

知识融合的多模态虚假新闻检测方法

王 茜, 韩冀中

中国科学院信息工程研究所 北京 中国 100093

摘要 随着互联网发展和社交网络的大流行, 虚假新闻和谣言能够迅速地传播并对公共产生巨大的负面影响。有效地识别出虚假新闻对阻断虚假新闻的快速传播有重要意义。由于新闻中除文字还常常包含图片等多模态的内容, 近年来的工作多着重于研究利用深度神经网络提取文本和视觉的特征表示进行虚假新闻检测的方法。然而, 却忽略了对新闻中所含的事实和知识的验证。在早期的工作中, 研究人员通常将新闻中的实体和关系, 与知识库中的实体关系进行比对, 计算出新闻的可信度。这类工作基于简单的比对验证, 往往难以覆盖海量信息内容的识别。但其利用外部知识库进行计算的思想仍有借鉴意义。我们提出一种知识融合的深度神经网络方法, 同时提取新闻中的知识文本、及视觉 3 个方面的特征表示, 用于虚假新闻识别任务。具体来说, 在我们的方法中, 我们首先利用大型的开放式知识图谱来计算新闻中实体之间的关系, 得到新闻的实体关系矩阵, 并用卷积神经网络提取关系矩阵中的特征, 作为新闻的知识特征表示。接着, 将知识特征与文本和视觉特征相融合, 进行虚假新闻的检测。在本文中, 我们在预训练的模型上为虚假新闻检测任务微调文本和图像特征。我们的知识特征也能够很方便地与其他特征一起串联使用。我们的方法在两个大型的多模态数据集上进行验证。数据集分别是从小红书采集的中文数据集和从推特上采集的英文数据集。丰富的实验证明了我们算法的有效性。在这两个数据集上, 我们的方法都得到了最优的虚假新闻检测结果。

关键词 虚假新闻识别; 知识图谱; 深度神经网络

中图法分类号 TP183 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.01.13

Multi-Domain Fake News Detection Fusing Knowledge Information

WANG Xi, HAN Jizhong

Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

Abstract Fake news and rumors can spread rapidly with the development of internet and the popularity of social network, and cause a huge negative impact on the public. Fake news detection is very important to blocking the spread of fake news. Since the wildly used of multi-media content in addition to text in posts, recent approaches had paid more attention to how to extract the feature representations of text and vision by using deep neural network. However, they ignored the verification of facts and knowledge information contained in news. In the early fact-checking articles, researchers usually matched the entity and relationship in news with the entity relationship in knowledge base, and calculated the credibility of news. This kind of works were based on simple comparison and verification, which were hard to cover large scale multi-media data. However, the idea of using external knowledge for calculation still is worth learning from. In this paper, we propose a knowledge fusion based deep neural network method, which extracts three domains features such as textual, visual and knowledge information in news for fake news detection task. Specifically, we first use the large open knowledge graphs to obtain the relationship between entities in news, and get the entity relationship matrix of news. Then we obtain the knowledge feature representation of news by extracting the features of the relation matrix with convolution neural network. Finally, knowledge features can be integrated with text and visual features to detect fake news. In this paper, the text and visual features are fun-tuning based on per-trained model for fake news detection task. Our knowledge features can also be conveniently concatenated with other features. Our method is validated on two large multi-modal datasets: The Chinese dataset collected from Weibo and the English dataset collected from Twitter. Extensive experiments show the significant improvement of using our algorithm, and our model achieved the state-of-the-art performance on both two datasets.

Key words fake news detection; knowledge graph; deep neural network

通讯作者: 韩冀中, 博士, 正高级工程师, Email: hanjizhong@iie.ac.cn。

本课题得到国家自然科学基金(No. 61702502)及国家留学基金委公派留学项目资助。

收稿日期: 2020-12-01; 修改日期: 2021-02-24; 定稿日期: 2023-03-02

1 引言

社交网络是人们共享信息的重要途径之一。任何人都能够在社交网络上发布信息, 这为虚假新闻的生产与传播提供了便捷。同时, 由于社交网络的流行, 虚假新闻可能被迅速的大规模传播, 从而对公共安全产生巨大的负面影响。例如, 2020 年发生的新冠病毒席卷全球事件中, 有层出不穷的假新闻在网络上发布和传播^[1], 造成恶劣影响。其中, 有的新闻利用大众恐惧心理虚假宣传产品引发疯狂抢购; 有的发布未经证实的错误治疗方法损害人身健康; 还有的甚至恶意挑起种族对抗, 引发暴力事件危害公共安全。为了遏止虚假新闻的发布和传播, 马来西亚政府甚至逮捕了 4 名在 Facebook 和 Twitter 上传播有关新冠肺炎假消息的人。因此, 亟须研究有效的虚假新闻自动识别方法, 为阻止虚假新闻快速传播提供技术手段。

近年来, 学者致力于虚假新闻检测问题的研究。虚假新闻检测中的主要研究路线包括基于传播的检测和基于内容的检测两类。其中, 基于传播的检测方法主要利用新闻在传播中的特征, 如用户、转发、评论等进行检测。Jin 等^[2-3]提出的基于图的优化方法中, 将每条推文看作上图上的一个节点。用当前节点的相邻节点的可信度来估计当前节点的可信度。这类型方法对虚假新闻的辨别更依赖新闻之间的相关性而非新闻的内容本身。另一类方法更关注对新闻的内容进行辨别, 这类方法称为基于内容的虚假新闻识别方法^[4-5]。传统的基于内容的方法^[6-7]通常使用一些手工设计的特征来识别虚假新闻。例如, 统计虚假新闻中的常用词分布来分析新闻的写作语言风格; 计算新闻中一对实体在知识图谱中的路径长度用于事实验证, 等。然而, 这类手工设计的特征难以应对丰富的具有多样性的大规模的数据的检测要求。近年来, 深度神经网络被应用于虚假新闻识别任务中去^[8]。Wang 等^[9]使用卷积神经网络(Convolutional Neural Network, CNN)自动地提取文本的写作风格特征, 利用深度神经网络发掘文本的隐含的写作模式。文献[10-12]利用深度神经网络同时提取新闻中的文本和图像的特征, 融合两种模态的特征进行虚假新闻识别, 并提高了虚假新闻识别任务的效果。然而, 这类基于深度学习的方法不能考虑到外部变化的知识和信息, 它们的方法将每个新闻看成一个独立的与外部事物隔离开的个体。然而, 外部的知识和信息对虚假新闻识别是十分重要的。例如, 新闻“2020 东

京奥运会将延迟举办”, 在 2020 年 3 月 24 日之前是谣言, 奥组委还多次发布辟谣声明。然而, 在 2020 年 3 月 24 日奥组委正式宣布之后, 该新闻内容属实。由此可见, 在判定新闻真伪时应考虑到外部的知识和事件信息。

