引入知识表示的图卷积网络谣言检测方法*

郭秋实†，李晨曦，刘金硕
（武汉大学 国家网络安全学院 空天信息安全与可信计算教育部重点实验室，武汉 430072）

摘 要： 在谣言检测的问题上，现有的研究方法无法有效地表达谣言在社交网络传播的异构图结构特征，并且没有引入外部知识作为内容核实的手段。因此，提出了引入知识表示的图卷积网络谣言检测方法，其中知识图谱作为额外先验知识来帮助核实内容真实性。采用预训练好的词嵌入模型和知识图谱嵌入模型获取文本表示后，融合图卷积网络的同时，能够在谣言传播的拓扑图中更好地进行特征提取以提升谣言检测的精确率。实验结果表明，该模型能够更好地对社交网络中的谣言进行检测。与基准模型的对比中，在 Weibo 数据集上的精确率达到 96.1%，在 Twitter15 和 Twitter16 数据集上的 F_1 值分别提升了 3.1% 和 3.3%。消融实验也表明了该方法对谣言检测皆有明显提升效果，同时验证了模型的有效性和先进性。

关键词： 知识表示；图卷积网络；谣言检测；知识图谱
中图分类号： TP391 **文献标志码：** A **文章编号：** 1001-3695(2022)07-017-2032-05
doi:10.19734/j.issn.1001-3695.2022.01.0003

Rumor detection with knowledge representation and graph convolutional network

Guo Qiushi†, Li Chenxi, Liu Jinshuo
(*Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan* 430072, *China*)

Abstract: Existing research methods have not effectively expressed the structural features of the heterogeneous graph of rumors and not introduced external knowledge as a means of content verification. Therefore, this paper proposed a rumor detection method with knowledge representation and graph convolutional network. It introduced the knowledge graph as additional prior knowledge to verify the authenticity of the content. After applying pre-training word embedding and knowledge graph embedding models to obtain text representation and fusing graph convolutional network simultaneously, it could extract features well in the topological graph of rumor propagation to improve the accuracy of rumor detection. Compared with the baseline methods, experimental results show that the proposed model reaches 96.1% at precision on Weibo, 3.1% improvement at F_1 on Twitter15 and 3.3% improvement at F_1 on Twitter16 respectively. The ablation experiment also shows this method has a significant improvement on rumor detection, and simultaneously verifies the effectiveness and progressiveness of this model.

Key words: knowledge representation; graph convolutional network; rumor detection; knowledge graph

0 引言

过去的数年，互联网用户数量一直保持着激增的势头，国内新浪微博和国外推特等社交媒体被广泛应用。社交媒体中的谣言具有传播速度快、涉及范围广且辟谣成本大的特点，严重影响了网民获取信息的真实性以及社会网络健康生态[1]。近年来，国内外研究团队针对谣言识别这一研究热点提出了许多理论和方法。传统的检测方法主要采用了有监督学习的思路，例如人工构造用户特征、文本内容和谣言传播模式。Qazvinia 等人[2] 对 Twitter 中产生的谣言进行检测，选取了推特内容特征、用户行为特征和推特标签（"#"、短链接等）特征。Cao 等人[3] 将谣言信息的研究分为了基于手工特征、基于传播以及基于深度神经网络的研究方法。然而这些基于人工构造特征的模型耗时耗力且忽略了先验知识的引入，无法提取更精细的文本特征。

随着深度学习的高速发展，有学者使用深度学习方法从社交媒体谣言传播路径或者网络中，捕获谣言随时间传播的序列特征，以构建时间序列特征模型。Jin 等人[4] 提出了一种新颖的递归神经网络保持机制，融合文本和社交网络特征进行有效的谣言检测。Bao 等人[5] 利用 BP 神经网络模型对微博谣言进行检测，选取了相关微博的数量、客观评论数和话题类型，通过引入参数调控的激活函数和加快网络学习速度的冲量项来提升模型效果。文献[6] 对于带有疑问的帖子进行聚类，进而构建谣言分类器，通过社交媒体转发的跟帖中的质疑信号来判断谣言与否。Yang 等人[7] 通过多重卷积提取的文字和图像的潜在特征，并最终结合图像和文字的显示特征，将所有特征投影到统一的特征空间中进行训练分类。虽然上述深度神经网络的方法可以得到局部邻域内的相关特征，但是不能处理图或树上的全局结构关联，即忽视了谣言散布的全局结构特征。

