验 1:虚假新闻检测性能比较

3.3.1　对比方法

为了验证本文提出方法的有效性,我们实现了 3 类代表性的方法进行性能对比.其中,attRNN 方法由参考文献[7]作者提供,其他方法由本文作者根据论文描述复现.

1) 基于单文本模态的方法

① TextCNN.采用文献[40]提出的卷积神经网络进行文本分类.使用了 3 种不同大小的卷积核,高度分别为 3,4,5.每一种卷积核的数目均设定为 100.

② BiLSTM-Att.循环神经网络是文本分类任务中一种经典的建模方法.本文选择双层的 LSTM[41]和注意力机制堆叠成的网络作为对比方法.其中,网络的隐层单元数被设定为 128.

③ BERT.预训练语言模型近年来在各类自然语言处理任务中表现优越.采用在本文任务数据集上微调后的 BERT 模型作为对比.预训练的 BERT 模型 bert-base-chinese 来自 GitHub 开源项目 Transformers[38].

④ ERNIE.采用在本文任务数据集上微调后的 ERNIE 模型作为对比.预训练的 ERNIE 模型 nghuyong/ernie-1.0 来自 GitHub 开源项目 Transformers[38].

2) 基于单视觉模态的方法

① VGG19[34].在目前的多模态虚假新闻研究中,VGG19 被广泛用作视觉特征提取器.本文将在 ImageNet 数据集[37]上预训练的 VGG19 模型在本文任务数据集上进行微调.

② ResNet152[35].将在 ImageNet 数据集上预训练的 ResNet152 模型在本文任务数据集上进行微调.

3) 基于多模态的方法

① attRNN.文献[7]提出了一种基于注意力机制的循环神经网络,用于融合文本、视觉及社交上下文 3 种模态的特征.其中,文本部分采用 LSTM 进行建模,图片部分采用预训练的 VGG19 进行特征提取.为了对比的公平性,在具体实现时,我们移除了处理社交特征的部分.

② EANN.文献[8]提出了一种基于事件对抗机制的神经网络.通过引入事件分类器作为辅助任务,引导模型学习到与事件无关的多模态特征.该模型分别采用 TextCNN 和预训练的 VGG19 进行文本及视觉模态特征提取,并将 2 种模态特征进行拼接,作为虚假新闻的多模态特征表达,输入到虚假新闻分类器及新闻事件分类器中.

③ MVAE.文献[9]提出了一种结合多模态变分自动编码器和虚假新闻检测器的多任务模型.其中,文本和图片分别通过双向 LSTM 及预训练的 VGG19 进行特征提取,两者的拼接特征被编码为一个中间表达,用于重构输入特征及虚假新闻分类.

④ KMGCN.文献[28]提出了一种知识引导的多模态图卷积网络.该方法从外部的百科知识图谱中提取文本中出现的命名实体所对应的概念作为外部知识.该方法对每条输入的多模态新闻都会构建一个图,图的节点包括文本中的单词、文本实体所对应的概念以及图片中识别到的物体名称,节点通过预训练的 Word2Vec 词向量进行初始化,边的权重设置为 2 个单词的 PMI 值.通过 2 层图卷积网络及最大池化得到图表达用于虚假新闻分类.

3.3.2 结果分析

表 2 列出了对比实验的结果,观察可得到结论:

1) 我们的方法在分类准确率上显著超过其他对比方法,说明本文提出的这种语义增强的多模态模型确实能够有效提升虚假新闻检测的性能.尤其在虚假新闻的召回率上,我们的方法超出其他方法 7 个百分点以上,说明我们的模型可以通过充分挖掘多模态语义线索,检测到被现有方法遗漏的虚假新闻.

2) 在基于多模态的方法中,KMGCN 显著低于其他对比方法.主要的原因可能是 GCN 对于微博这类短文本的建模能力较差,在此基础上无法很好地体现外部知识的作用.另外,KMGCN 仅提取了图片中的物体标签信息,对于图片语义建模不充分.