在本文中, 我们提出一种新颖的基于知识融合的深度神经网络方法, 通过利用知识图谱的信息将外部环境的特征考虑进来, 同时融合了新闻的文本、视觉及知识 3 个方面的特征用于虚假新闻识别任务。我们的方法一方面考虑到了传统的基于事实验证的方法中对知识、事实、事件信息的使用, 另一方面又能够利用深度学习的方法对大规模数据进行自动的特征提取。具体来说, 在我们的方法中, 我们将一条新闻看作一个图, 新闻中的实体对应着图中的节点。我们计算节点之间的关系, 比对节点之间在大型的开放知识图谱中是否有连接。如果两个节点在知识图谱中有连接, 我们认为这一对节点相关, 反之则为不相关。从而得到了每条新闻的实体邻接关系矩阵。该关系矩阵中就包含了外部环境中的知识、事实、事件信息。接着, 我们对新闻的实体邻接关系矩阵进行特征提取。在下文中, 我们简称提取出的特征为知识特征。为了使知识特征能够更好地和其他域上的特征进行融合, 我们利用 CNN 来提取实体关系矩阵的特征。同时, Text-CNN 和 VGG-19 被分别用来提取新闻的文本和图像特征。最终, 我们将知识特征、文本特征及图像特征进行融合, 送到多层感知分类器中, 判别新闻的真假。我们工作的主要贡献总结如下:

1) 我们提出了一种融合文本、视觉及知识 3 个特征的虚假新闻识别方法; 据我们所知, 我们是第一个将外部知识与文本、视觉同时用深度神经网络融合用于虚假新闻识别任务的方法。

2) 传统的基于神经网络的特征提取的方法中, 邻接关系矩阵通常被用来调整特征的权重, 或简单的计算关系矩阵的度来作为其特征表示。在我们的方法中, 尝试一种新的思路, 直接使用 CNN 来提取邻接关系矩阵的特征。该提取方法得到的知识特征可以被应用到其他的如文本分类、文本情感分析等文本处理任务中去。

3) 我们使用一种简单鲁棒的方法来引入外部的大型知识图谱中的信息。

4) 为了验证本文方法的有效性, 我们分别在两个大规模的真实场景下的数据集: Weibo 数据集和 Twitter 数据集上进行实验。实验表明, 在两个数据集上我们均取得了最好的实验结果。

2 相关工作

虚假新闻检测是一个有重要意义的研究课题,也是一个具有重大挑战性的课题。近年来,学者尝试多种思路来解决这个问题^[13]。

社交媒体上的推文通常包括用户、发表时间、标题、正文、图像和视频等多种信息。基于图的方法通常更关注推文之间的关系。这种方法主要从新闻的传播特性来研究虚假新闻的识别,将新闻看作是图上的节点。若其中一个节点被识别成假新闻,那么在图中与该节点更相关更相似的节点被认为有较大概率也属于假新闻。Ishida 和 Kuraya^[14]提出使用动态的关系网络来进行假新闻识别。文献[15-16]通过社交信息来捕捉用户之间的互动用于计算推文的可信度。另一类基于新闻内容的检测方法更多地关注新闻内容的特征。在文献[10-11,17]中利用新闻的文本和图像内容来识别是否是假新闻。由于我们的方法属于基于内容的检测方法,因此在本小节中,我们主要介绍基于新闻内容的虚假新闻识别的方法的研究进展。在此类方法中,最关键的部分在于如何提取内容的特征,如文本特征和视觉特征等。接下来的部分,我们重点围绕新闻内容的特征提取展开相关工作介绍。

新闻文本特征的提取方法主要可分为两类:基于知识的特征提取方法和基于语言风格的特征提取方法。基于知识的方法通常采用事实验证的手段进行虚假新闻检测。其基本思想是利用知识检查文本中声明的内容的真实性。Shiralkar 等^[7]提出一种称为关系知识链接器(Relational Knowledge Linker)的方法。该方法通过计算新闻中两个相关节点在知识图谱(Knowledge Graph, KG)中的最短路径的长度来判别谣言。文献[18]中,作者提取新闻文本中的三元组,并将其与 KG 中的三元组对比,利用知识图谱嵌入向量计算偏差,用于分类器进行真假新闻分类。基于语言的虚假新闻检测方法^[6,9,19]通常通过分析新闻的写作风格来进行真假新闻辨别。文献[6]中认为假新闻和垃圾邮件类似。作者统计文章中的语法错误,用词量等特征来进行虚假新闻检测。文献[19]统计单词、大写字符和标点符号等作为谣言检测的特征。文献[9]中通过分析文章的语言模式来进行判别。这些手工设计的特征往往高度依赖于特定领域的知识^[20]或特定的数据集,往往泛化性能不佳,难以覆盖互联网上海量数据的检测。

现在的新闻和推文中经常包含一些图像、视频等多媒体的内容^[20],这些内容直观形象,同时能够增加新闻的可信度。在文献[16,21-23]中,作者们使用了一些手工设计的图像特征来进行谣言的判别。这些手工

设计的图像特征难以很好地表示复杂的视觉内容。

深度神经网络已被证明在文本和视觉等多个领域具有强大的表示能力。在虚假新闻检测的任务中,Jin 等^[10]首次提出了一种融合文本、图像和社交信息的深度学习框架。与传统方法相比,该方法显著地提高了检测的性能。Wang 等^[11]在融合文本和图像模态的基础上,在深度神经网络框架中增加了分类新闻事件的任务。Qi 等^[12]利用深度学习分析图像频域的特征,从频域的角度上改进了图像的特征表示。然而,上面这些方法仅仅关注如何利用深度网络得到更好的文本和视觉特征表示,却没有很好地考虑到知识方面的信息。而知识信息对区分真假新闻是至关重要的。

在本文中,我们提出了一个融合文本、视觉和知识域信息的统一的深度神经网络框架。不同于传统的如利用最短路径长度进行分类的基于知识的方法,我们设计了一个基于 CNN 的网络来提取基于知识图谱的关系矩阵的特征表示。所得到的关系特征具有易于与其他域特征相融合的特点。

3 模型

在本文中,我们旨在同时利用新闻的文本及图像等多模态的内容信息和大型开放知识图谱中丰富的知识信息来完成虚假新闻检测任务。模型既需要考虑到新闻内容的特点,又能利用外部的知识进行综合判断。基于此思想,我们提出一个新颖的神经网络框架同时提取新闻的文本、图像和知识 3 个域上的特征进行虚假新闻检测。同时,我们提出一种鲁棒的方法来利用 KG 中的信息。具体细节如下所示。