基于此，本文充分考虑了引入知识图谱作为额外先验知识的重要性，并针对社交网络的异构图结构使用图卷积挖掘节点传播行为，以更好地进行文本表示。本文提出的方法在三个开源数据集上的检测结果均有明显提升。模型的整体框架如图 1 所示。

收稿日期：2022-01-02；修回日期：2022-03-01 基金项目：国家自然科学基金资助项目(U193607)；国家重点研发计划资助项目(2020FYA0607902)

作者简介：郭秋实(1997-)，男(通信作者)，吉林长春人，硕士研究生，主要研究方向为自然语言处理和数据挖掘(1415142293@qq.com)；李晨曦(1995-)，男，湖北武汉人，硕士，主要研究方向为人工智能和知识图谱；刘金硕(1974-)，女，吉林辽源人，教授，博导，博士，主要研究方向为信息安全和数据挖掘.

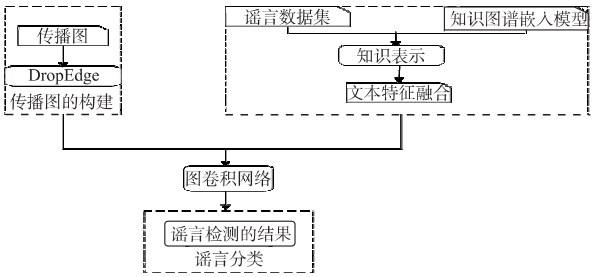

图 1　模型的整体框架

Fig. 1　Overall structure of the model

1　引入知识图谱的文本表示

在社交媒体谣言检测任务中，本文考虑了外部知识的引入，添加知识图谱作为额外的先验知识。在知识图谱嵌入的模型基础上，首先提取文本的实体表示，并进行实体嵌入和文本词嵌入。接着融合多头注意力层获取文本高阶表示，最后将两种表示结果相结合。

1.1　知识图谱嵌入

本文为引入知识图谱作为额外先验知识表示，提出了生成式对抗网络的知识图谱嵌入模型[8]。其中生成式对抗网络的生成器产生负样本集合并产生候选实体的类型；判别器则同时接受正负样本并计算它们的得分。

具体地，本文选取了 TransE 模型来作为独立的生成器和判别器。TransE 是基于实体和关系的向量表示，并利用嵌入向量之间的距离进行建模[9]。给定一组正样本三元组实例向量 $(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$，分别表示头部实体向量、关系向量和尾部实体向量，核心目标是使 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$，即模型的打分函数如下所示。

$$f_r(\boldsymbol{h},\boldsymbol{t}) = \|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_2 \tag{1}$$

其中：$\|x\|$ 表示欧几里德范数。

当生成器以正样本作为输入时，本文按照三元组的类型来决定替换头部实体和尾部实体的概率。当三元组类型为多对一时，即替换尾部实体；当三元组类型为一对多时，即替换头部实体[10]。形式化表示来说，对于随机采样的包含正负样本的候选实体集合，生成器初始化一个 TransE 模型，利用式(1)的打分函数计算所有候选正样本得分。最后通过归一化表示，得到对于正样本的采样概率，如式(2)所示。

$$p(t_i|(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})) = \frac{p(t_i|(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}))}{\sum_j p(t_j|(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}))} \tag{2}$$