3) 基于单文本模态的方法要优于基于单视觉模态的方法,说明虚假新闻检测主要依靠文本线索.基于多模态的方法要优于具有相同子网络结构的单模态方法,说明文本和图片模态能够为虚假新闻检测任务提供互补的线索.其中,我们提出的方法与 ERNIE 相比,准确率提升了 4.3 个百分点,进一步证明了图片语义特征的重要性.

4) 在基于单文本模态的方法中,预训练语言模型要优于 CNN,RNN 等传统的文本建模方法.这种提升一方面来源于 Transformer 更强大的建模能力,另一方面受益于预训练语言模型从大量预训练语料中学习到的语言学知识.ERNIE 的效果要优于 BERT,这说明增加实体概念知识可以增强对新闻的语义理解,进而提升虚假新闻的检测效果.

Table 2 Performance Comparison of Different Methods

表 2 不同方法的性能比较

分类	方法	准确率	F1 值	精确率	召回率
单文本模态	TextCNN	0.764	0.722	0.880	0.612
	BiLSTM-Att	0.785	0.763	0.851	0.692
	BERT	0.830	0.798	**0.977**	0.675
	ERNIE	0.852	0.830	0.970	0.725
单视觉模态	VGG19	0.730	0.698	0.789	0.626
	ResNet152	0.688	0.675	0.705	0.647
多模态	attRNN	0.808	0.787	0.882	0.711
	EANN	0.803	0.776	0.899	0.682
	MVAE	0.797	0.787	0.827	0.751
	KMGCN	0.714	0.677	0.599	0.777
	Ours	**0.895**	**0.890**	0.936	**0.847**

注:黑体表示该列性能的最优值

3.4 实验 2:消去分析

3.4.1 对比方法

为验证不同的模型组件对实验结果的影响,我们设计了 5 种模型的变体,对模型进行消去分析.

1) 去掉 ERNIE.对文本及图片文本进行建模时,用双向 LSTM 结合注意力机制的网络结构替换 ERNIE;获取视觉实体表示时,用预训练的 Word2Vec 词向量替代 ERNIE 生成的词向量表示.

2) 去掉 OCR 文本.移除提取及处理图片文字的部分.此时输入信息的多模态表示由原始文本的特征表达和原始文本引导下的视觉特征向量、视觉实体向量拼接而成.

3) 去掉视觉实体.移除提取及处理图片中视觉实体的部分.此时输入信息的多模态表示由原始文本和图片文本的联合表示及原始文本引导下的视觉特征向量拼接而成.

4) 去掉特征向量.移除处理图片视觉特征向量的部分.此时输入信息的多模态表示由原始文本和图片文本的联合表示及原始文本引导下的视觉实体向量拼接而成.

5) 去掉注意力机制.移除视觉实体及视觉特征向量在文本引导下的注意力机制.此时多个视觉实体向量和视觉特征向量分别通过平均操作进行融合.

3.4.2 结果分析

表 3 列出了消去分析的实验结果,可以得到 2 个结论:

1) 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,这说明了模型各元素的有效性.

2) 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排序如下:ERNIE>图片文本>视觉实体>视觉特征向量=注意力机制.这说明对于虚假新闻检测任务,文本比图片发挥的作用更重要,图片的高层语义比低层语义更重要.

Table 3 Ablation Study

表 3 消去分析

方法	准确率	F1 值	精确率	召回率
本文方法	**0.895**	**0.890**	0.936	0.847
去掉 ERNIE	0.806	0.799	0.830	0.771
去掉 OCR 文本	0.873	0.872	0.877	**0.866**
去掉视觉实体	0.877	0.870	0.929	0.817
去掉特征向量	0.881	0.868	**0.971**	0.784
去掉注意力机制	0.881	0.870	0.963	0.793

注:黑体表示该列性能的最优值

3.5　样例分析

为了更加直观地展示本文方法的优越性,我们对比了本文模型和表 2 列出的对比方法中性能最好的 ERNIE 模型在测试集上的预测结果,并对 ERNIE 模型无法检测但本文模型能够成功检测到的多模态虚假新闻进行分析.图 4 展示了 3 条代表性的样例,分别体现了图片的视觉特征向量、视觉实体和图片文本对于虚假新闻检测的重要性.