3.1 模型概述

在新闻和推文中通常含有多种模态信息,包括文本、图像及视频等。在我们的模型中,主要关注其中的文本和图像信息。我们将一条新闻定义为 $P = \{T, I\}$, 其中, T 表示新闻的文本信息, I 表示图像。我们模型的框图如图 1 所示。

模型的输入是新闻的文本和图像内容。通过联合特征提取模块得到新闻 3 个域上的联合特征。3 个域的特征分别用 3 个分支进行提取,分别是:通过词嵌入(Word Embedding, WE)和 TEXT-CNN 提取文本特征;通过视觉神经网络 VGG-19 提取图像特征;从知识图谱得到新闻文本中实体关系矩阵,并用 CNN 对关系矩阵进行特征提取,我们称该分支为知识特征提取。接着,将 3 种特征进行融合,得到新闻特征的联合表示。最后,联合特征通过全连接神经网络多层感知机进行最终分类。

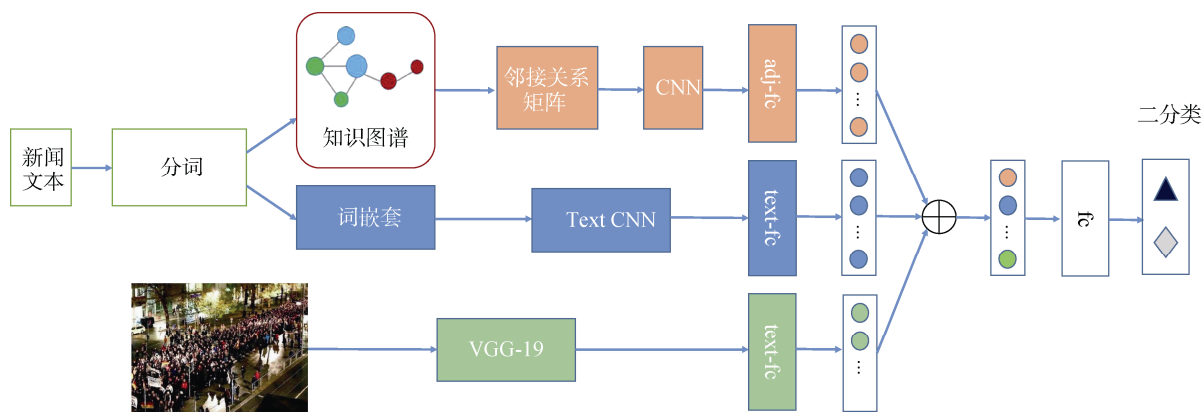


图 1 我们提出的虚假新闻检测模型框架图。在特征提取部分包含 3 个分支。图中顶部分支表示计算从开放式知识图谱得到关系矩阵的特征提取。中间分支和下部分支分别表示本文和视觉特征的提取。

Figure 1 The framework of our proposed fake news detection model. It contains three branches in the feature extractor. The top branch computes the features of relationships extracting from the open knowledge graph. The middle branch and bottom branch extract the textual and visual features, respectively.

下面, 我们详细介绍模型的各个部分。

3.2 知识特征提取

在传统方法中^[18,24], 知识图谱往往被作为事实验证的外部比对库。研究人员通常计算新闻中两个概念之间的距离或相关性值, 例如两个概念对应到知识图谱中实体之间的最短路径等, 然后将这些距离或相关性值与一个设定的阈值相比较, 作为分类的依据。这些方法通常需要一个比较完备的知识图谱, 或为任务建立特殊的知识图谱。同时也难以和其他的文本特征相融合。

在本节中, 我们提出一种简单鲁棒的方法来融合外部的大型开放知识图谱信息, 同时, 能够表征出文本的整体的知识特征。我们的模型得出的特征能够很容易与其他模态的特征相结合。

在我们的模型中, 我们将新闻的文本看作是一个图, 定义为 $G=(V, E)$ 。其中 V 是构成新闻的词组的集合, 每个词组被看作一个节点。 E 是词组之间的关系集合。首先, 我们使用文本分割得到新闻文本中的词。在这里我们没有使用知识图谱中的三元组提取方法。因为三元组提取通常会对句子有更高的语法要求。而社交媒体上的句子通常是自由书写风格随意, 我们直接使用文本分割能够加大算法的鲁棒性, 更多地保留句子中的词组。在 KG 中, 一个三元组可以表示为 $Triplet=(entity_{start}, relation, entity_{end})$ 。那么, 文本 G 的邻接关系矩阵 A 中的元素由计算节点间的关系得到, 为

$$e_{ij} = \begin{cases} 1, (v_i, v_j) \in (entity_{start}, entity_{end}) \\ 0, (v_i, v_j) \notin (entity_{start}, entity_{end}) \end{cases} \quad (1)$$

其中, e_{ij} 是 A 中的元素。这样, 我们得到了一个 2 维

的邻接关系矩阵 A 。在传统的基于知识的方法中, 他们总是使用正则表达式来分析 A 中的关系, 在本方法中, 我们使用了 2 维 CNN 来获取邻接矩阵的特征, 具体的网络结构如图 2 所示。

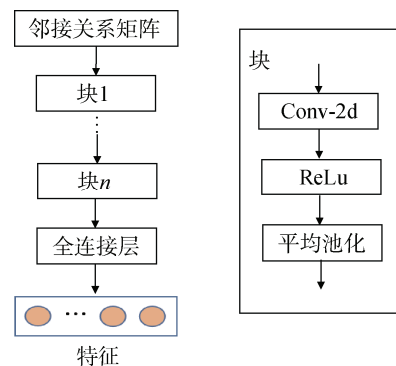


图 2 邻接矩阵特征提取的网络结构

Figure 2 The architecture of extracting the feature of the adjacency matrix

网络的输入是包含知识信息的 2 维邻接矩阵。如图 2 所示, 一个块的计算过程是, 先经过 2 维卷积层, 接着通过激活层: $y = \sigma(f(x, w))$, 最后通过平均池化层得到输出。邻接矩阵特征提取网络由 n 个块和一个全连接层组成。2 维邻接矩阵通过该网络后得到一个 1 维的知识特征表示。我们将该知识特征表示标记为 R_A 。 R_A 与其他特征融合的细节将在 3.5 节中介绍。

3.3 文本特征提取

文本在虚假新闻的识别中至关重要。传统方法中^[25]经常使用如循环神经网络(Recurrent Neural Network, RNN)这类处理序列数据的模型来提取文本