其中：t_i 和 t_j 表示候选实体集合中不同的尾部实体。对于计算得到的所有采样概率，概率最大值即正样本输出。

对于采用 TransE 模型的判别器来说，力求正样本的打分函数 $f_r(\boldsymbol{h},\boldsymbol{t})$ 迫近于 0 而负样本的打分函数 $f_r(\boldsymbol{h}',\boldsymbol{t}')$ 不为 0[11]，这样可以使模型在训练过程中有效清晰地分离正负样本。然而，对于效果优秀的生成器来说，其产生的正样本集合 $(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$ 对应的打分函数 $f_r(\boldsymbol{h},\boldsymbol{t})$ 应当更小，这是因为模糊难以辨别的正负样本可以提高模型的实例嵌入效果，为后续的知识表示提供良好的预训练效果。

1.2　知识表示

在知识表示部分，本文采用预训练好的词嵌入模型和知识图谱嵌入模型分别进行文本词向量表示和文本的实体表示；然后融合多头注意力层提取词向量和实体向量的高阶表示；最后结合谣言传播图输送到图卷积网络进一步特征提取。

对于给定的社交媒体输入文本词序列 $T=t_1,t_2,\cdots,t_n$，其中 n 为本序列长度。首先获得文本的词嵌入结果 $\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2^{\mathrm{T}},\cdots,\boldsymbol{x}_n^{\mathrm{T}}$，其中 $\boldsymbol{x}_i^{\mathrm{T}}\in\mathbb{R}^{d_t}$。联结节点词嵌入和位置嵌入作为嵌入层的表达式为 $\boldsymbol{X}=\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_n$，$\boldsymbol{x}_i=\boldsymbol{x}_i^{\mathrm{T}}\oplus\boldsymbol{x}_i^{P}$，其中⊕为向量拼接运算；$\boldsymbol{X}^p=\boldsymbol{x}_1^p\boldsymbol{x}_2^p\cdots\boldsymbol{x}_n^p$，$\boldsymbol{x}_i^P\in\mathbb{R}^{d_p}$ 代表位置嵌入表示结果。接着 $\boldsymbol{X}$ 编码层的输入，编码为多头注意力机制。本模型使用的多头注意力机制计算公式如下：

$$\hat{\boldsymbol{H}}=\hat{\boldsymbol{h}}_1\hat{\boldsymbol{h}}_2\cdots\hat{\boldsymbol{h}}_n=(MH_1\oplus MH_2\oplus\cdots\oplus MH_n)\boldsymbol{T}_{\hat{H}} \tag{3}$$

其中：$\hat{\boldsymbol{h}}_1\hat{\boldsymbol{h}}_2\cdots\hat{\boldsymbol{h}}_n$ 为编码结果；$\boldsymbol{T}_{\hat{H}}$ 是参数矩阵；MH_i 为第 i 个注意力机制的输出结果，其计算结果如式(4)所示。

$$MH_i=\mathrm{softmax}\left(\frac{QT_i^Q(KT_i^K)^{\mathrm{T}}}{\sqrt{d_x}}\right)VT_i^V \tag{4}$$

其中：$Q=K=V=X$；$\sqrt{d_x}$ 为特征维数。然后在多头注意力机制的输出 $\hat{H}$ 基础之上，引入两层全连接前馈神经网络，该全连接层的表达式如下：

$$\tilde{\boldsymbol{H}}=\tilde{\boldsymbol{h}}_1\tilde{\boldsymbol{h}}_2\cdots\tilde{\boldsymbol{h}}_n=\sigma(\hat{\boldsymbol{H}}\boldsymbol{T}_{H_1}+\boldsymbol{b}_{H_1})\boldsymbol{T}_{H_1}+\boldsymbol{b}_{H_2} \tag{5}$$

其中：$\boldsymbol{T}_{H_1}$、$\boldsymbol{T}_{H_2}$、$\boldsymbol{b}_{H_1}$、$\boldsymbol{b}_{H_2}$ 为参数矩阵和偏置矩阵；$\tilde{\boldsymbol{h}}_1\tilde{\boldsymbol{h}}_2\cdots\tilde{\boldsymbol{h}}_n$ 为全连接层表达结果；$\sigma(x)=\max(0,x)$ 表示 ReLU 激活函数。