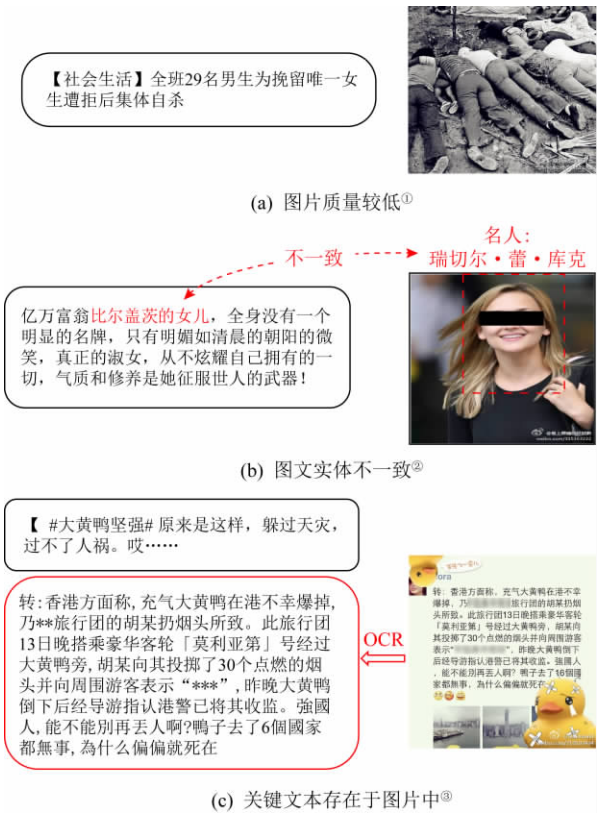

Fig. 2　Examples of multi-modal fake news detected by our model

图 2　本文模型成功检测的多模态假新闻示例

图 4(a)中的新闻配图清晰度较低,重压缩造成的块状效应明显,说明该张图片很可能是从网上下载的来自其他事件的过时图片,而非在本次事件中现场拍摄的图片.本文模型通过提取图片的视觉特征向量对图片质量进行建模,可以判断该条新闻为假的可能性较大.图 4(b)中,通过提取图片中的视觉实体,可以识别出该条新闻配图里的人物为女演员瑞切尔·蕾·库克(Rachael Leigh Cook),而非文本中所说的比尔盖茨的女儿.本文模型通过提取图片中的视觉实体并利用预训练语言模型中隐含的事实知识,可以发现图文语义的冲突点,作为虚假新闻的线索.图 4(c)为以图片为主体的多模态新闻,其原始文本中包含的信息量较少,不足以提供假新闻判定的线索.只有对图片进行文字提取,才能充分理解该条新闻的语义,从而发现虚假新闻的线索.

4　总　　结

针对现有方法对于多模态新闻语义理解不足的问题,本文提出了一种语义增强的多模态虚假新闻检测方法.通过隐式利用外部模型中存储的大量事实知识,更好地理解多模态新闻的深层语义.提取不同语义层次的视觉特征,并采用文本引导的注意力机制建模图文之间的语义交互,从而更好地融合多模态异构特征.实验结果表明:本文提出的方法在准确率上大幅超越当前最好的方法,证明了基于语义增强方法的有效性.