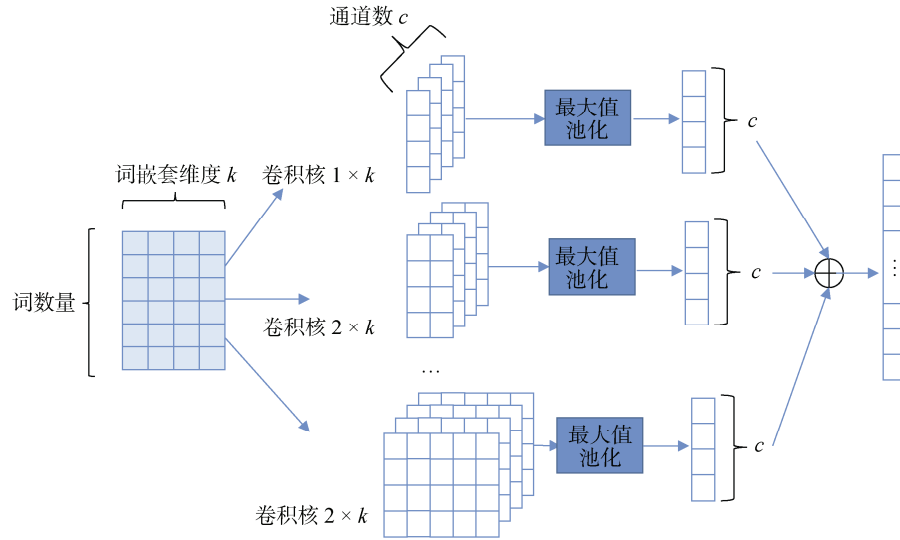


图 3 Text-CNN 的网络结构图

Figure 3 The network structure of Text-CNN

的特征。在这些方法中, 文本可以看作是一系列词语的序列。在近期工作中, CNN 被证实对文本的分类任务有效^[26-27]。由于 CNN 的结构更适合进行快速的批量训练, 因此我们采用文献[11,26]中的 Text-CNN 的方法进行文本的特征提取。其网络结构如图 3 所示。

在 Text-CNN 中, 将词向量矩阵作为模型的输入。词向量矩阵定义为 $C \in R^{n \times k}$ 。其中, n 是一个给定的值, 表示最多能接受的单词的数量。 k 是词嵌入后得到的特征的维度。接着, 多组卷积核被用于输入向量矩阵的处理。在计算机视觉任务中, 卷积核大小通常被设置为正方形 $m \times m$ 。然而, 在自然语言处理 (Natural Language Processing, NLP) 任务中, 词向量特征通常为一维, 且有其特殊的含义。因此, 在本任务中, 我们使用长方形卷积核代替正方形的卷积核。卷积核被定义为 $Kernel \in R^{k \times window_size}$ 。在同一组的卷积核拥有相同大小的 $window_size$ 。接着, 每组卷积核的输出被串联起来, 通过一个全连接层得到最终的文本特征输出 R_T 。

3.4 图像特征提取

我们使用一个预训练的 VGG-19 网络^[28]来提取新闻的图像特征。同时, 我们在 VGG-19 网络的最后一层增加一个全连接层, 来调整最终输出的视觉特征的维度。视觉特征标记为 R_V 。其计算公式如下:

$$R_V = \sigma(W_v \cdot F_{vgg}) \quad (2)$$

其中, F_{vgg} 是 VGG-19 网络最后一层的输出, W_v 是添加的全连接层的参数。在整个端对端的训练过程中, 为了节省内存空间, 和 GPU 的计算资源, 我们保持 VGG-19 网络中已预训练部分的参数不变。

3.5 虚假新闻分类

通过上述步骤, 我们分别得到了知识、文本和图像 3 个域上的特征。我们将这 3 个域上的特征串联起来得到新闻的联合特征表示 R_F , 如下:

$$R_F = R_T \oplus R_V \oplus R_A \quad (3)$$

最后, 联合特征表示 R_F 通过一个全连接层的多层感知机分类器, 用逻辑回归函数 Softmax 得到最终的二分类结果, 判断新闻是真或假。

$$p = \text{Softmax}(W_{fc} \cdot R_F) \quad (4)$$

其中, W_{fc} 为多层感知机中全连接层的参数。最后, 在目标函数中我们使用交叉熵作为损失函数。

$$L = -[y \log p + (1 - y) \log(1 - p)] \quad (5)$$

其中, y 是新闻的真实标记类别。

模型进行训练时使用批量随机梯度下降的算法来最小化损失函数。

4 实验验证

本小节中, 首先介绍了 2 个来自真实世界社交媒体平台的虚假新闻数据集: 微博数据集和推特数据集。接着, 介绍了在本任务上的基线算法和近期的代表性方法, 作为本文方法的对比算法。最后, 对我们算法的实验结果进行分析。在两个真实数据集上丰富的实验证明了我们算法的有效性。

4.1 数据集

为了验证我们方法的有效性, 我们使用两个目前主流的真实世界采集的数据集, 分别采集自微博数据和推特数据。数据集的数据统计情况如表 1 中所示。

表 1 两个真实世界数据集的统计
Table 1 Statistics of two Real-World datasets

	微博数据集			推特数据集		
	假新闻	真新闻	图片	真新闻	假新闻	图片
训练集	3749	3783	10819	7334	5599	420
测试集	1000	996	2453	564	427	104
总计		9528	13272		13924	514

4.1.1 微博数据集

微博数据集由 Jin 等^[10]在 2012 年提出。其中的虚假新闻收集自 2012 年 5 月至 2016 年 1 月期间微博官方谣言举报系统,并由信誉用户委员会确认的谣言信息。真实新闻是从中国新闻的权威机构新华社认证的消息中采集。在该数据集中,每条新闻至少包含一张图片。如表 1 所示,该数据集总计包含新闻 9528 条,其中训练集新闻 7532 条,测试集新闻 1996 条,真新闻与假新闻的比例为 1:1。含图片 13272 张。

4.1.2 推特数据集

推特数据集在竞赛 MediaEval^[29]中被启用,目的是能够自动地检测推特上多种媒体形式的虚假新闻。在该数据集中,包含推文的文本、图像、视频链接、用户、发布等信息。由于我们的方法关注在文本及图像。因此,不使用数据集中仅含有视频链接的推文。该数据集含有新闻 13924 条,其中训练集 12933 条,测试集 991 条。真假新闻的比例接近 1:1.3。数据集含图像 514 张,有多条相同话题的新闻使用同一图片的情况。