2　融合图卷积网络的谣言检测方法

由于社交媒体谣言传播路径为一种异构图结构[12]，所以，本文利用图卷积网络来获取社交媒体谣言的传播特征。融合图卷积网络的谣言检测方法如图 2 所示。

2.1　建立传播图

定义社交媒体谣言检测数据集为集合 $C=\{c_1,c_2,\cdots,c_m\}$，其中 c_i 第 i 个谣言事件，m 表示数据集中的谣言事件数量。每个谣言事件 c_i 具体地表示为 $c_i=\{m_i,n_1^i,n_2^i,\cdots,n_{n_i-1}^i,G_i\}$，其中 m_i 表示谣言事件的原始帖子，n_j^i 表示第 j 个转发帖子，n_i 表示谣言事件 c_i 包含的初始帖子数量，G_i 为谣言传播图。传播图 G_i 形式化表示为 $G_i=\{E_i,V_i\}$，其中节点集合为 $V_i=\{m_i,n_1^i,n_2^i,n_{n_i-1}^i\}$，边集合表示为 $E_i=\{b_{pq}^i\,|\,p,q=0,1,\cdots,n_i-1\}$。在传播图的表示中，$m_i$ 代表传播图的根节点，b_{pq}^i 表示节点之间社交媒体的传播状态。例如，节点 n_1^i 与 n_2^i 的微博或者推特存在转发或者评论状态，在传播图中则存在一条有向边 b_{21}^i。

本文对于每个谣言事件 c_i 设定了一个分类标签 v_i，定义其标签 $v_i\in\mathcal{L}$，其中 $\mathcal{L}=\{N,F,T,U\}$ 为标签集合，N 为非谣言事件，F 为假谣言事件，T 为真谣言事件，U 表示被证实的谣言事件。因此，社交媒体谣言检测任务可以表示为一个分类任务，即

$$f:C\rightarrow\mathcal{L} \tag{6}$$

其中：C 表示社交媒体谣言事件集合；$\mathcal{L}$ 表示标签集合；f 为一个判别模型。

对于传播图 $G_i=\{E_i,V_i\}$，构建其邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{N\times N}$，$N$ 为节点数目，然后对于每个节点刻画其特征表示。对于传播图中的任意节点，其特征表示包含了用户特征和节点特征两种特征模式[13]。其中，用户特征属于用户自有属性，这些属性可以在一定程度上反映该用户的信誉程度、账号使用风格等。具体包含：a) 用户是发帖原用户还是转发用户；b) 用户等级；c) 用户粉丝量；d) 用户发帖量；e) 用户关注量。

在谣言传播图矩阵当中，邻接矩阵 A 涵盖了从根节点到下游节点的传播文本信息。为了缓解过拟合的问题，本文采用去边(DropEdge)[14]操作，即在每一次训练过程中，随机在输入图信息当中除去一些边用来形成不同的传播图，从而提高模型的泛化效果。形式化表示为对于谣言传播图的邻接矩阵 $\boldsymbol{A}$，在 DropEdge 操作之后，新的邻接矩阵 $\boldsymbol{A}'$ 可表示为

$$\boldsymbol{A}'=\boldsymbol{A}-\boldsymbol{A}_p \tag{7}$$

其中：$\boldsymbol{A}_p$ 为被去边之后的边矩阵的集合，意义为在原始图结构中以概率 p 随机去掉边以构成新的图结构。因此，邻接矩阵 $\boldsymbol{A}'$ 和其特征矩阵 $\boldsymbol{X}\in\mathbb{R}^{N\times C}$ 可以作为图卷积网络的特征输入。

其中矩阵的行数即节点个数为 N，C 代表向量维度。

2.2 图卷积网络结构

由于谣言在社交网络特有的拓扑图传播结构，图卷积网络恰好可以表示节点与节点之间的边形成的拓扑图[15]。本文使用的图卷积网络如图 3 所示，其输入层由特征矩阵 $X \in \mathbb{R}^{N \times C}$ 和邻接矩阵 $A' \in \mathbb{R}^{N \times N}$ 组成。