参　考　文　献

[1] Tang Xujun, Huang Chuxin, Wu Xinxun. Annual Report on the Development of New Media in China No.11 [M]. Beijing: Social Sciences Academic Press, 2020 (in Chinese)
(唐绪军, 黄楚新, 吴信训. 中国新媒体发展报告 No.11 [M]. 北京: 社会科学文献出版社, 2020)

[2] Boididou C, Andreadou K, Dang-Papadopoulos S, et al. Verifying multimedia use at MediaEval 2015 [G/OL] //Proc of the MediaEval 2015 Workshop. Wurzen, Germany: CEUR-WS, 2015 [2020-10-08]. http://ceur-ws.org/Vol-1436/Paper4.pdf

[3] Sunstein C R. On Rumors: How Falsehoods Spread, Why We Believe Them, and What Can Be Done [M]. Princeton: Princeton University Press, 2014

[4] Shu Kai, Sliva A, Wang Suhang, et al. Fake news detection on social media: A data mining perspective [J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36

① 该假新闻示例已被新浪微博社区管理中心判定为不实信息 https://service.account.weibo.com/show?rid=K1CaJ6g5f7aYd.新闻文本已在不影响语义的前提下进行删减展示

② 该假新闻示例已被新浪微博社区管理中心判定为不实信息 https://service.account.weibo.com/show?rid=K1CaJ6wpc66kl.新闻文本已在不影响语义的前提下进行删减展示

③ 该假新闻示例已被新浪微博社区管理中心判定为不实信息 https://service.account.weibo.com/show?rid=K1CaJ7Apk7aci.为避免引起读者不适,新闻图片及 OCR 文字中的谩骂言语已被隐去

[5] Jin Zhiwei, Cao Juan, Luo Jiebo, et al. Image credibility analysis with effective domain transferred deep networks [EB/OL]. (2016-11-16) [2020-10-08]. https://arxiv.org/pdf/1611.05328.pdf

[6] Qi Peng, Cao Juan, Yang Tianyun, et al. Exploiting multi-domain visual information for fake news detection [C] //Proc of the 2019 IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2019: 518-527

[7] Jin Zhiwei, Cao Juan, Guo Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs [C] //Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 795-816

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

[9] Khattar D, Goud J S, Gupta M, et al. MVAE: Multimodal variational autoencoder for fake news detection [C] //Proc of the Web Conf 2019. New York: ACM, 2019: 2915-2921

[10] Guo Bin, Ding Yasan, Yao Lina, et al. The future of false information detection on social media: New perspectives and trends [J]. ACM Computing Surveys, 2020, 53(4): No.68

[11] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter [C] //Proc of the Web Conf 2020. New York: ACM, 2011: 675-684

[12] Qazvinian V, Rosengren E, Radev D, et al. Rumor has it: Identifying misinformation in microblogs [C] //Proc of the 2011 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1589-1599

[13] Pérez-Rosas V, Kleinberg B, Lefevre A, et al. Automatic detection of fake news [EB/OL]. (2017-08-23) [2020-10-08]. https://arxiv.org/pdf/1708.07104.pdf

[14] 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. Menlo Park, CA: AAAI, 2016: 3818-3824

[15] Shu Kai, Wang Suhang, Liu Huan. Understanding user profiles on social media for fake news detection [C] //Proc of the 2018 IEEE Conf on Multimedia Information Processing and Retrieval. Piscataway, NJ: IEEE, 2018: 430-435

[16] Li Quanzhi, Zhang Qiong, Si Luo. Rumor detection by exploiting user credibility information, attention and multi-task learning [C] //Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 1173-1179

[17] Liu Bo, Li Yang, Meng Qing, et al. Evaluation of content credibility in social media [J]. Journal of Computer Research and Development, 2019, 56(9): 1939-1952 (in Chinese)
(刘波, 李洋, 孟青, 等. 社交媒体内容可信性分析与评价 [J]. 计算机研究与发展, 2019, 56(9): 1939-1952)

[18] Jin Zhiwei, Cao Juan, Jiang Yugang, et al. News credibility evaluation on microblog with a hierarchical propagation model [C] //Proc of the 2014 IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2014: 230-239

[19] Jin Zhiwei, Cao Juan, Zhang Yongdong, et al. News certification by exploiting conflicting social viewpoints in microblogs [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 2972-2978

[20] Shu Kai, Wang Suhang, Liu Huan. Beyond news contents: The role of social context for fake news detection [C] //Proc of the 20th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2019: 312-320