4.2 实验设置

对于新闻中的文本,我们使用现有的文本分割工具进行分词,这个步骤让我们获得新闻中的实体,用于在知识图谱中的匹配。在我们的方法中,使用 Jiagu 处理微博数据集中的文本,用 SpaCy 处理推特数据集中的文本。在词嵌套过程中,我们使用预训练的词向量来初始化词嵌套模型。在推特数据集上,我们使用腾讯 AI 词向量^[30],其输出的词向量维度是 200。在推特数据集上,我们使用的是 Google 的 word2vec 词向量^[31],其特征维度为 300。在后面的训练过程中,对词嵌套模型进行微调。在 Text-CNN 中,我们设置通道数为 24,使用 5 种不同的卷积核,大小为 $n \times k, n \in [1, 5], k$ 是词向量的维度。文本特征提取中,最后输出的特征维度为 128 维。

在图像特征提取中,我们使用的是在 ImageNet^[32]上的 VGG-19^[28]预训练模型的最后一层的输出,其特征的维度是 1000。该 1000 维特征通过一个全连接层输出 128 的图像特征。在训练时,为了

节省内存,我们固定 VGG-19 部分的参数,仅训练最后的全连接层部分。

在知识特征提取中,句子中的每个单词被看作是 1 个节点。鉴于代词等词组的意义过于广泛,我们仅考虑名词和其他词语之间的关系。对于中文,我们使用的是 Ownthink^[33]知识图谱,其中包含了 1.4 亿个三元组。用于英文的是 Freebase^[34],其中包含 1.9 亿个三元组。知识特征模块最终输出的是 16 维度的知识特征。我们将在 4.5 节中给出更多的细节。

最终,3 个域上联合特征的维度是 $2 \times 128 + 16$ 。我们先训练知识特征提取,并固定它的参数。更多的分析在 4.5 节给出。然后,整个网络一起进行批训练,批量大小 $batch_size = 128$ 。

4.3 对照方法

为了验证我们的模型,我们给出一系列的对照方法作为对比。

1) 仅文本特征的检测方法。我们将图像特征提取和知识特征提取从模型中移除,仅留下文本特征提取分支,进行虚假新闻的二分类。

2) 仅视觉特征的检测方法。文本和知识特征从模型中移除,仅保留图像特征提取的检测模型。

3) 仅知识特征的检测方法。从模型中去掉文本和图像特征的提取,仅使用知识特征进行虚假新闻检测。

4) VQA 模型检测方法。原始的 VQA 模型^[35]被设计用来计算给定输入图像和问题的答案。它可以被调整为面向虚假新闻检测任务,仅需要将原始 VQA 中最后的多类别分类器改为二分类器。改动后的模型结果在文献[10]中给出。在本论文中使用相同模型的结果作为比对。

5) NeuralTalk 模型检测方法。原始 NeuralTalk 模型^[36]用于图片图像描述任务,即自动生成一句话来描述给定的图片。该方法同样涉及文本和图像的处理。也可以迁移到虚假新闻识别任务。仅需要将最后 RNN 的输出作为联合特征。

6) Att-RNN 检测模型。Att-RNN 模型^[10]是基于注意力机制的多模态虚假新闻检测方法。

7) EANN 检测模型。EANN 模型^[11]是同时利用文本、图片和事件的虚假新闻检测方法。

8) 去掉知识特征的模型。我们的模型中去掉知识特征提取模块,仅使用文本和图像特征的虚假新闻检测。

4.4 性能比较

表 2 中给出了我们方法和 4.3 节所述方法的性能比较。我们分别计算了在两个数据集上的准确率,和

谣言类、非谣言类新闻的精确率、召回率和 F_1 值。

从表 2 可看出, 在两个数据集上, 我们的方法都取得了显著的提升。在微博数据集上, 我们的方法达到了 87.1% 的准确率, 比此前的最好方法 EANN 提高了 4.4% 的准确率。同时, 我们注意到, 在消融实验中, 在融入知识特征后, 大幅提升了假新闻的召回率和真新闻的精确率。与去掉知识特征的模型相比, 假新闻类别中的召回率提高了 14.3%, F_1 值提高了

4.2%。在推特数据集上, 我们同样得到了最高的 73.5% 的准确率。从实验看出, 在两个数据集上, 当与其他特征融合使用时, 知识特征显著提高了虚假新闻检测的效果。在所有的实验中, Softmax 的分类阈值都被设为 0.5。在仅使用知识特征进行分类时, 在谣言类新闻上的 F_1 值要大幅高于非谣言类的 F_1 值。在本文的 4.5 节中, 我们更进一步对知识特征的效果进行详细的分析。

表 2 两个数据集上的现有工作对比
Table 2 Comparison of different methods on two datasets

数据集名称	方法名称	准确度	假新闻			真新闻		
			精确率	召回率	F_1	精确率	召回率	F_1
微博数据集	仅视觉特征	0.648	0.64	0.70	0.668	0.658	0.597	0.626
	仅文本特征	0.828	0.905	0.738	0.813	0.774	0.92	0.841
	仅知识特征	0.671	0.627	0.869	0.728	0.777	0.468	0.584
	VQA ^[34]	0.736	0.797	0.634	0.706	0.695	0.838	0.76
	NeuralTalk ^[35]	0.726	0.794	0.613	0.692	0.684	0.84	0.754
	Att-RNN ^[10]	0.788	0.862	0.686	0.764	0.738	0.89	0.807
	EANN ^[11]	0.827	0.835	0.783	0.811	0.792	0.902	0.843
	去掉知识特征	0.848	0.914	0.773	0.837	0.799	0.926	0.857
	我们的方法	0.871	0.844	0.916	0.879	0.905	0.826	0.864
	仅视觉特征	0.674	0.812	0.558	0.663	0.581	0.834	0.685
推特数据集	仅文本特征	0.66	0.707	0.70	0.703	0.599	0.606	0.602
	仅知识特征	0.62	0.61	0.937	0.74	0.687	0.19	0.297
	VQA ^[34]	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	NeuralTalk ^[35]	0.61	0.728	0.504	0.595	0.534	0.752	0.625
	Att-RNN ^[10]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN ^[11]	0.715	0.822	0.638	0.719	0.647	0.779	0.709
	去掉知识特征	0.703	0.801	0.644	0.714	0.619	0.784	0.691
	我们的方法	0.735	0.816	0.699	0.753	0.658	0.786	0.716

4.5 知识特征分析

在本节中, 我们给出了更详细的知识特征提取模块的说明和分析。

在两个数据集中, 句子的长度分布并不相同。大多数微博数据集中的句子长度要远远长于推特数据集中的句子长度。因此, 在知识特征提取过程中, 我们在两个数据集上分别设置不同大小的知识邻接关系矩阵。在微博数据集上使用的邻接关系矩阵 A 大小为 128×128 , 并且 $x \in \{0, 1\}$ 。在推特数据集上, A 的大小为 56×56 , $x \in (-1, 1)$ 。 x 是邻接关系矩阵 A 中的元素。