图卷积网络最后的输出层为节点输出，是一个向量矩阵 $Z \in \mathbb{R}^{N \times g}$，$g$ 为分类模型中的分类数目。在谣言传播的图卷积结构中，经激活函数输出的输出层可用非线性函数来表示，具体形式如下：

$$Z^{(d+1)} = g(Z^{(d)}, A') = \sigma(\hat{A}Z^{(d)}T^{(d)}) = \hat{A}\,\sigma(\hat{A}XT^{(d)})\,T^{(d+1)} \tag{8}$$

其中：d 是神经网络层数，$\sigma(x)$ 为非线性激活函数，根据模型构造可调整为 ReLU 函数或 sigmoid 函数，T 为权重矩阵，$\hat{A}$ 表示邻接矩阵 A' 归一化后的矩阵结果。而 $\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} = \tilde{D}^{-\frac{1}{2}}(A' + I)\tilde{D}^{-\frac{1}{2}}$，$\tilde{D}$ 为矩阵 $\hat{A}$ 的对角矩阵，I 为单位矩阵。

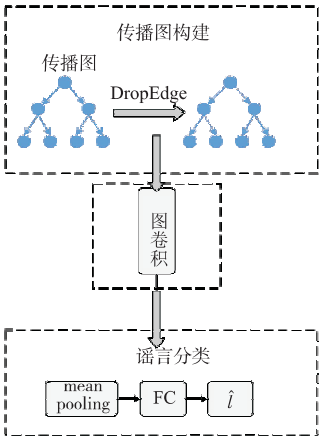

图 2 融合图卷积的谣言分类模型

Fig. 2 Rumor classification model with graph convolution

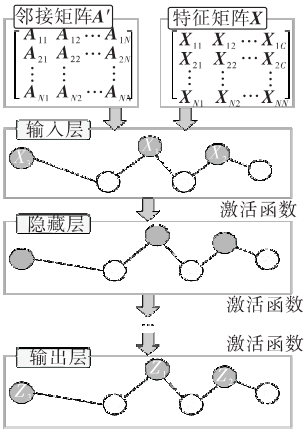

图 3 图卷积网络结构示意图

Fig. 3 Diagram of graph convolutional network

2.3 融合图卷积的谣言分类

本文利用知识表示和图卷积网络来刻画谣言分类模型，用谣言事件属于标签的最大概率结果来判别是否为谣言。由于图卷积输出的特征矩阵 Z 包含长度不一的文本序列长度，所以，本文对图卷积层输出的特征矩阵进行均值池化以获取所有节点的传播信息。均值池化公式如下所示。

$$M = \text{meanpooling}(Z) \tag{9}$$

其中：$M \in \mathbb{R}^{1 \times m_k}$ 表示在图卷积网络上均值池化的结果；m_k 为第 k 层图卷积输出的特征维度。

为了后续的谣言判别分类任务，本文在均值池化的基础上添加一个全连接层，将特征向量映射成 $|\mathcal{L}|$ 维度的特征，其中 $|\mathcal{L}|$ 为谣言标签集合 $\mathcal{L}$ 的大小。全连接层的表达式如下：

$$a = \sigma(W_a M^f + b_a) \tag{10}$$

其中：W_a 和 b_a 分别表示参数矩阵和偏置向量，$\sigma(x)$ 为 sigmoid 激活函数，$a \in \mathbb{R}^{1 \times |\mathcal{L}|}$ 为预测标签类别的概率。为了方便表示预测概率结果，本文使用 softmax 函数进行归一化，得到归一化后的概率结果 $\hat{a} = \text{softmax}(a)$。那么找到概率最大的标签类别 $\hat{l}$ 即可判别事件是否为谣言，即需满足 $\hat{l} = \text{argmax}(\hat{a})$。