[21] Ma Jing, Gao Wei, Wong K F. Rumor detection on Twitter with tree-structured recursive neural networks [C] //Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 1980-1989

[22] Jin Zhiwei, Cao Juan, Zhang Yongdong, et al. Novel visual and statistical image features for microblogs news verification [J]. IEEE Trans on Multimedia, 2016, 19(3): 598-608

[23] Boididou C, Papadopoulos S, Dang-Nguyen D T, et al. The CERTH-UNITN participation @ verifying multimedia use 2015 [G/OL] //Proc of the MediaEval 2015 Workshop. Wurzen, Germany: CEUR-WS, 2015 [2020-10-08]. http://ceur-ws.org/Vol-1436/Paper56.pdf

[24] Sun Shengyun, Liu Hongyan, He Jun, et al. Detecting event rumors on sina weibo automatically [G] //LNCS 7808: Proc of the 15th Asia-Pacific Web Conf 2013. Berlin: Springer, 2013: 120-131

[25] Zlatkova D, Nakov P, Koychev I. Fact-checking meets fauxtography: Verifying claims about images [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2019: 2099-2108

[26] Cao Juan, Qi Peng, Sheng Qiang, et al. Exploring the role of visual content in fake news detection [G] //LNSN: Disinformation, Misinformation and Fake News in Social Media. Berlin: Springer, 2020: 141-161

[27] Zhang Huaiwen, Fang Quan, Qian Shengsheng, et al. Multi-modal knowledge-aware event memory network for social media rumor detection [C] //Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1942-1951

[28] Wang Youze, Qian Shengsheng, Hu Jun, et al. Fake news detection via knowledge-driven multimodal graph convolutional networks [C] //Proc of the 2020 Int Conf on Multimedia Retrieval. New York: ACM, 2020: 540-547

[29] Trieu H T, Quoc V L. Do language models have common sense? [EB/OL]. (2018-12-21) [2020-11-20]. https://openreview.net/forum?id=rkgfWh0qKX

[30] Joe D, Joshua F, and Alexander R. Commonsense knowledge mining from pretrained models [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing. Stroudsburg, PA, 2019: 1173-1178

[31] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: ACL, 2019: 4171-4186

[32] Sun Yu, Wang Shuohuan, Li Yukun, et al. ERNIE: Enhanced representation through knowledge integration [EB/OL]. (2019-04-19) [2020-10-08]. https://arxiv.org/pdf/1904.09223.pdf

[33] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, Inc., 2017: 5998-6008

[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2020-10-08]. https://arxiv.org/pdf/1409.1556.pdf

[35] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[36] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] //Proc of the 2015 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1-9

[37] Deng Jia, Dong Wei, Socher Richard, et al. ImageNet: A large-scale hierarchical image database [C] //Proc of the 2009 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248-255

[38] Wolf T, Chaumond J, Debut L, et al. Transformers: State-of-the-art natural language processing [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, PA: ACL, 2020: 38-45

[39] Kingma D P, Ba J. Adam: A method for stochastic optimization [EB/OL]. (2017-01-30) [2020-10-08]. https://arxiv.org/pdf/1412.6980.pdf

[40] Kim Y. Convolutional neural networks for sentence classification [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1746-1751

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

Qi Peng, born in 1996. PhD candidate. Student member of CCF. Her main research interests include false information detection and multimedia content analysis.

亓　鹏,1996 年生.博士研究生.CCF 学生会员.主要研究方向为虚假信息检测、多媒体内容分析.

Cao Juan, born in 1980. PhD, professor, PhD supervisor. Member of CCF. Her main research interests include multimedia content analysis and artificial intelligence security.

曹　娟,1980 年生.博士,研究员,博士生导师.CCF 会员.主要研究方向为多媒体内容分析、人工智能安全.

Sheng Qiang, born in 1995. PhD candidate. Student member of CCF. His main research interests include false information analysis and detection, and automatic fact-checking.

盛　强,1995 年生,博士研究生.CCF 学生会员.主要研究方向为虚假信息分析与检测、事实核查.