在微博数据集上, 我们首先分析了谣言类和非谣言类的新闻的邻接关系矩阵 A 的统计特性。在

微博数据集上 $x_{ij} = 1$ 表示在知识图谱中存在同时包含第 i 个单词与第 j 个单词的三元组。我们分别计算谣言类和非谣言类中 A 的元素的均值。在谣言类新闻中, 矩阵的均值为 0.008。而在非谣言类中, 该均值为 0.011。可以看出, 新闻的邻接关系矩阵是一个较为稀疏的矩阵。同时, 非谣言类的矩阵均值大于谣言类的均值, 说明了真新闻中的实体之间的相关性要强于假新闻之间的实体相关性。在推特数据集上, 真新闻的矩阵的元素均值高于假新闻 (37.5%)。由此我们可以推断出, 新闻的邻接关系矩阵 A 在真假两类新闻上分布不同, 具有一定的辨别能力。

接着, 我们分析知识特征在不同长度的新闻上

的识别能力。通常来说,人们会认为更长的句子含有更多的信息。由于微博数据集的新闻文本较长,在本节中,通过设置最小文本长度来划分数据集,并训练和测试我们的知识特征检测方法在不同类别上的效果。由于在微博数据集上,文本中词语数超过 100 的句子数量很少,因此,我们将最大的类别的长度设为 80。这里的计数单位是词语数量,即经过文本分割后的词语数,而不是字符的数量。我们设置 5 类长度不同的数据集,为 $sentence_length \in \{> 2, > 20, > 40, > 60, > 80\}$ (当单词数小于 2 时 A 是一个零矩阵,因此,考察 A 的效果时,从长度大于 2 开始)。 $sentence_length > 2$ 表示过滤掉数据集中单词数小于等于 2 的文本后留下的子集。实验结果如表 3 所示。

表 3 仅使用知识特征在不同的微博数据集的子集上的实验结果

Table 3 The experimental results of only using knowledge feature on different subsets of Weibo dataset				
长度	准确率	精确率	召回率	精准性
> 2	0.67	0.70	0.67	0.66
> 20	0.69	0.70	0.69	0.68
> 40	0.72	0.72	0.72	0.72
> 60	0.73	0.73	0.74	0.73
> 80	0.78	0.80	0.77	0.78

从表 3 可看出,随着文本长度的增加,用知识特征识别新闻的准确率也逐步提升。当子集中文本长度大于 2 时,仅使用知识特征的虚假新闻分类准确率为 67%。当文本长度大于 80 时,分类准确率达到 78%。这说明随着文本长度增加,邻接关系矩阵 A 能更有效地提取出文本中的信息。

接着,我们研究使用不同的神经网络的组件对知识特征效果的影响。在知识特征提取模块中在微博数据集上使用的卷积神经网络结构如表 4 所示。

我们使用了 2 维卷积层,2 维的平均池化层和激活层 ReLU。我们同样尝试了使用 2 维最大化池化层,和正则化组件。实验结果如图 4 所示。我们可以看出,当使用最大池化层代替平均池化层时,检测的准确率有所下降。联系到本小节开头的邻接关系矩阵 A 的平均值的计算中谣言类的平均值小于非谣言类,我们可以有所推断。使用最大池化层造成准确率下降是由于在池化过程中仅关注了窗口中最大值的信

息,而忽略了窗口中其他位置的信息,造成了信息的丢失。而在邻接关系矩阵中相邻位置的元素不同于视觉图像中像素具有连续性,邻接关系矩阵中的元素是较为稀疏的。因此,平均池化操作更为适用。

表 4 在微博数据集上知识特征提取网络配置
Table 4 ConvNet configurations of knowledge feature extraction for the Weibo dataset

ConvNet 配置	
Conv 7×7, 16, stride 3	
ReLU	
3×3 Avg Pool, stride 2	
Conv 3×3, 32, stride 2, padding 1	
ReLU	
3×3 Avg pool, stride 2, padding 1	
Conv 3×3, 64, stride 1	
ReLU	
FC-16	
FC-2, Softmax	

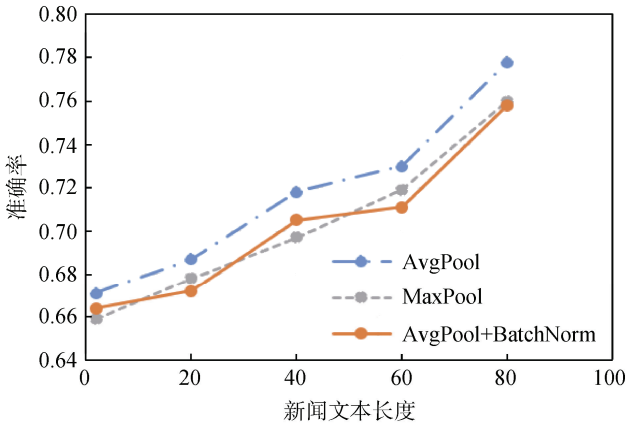


图 4 在知识特征提取网络中使用不同的组件,在微博数据集上的比较

Figure 4 Comparison of using different components in knowledge feature extraction networks on Weibo dataset

同理,在添加了正则化操作之后,模型的准确率也有所下降。说明在正则化过程中减少了邻接关系矩阵的区分度。从上面两个例子也可以说明,我们的模型提取出的知识特征通过处理邻接关系矩阵保留了不同类别的新闻的内在的知识和关系。

在推特数据集上,我们使用的类似的网络结构提取知识特征。由于推特数据集上的文本长度远短于微博数据集上的文本长度,我们也使用了不同大小的邻接关系矩阵,为 $A \in R^{56 \times 56}$ 。其对应的知识特征提取网络配置如表 5 所示。

表 5 在推特数据集上知识特征提取网络配置
Table 5 ConvNet configurations of knowledge feature extraction for the Twitter dataset

ConvNet 配置
Conv 5×5, 16, stride 2, padding 1
ReLU
3×3 Avg Pool, stride 2
Conv 3×3, 32, stride 2
ReLU
Conv 3×3, 64, stride 1
ReLU
FC-16
FC-2, Softmax

由于推特数据集上的文本长度有限,可能会造成邻接关系矩阵 A 十分稀疏。为了避免这种情况发生,在推特数据集上,我们不直接匹配实体在知识图谱中的三元组,而是利用在知识图谱上的知识嵌入(Knowledge Embedding, KE)技术,来计算两个节点之间的关系。此过程中,我们直接使用 Google 发布的已训练完成的 Freebase^[34]命名空间中的词向量,来计算两个实体之间的关系。邻接关系矩阵中元素