在谣言分类模型中，本文使用交叉熵损失函数 ℓ 来刻画整个模型的损失函数，损失函数的数学表达式如下：

$$\mathcal{L} = -\sum_i \sum_j Y_{ij} \ln(\hat{a}_{ij}) \tag{11}$$

其中：模型的优化目标为最小化损失函数 ℓ。$Y_{ij} \in \{0,1\}^{|\mathcal{L}|}$ 表示谣言事件 c_i 对应的真实标签，$\hat{a}_{ij}$ 表示对于谣言事件 c_i，模型预测其标签为 j 的概率值。

3 实验与分析

3.1 实验数据集

本文选择三种社交媒体谣言数据集作为实验数据集。第一个数据集是 Ma 等人[16]创建的 Weibo 数据集，经整理该数据集包含 2 212 个谣言事件和 2 248 个非谣言事件；第二个数据集是 Liu 等人[17]创建的 Twitter15 数据集，经整理其中包括 94 个非谣言事件和 446 个谣言事件；第三个数据集是 Ma 等人[18]创建的 Twitter16 数据集，经整理其中包含了 499 个谣言事件和 493 个非谣言事件。在这三种数据集中，节点表示社交媒体用户，边表示转发或者评论之间的关系。其中，Twitter15 和 Twitter16 数据集的每个事件的标签都是根据辟谣网站中文章的真实性标签标注（snopes. com 和 Emergent. info）。三种数据集包含的关键参数如表 1 所示。

表 1 三种数据集的关键信息

Tab. 1 Key information of three datasets

参数	Weibo	Twitter15	Twitter16
#posts	3 805 656	331 614	204 820
#events	4 460	1 289	1 400
#true rumors	2 212	446	499
#false rumors	2 248	94	493
#unproven rumors	0	375	203
#not rumors	0	374	205

3.2 创建实验知识图谱

在创建实验所需的知识图谱中，本文针对不同的数据集相应地构建了不同的知识图谱类型。针对中文 Weibo 数据集，本文利用 CN-DBPedia 的公开 API 获取数据创建中文知识图谱。针对英文 Twitter15 和 Twitter16 数据集，利用 DBPedia 创建相应的英文知识图谱。具体地，本文从谣言数据集提取谣言类型的实体，将实体连接到开放的知识图谱当中并提取与这些实体相连接的关系，基于提取的关系和实体创建实验所需的知识图谱类型。

在本文的实验当中，实验参数设置如下：使用随机梯度下降算法更新实验参数，并采用 Adam 算法来优化本文模型；在 Twitter15 和 Twitter16 数据集的学习率设置为 0. 000 2 和 0. 000 5，Weibo 数据集的学习率设置为 0. 000 5。每个节点的隐层特征向量维度为 64；在防止过拟合 DropEdge 操作中，下降率和丢弃率分别设置成 0. 2 和 0. 5，迭代次数为 200 轮，其中验证损失函数不再下降时使用早停法（early stopping）。

3.3 实验设置

3.3.1 评价指标

本文将社交媒体谣言检测任务视为二分类问题，为评估模型性能，采用经典的精确率（precision）、召回率（recall）和 F_1 值作为评价指标。其中三种评价指标的定义式如下：

$$precision = \frac{TP}{TP + FP} \tag{12}$$

$$recall = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{14}$$

其中：TP 表示预测标签和真实标签皆为谣言事件的概率；FP 表示真实标签为非谣言事件但预测标签为谣言事件的概率；FN 表示真实标签为谣言事件但预测标签为非谣言事件的概率。