由计算两个向量之间的余弦距离得到。

在推特数据集上,仅用知识特征进行虚假新闻检测的实验结果如表 2 所示。我们可以看出,在推特数据集上仅使用知识特征进行分类达到了和仅使用文本特征和仅使用图片特征相似的效果。仅使用知识特征、仅文本特征和仅使用图片特征的检测准确率分别是 0.62%、0.66%和 0.674%。并且,在这 3 种特征中,知识特征在谣言类数据上得到了最高的 F_1 值 0.74%。

最后,我们进行了知识特征的消融实验。分别计算了在两个数据集上的实验结果。我们给出了 3 组消融实验的比对结果。分别是 1) 仅文本特征对比文本&知识特征; 2) 仅图像特征对比图像&知识特征; 3) 文本&图像特征对比文本&图像&知识特征。实验结果如图 5 所示。

左边蓝色柱表示没有使用知识特征的虚假新闻检测结果,右边橙色柱表示融入知识特征后的检测结果。图 5 中我们给出了在准确率和 F_1 值上的对比。我们可以看到,在各种情况下,融合知识特征均能得到更好的实验结果。

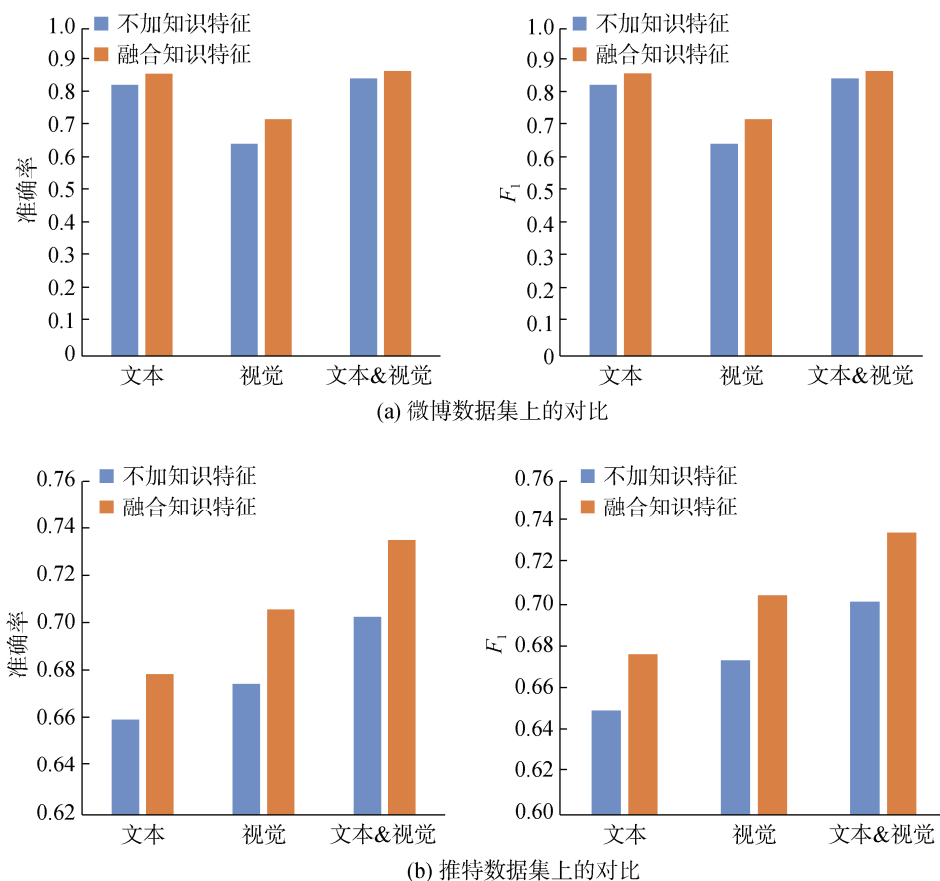


图 5 使用与未使用知识特征的模型效果对比

Figure 5 Performance comparisons for models with and without the knowledge feature extraction

以上的实验结果说明了文本所提出的知识特征提取的方法的有效性。

4.6 实例分析

本小节中,给出了一些使用文本&图像不能正确检测出而加入知识特征能检出的虚假新闻实例。

图 6 中展示了一些仅通过加入知识特征才能识别出的虚假新闻实例。如图 6(a)中第一个实例所示:在该条假新闻中,出现了较多的医疗病症相关内容,且图片与文字内容相关性弱。融入知识信息后能判断出其中信息的虚假成分,将其识别为假

新闻。在第二个实例中,赚钱、板凳、钉子出现在同一新闻中,新闻的整体统计特征上,实体相关性不大,增加了识别为假新闻的概率。图 6(b)中展示了部分推特数据集上的实例。例如,“不可置信的黑色狮子”,其图片搭配的是黑色狮子,新闻的图片与文本一致,仅使用图片和文本信息也不能将其识别出来。结合知识信息后,可判断出该新闻不符合常识,将其识别为虚假信息。从上述例子可以看出,引入知识特征后,能有效提高虚假新闻的识别效果。



图 6 一些被文本&图像模型漏掉而被我们方法检出的虚假新闻实例
Figure 6 Some fake news detected by our method but missing by text-image model

5 结论

在文本中,我们提出了一种融合知识的多模态虚假新闻检测方法。为了能更好地利用外部大型开放知识图谱中丰富的信息,我们提出了一种简单鲁棒的方法来提取新闻中的知识特征。我们将新闻看作是一个图,新闻文本经过文本分割后得到的词语

则看成图中的节点。将新闻中的节点匹配到知识图谱中的三元组,计算出两两节点之间的关系,从而形成能够表征新闻知识内容的邻接关系矩阵。接着,我们使用基于 CNN 的神经网络处理新闻的邻接关系矩阵,得到新闻的知识特征表示。将其与新闻中其他模态的特征相融合,得到了包含文本、视觉和知识 3 个层面的联合特征表示。最终,将联合特征输入到多

层感知机分类器中得到虚假新闻的检测结果。在微博和推特两个真实世界的数据集上, 我们的方法都得到了目前最优的虚假新闻检测结果。同时, 丰富的实验结果表明, 文本提出的知识特征能和其他的特征结合, 显著地提高虚假新闻检测的效果。证明了知识特征的有效性。下一步工作中, 我们将继续探索文本的知识特征在其他的文本分类任务中的应用。

参考文献

- [1] Sharma K, Seo S, Meng C Z, et al. COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations[EB/OL]. 2020: arXiv: 2003.12309. <https://arxiv.org/abs/2003.12309>.
- [2] Jin Z W, Cao J, Jiang Y G, et al. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model[C]. *2014 IEEE International Conference on Data Mining*, 2015: 230-239.
- [3] Jin Z W, Cao J, Zhang Y D, et al. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs[C]. *The Thirtieth AAAI Conference on Artificial Intelligence*, 2016: 2972-2978.
- [4] Zhou X, Zafarani R. Fake news: A survey of research, detection methods, and opportunities[EB/OL]. 2018: ArXiv Preprint ArXiv: 1812.00315.
- [5] Bondielli A, Marcelloni F. A Survey on Fake News and Rumour Detection Techniques[J]. *Information Sciences*, 2019, 497: 38-55.
- [6] Granik M, Mesyura V. Fake News Detection Using Naive Bayes Classifier[C]. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering*, 2017: 900-903.
- [7] Shiralkar P, Flammini A, Menczer F, et al. Finding Streams in Knowledge Graphs to Support Fact Checking[EB/OL]. 2017: arXiv: 1708.07239. <https://arxiv.org/abs/1708.07239>.
- [8] Shu K, Sliva A, Wang S H, et al. Fake News Detection on Social Media[J]. *ACM SIGKDD Explorations Newsletter*, 2017, 19(1): 22-36.
- [9] Wang W Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection[EB/OL]. 2017: arXiv: 1705.00648. <https://arxiv.org/abs/1705.00648>.
- [10] Jin Z W, Cao J, Guo H, et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs[C]. *The 25th ACM international conference on Multimedia*, 2017: 795-816.
- [11] Wang Y Q, Ma F L, Jin Z W, et al. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection[C]. *The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 849-857.
- [12] Qi P, Cao J, Yang T Y, et al. Exploiting Multi-Domain Visual Information for Fake News Detection[C]. *2019 IEEE International Conference on Data Mining*, 2020: 518-527.
- [13] Jin Z W, Cao J, Wang B, et al. Rumor Detection on Social Media with Multimodal Feature Fusion[J]. *Journal of Nanjing University of Information Science & Technology (Natural Science Edition)*, 2017, 9(6): 583-592.
(金志威, 曹娟, 王博, 等. 融合多模态特征的社会多媒体谣言检测技术研究[J]. *南京信息工程大学学报(自然科学版)*, 2017, 9(6): 583-592.)
- [14] Ishida Y, Kuraya S. Fake news and its credibility evaluation by dynamic relational networks: A bottom up approach[J]. *Procedia Computer Science*, 2018, 126: 2228-2237.
- [15] Chen X Q, Zhou F, Zhang K P, et al. Information Diffusion Prediction via Recurrent Cascades Convolution[C]. *2019 IEEE 35th International Conference on Data Engineering*, 2019: 770-781.
- [16] Wu K, Yang S, Zhu K Q. False Rumors Detection on Sina Weibo by Propagation Structures[C]. *2015 IEEE 31st International Conference on Data Engineering*, 2015: 651-662.
- [17] Khattar D, Goud J S, Gupta M, et al. MVAE: Multimodal Variational Autoencoder for Fake News Detection[C]. *WWW '19: The World Wide Web Conference*, 2019: 2915-2921.
- [18] Pan J Z, Pavlova S, Li C X, et al. Content Based Fake News Detection Using Knowledge Graphs[C]. *International Semantic Web Conference*. Cham: Springer, 2018: 669-683.
- [19] Castillo C, Mendoza M, Poblete B. Information Credibility on Twitter[C]. *The 20th International Conference on World Wide Web*, 2011: 675-684.
- [20] Ruchansky N, Seo S, Liu Y. CSI: A Hybrid Deep Model for Fake News Detection[C]. *The 2017 ACM on Conference on Information and Knowledge Management*, 2017: 797-806.
- [21] Morris M R, Counts S, Roseway A, et al. Tweeting is Believing? : Understanding Microblog Credibility Perceptions[C]. *The ACM 2012 conference on Computer Supported Cooperative Work*, 2012: 441-450.
- [22] Gupta M, Zhao P X, Han J W. Evaluating Event Credibility on Twitter[C]. *The 2012 SIAM International Conference on Data Mining*, 2012: 153-164.
- [23] Jin Z W, Cao J, Zhang Y D, et al. Novel visual and statistical image features for microblogs news verification[J]. *IEEE Transactions on Multimedia*, 2017, 19(3): 598-608.
- [24] Ciampaglia G L, Shiralkar P, Rocha L M, et al. Correction: Computational fact checking from knowledge networks[J]. *PLoS ONE*, 2015, 10(10): e0141938.
- [25] Tang D Y, Qin B, Liu T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification[C]. *The 2015 Conference on Empirical Methods in Natural Language Processing*, 2015: 1422-1432.
- [26] Kim Y. Convolutional Neural Networks for Sentence Classification[EB/OL]. 2014: arXiv: 1408.5882. <https://arxiv.org/abs/1408.5882>.
- [27] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[EB/OL]. 2014: arXiv: 1404.2188. <https://arxiv.org/abs/1404.2188>.
- [28] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL]. 2014: arXiv: 1409.1556. <https://arxiv.org/abs/1409.1556>.
- [29] Boididou C, Papadopoulos S, Kompatsiaris Y, et al. Challenges of computational verification in social multimedia[C]. *The 23rd International Conference on World Wide Web*, 2014: 743-748.
- [30] Song Y, Shi S, Li J, et al. Skip-gram: Explicitly distinguishing left and right context for word embeddings[C]. *The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2018: 175-180.

- [31] Tomas M, Kai C, Greg C, et al. Efficient estimation of word representations in vector space[C]. *Proceedings of Workshop at The International Conference on Learning Representation*, <https://code.google.com/archive/p/word2vec/>. 2013: 147-158.
- [32] Deng J, Dong W, Socher R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C]. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 248-255.
- [33] Yener. Ownthink-knowledge graph data[Z]. <https://github.com/ownthink/KnowledgeGraph>. 2019.
- [34] Freebase, <https://developers.google.com/freebase>. 2020.
- [35] Agrawal A, Lu J S, Antol S, et al. VQA: Visual Question Answering[J]. *International Journal of Computer Vision*, 2017, 123(1): 4-31.
- [36] Vinyals O, Toshev A, Bengio S, et al. Show and Tell: A Neural Image Caption Generator[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 3156-3164.



王茜 于 2016 年在中国科学院计算技术研究所计算机应用技术专业获得博士学位。现任中国科学院微电子研究所副研究员。该论文为作者在中国科学院信息工程研究所工作期间的研究工作。研究领域为多媒体智能信息处理、知识图谱。研究兴趣包括知识图谱表示学习、跨媒体虚假新闻检测等。Email: wangxiboss@163.com



韩冀中 于 2001 年在中国科学院计算技术研究所获得博士学位, 现为中国科学院信息工程研究所第三研究室正研级高工、博士生导师。主要研究领域为大数据存储与管理、多媒体信息智能化处理。研究兴趣包括多媒体内容理解、深度伪造。Email: hanjizhong@iie.ac.cn