3.3.2 基准方法

本文选取了在谣言检测领域中取得过突出效果的模型作为基准方法，并在三种数据集上进行了对比实验。对比实验所采用的基准方法如下：

a) Bi-PG[12]。由 Bian 等人提出了一种双向传播图结构的

模型，能够在更细微的文本特征上进行谣言检测任务。

b) PPC [13]。一种通过提取谣言传播过程的用户特征，并结合深度神经网络的谣言检测模型。

c) CIAM [19]。将用户可信度信息和注意力机制加入到谣言检测层当中，并使用多任务学习框架进行训练。

d) DTC [20]。由 Castillo 等人提出的一种采取人工手段提取谣言特征，并构建分类决策树来对谣言进行判别的分类方法。

e) RVNN [21]。该方法使用了树状递归神经网络在 Twitter15 和 Twitter16 数据集上建模来进行谣言检测任务。

f) MKN [22]。一种从真实世界的知识图谱中检索外部知识，以补充谣言帖子短文本的语义表示的谣言检测模型。

3.4　实验结果及分析

3.4.1　对比实验结果及分析

表 2 ~ 4 展示了本文的模型和基准方法在三种数据集上的对比实验结果。根据实验数据集的类型，本文在二分类类型的 Weibo 数据集中给出了真谣言事件和假谣言事件的准确率、召回率和 F_1 值；对于 Twitter15 和 Twitter16 数据集包含了四种不同形式的谣言类别，即非谣言事件、假谣言事件、真谣言事件和未被证实的谣言事件。因此，本文在 Twitter15 和 Twitter16 数据集上采取 F_1 值评测指标来全面刻画模型的性能。

表 2　Weibo 数据集的对比实验结果

Tab. 2　Contrast results of Weibo

算法	分类	*precision*	*recall*	F_1
Bi-PG	false rumors	0.903	0.926	0.900
	true rumors	0.624	0.763	0.697
PPC	false rumors	0.915	0.815	0.831
	true rumors	0.815	0.824	0.819
CIAM	false rumors	0.762	0.658	0.706
	true rumors	0.559	0.711	0.626
DTC	false rumors	0.832	0.795	0.791
	true rumors	0.802	0.813	0.807
RVNN	false rumors	0.904	0.910	0.938
	true rumors	0.893	0.908	0.907
MKN	false rumors	0.876	0.929	0.918
	true rumors	0.913	0.883	0.905
本文算法	false rumors	0.950	0.959	0.955
	true rumors	0.961	0.942	0.952

表 3　Twitter15 数据集对比实验结果

Tab. 3　Contrast results of Twitter15

算法	not rumors	false rumors	true rumors	unproven rumors
Bi-PG	0.804	0.758	0.822	0.771
PPC	0.496	0.584	0.469	0.638
CIAM	0.525	0.323	0.623	0.228
DTC	0.423	0.374	0.745	0.309
RVNN	0.679	0.799	0.819	0.664
MKN	0.712	0.519	0.667	0.671
本文算法	0.851	0.811	0.848	0.789

表 4　Twitter16 数据集对比实验结果

Tab. 4　Contrast results of Twitter16

本文算法	not rumors	false rumors	true rumors	unproven rumors
Bi-PG	0.792	0.809	0.821	0.797
PPC	0.551	0.715	0.367	0.685
CIAM	0.617	0.436	0.694	0.556
DTC	0.216	0.128	0.203	0.412
RVNN	0.678	0.753	0.840	0.712
MKN	0.724	0.642	0.765	0.778
本文算法	0.798	0.828	0.869	0.804

例如，在 Twitter16 数据集中，有这样一段假谣言事件："Hillary Clinton and her State department were actively arming Islamic jihadists，which includes ISIS."，从中可以提取到一组负类三元组（Hillary Clinton，actively arming，Islamic jihadists）。对应的真谣言事件为 "Hillary Clinton-led state department had approved weapon shipments to Libya during the intervention in 2011，and that those weapons had later ended up in the hands of jihadists."，这里提取到的三元组包括（Hillary State Department，had approved，weapon shipments）以及（weapon shipments，is to，Libya）。本文即利用生成式对抗网络的知识图谱嵌入模型对正负样本进行打分并计算采样概率，将概率最大化的输出值作为实体表示。在社交媒体的转发或评论中，本文通过构建传播图获取谣言的分类标签，并将本方案和取得良好效果的基准模型进行了对比。

根据对比实验结果，可以得出以下结论：

a) 相较于人工构造特征的方法，基于深度学习的模型在三种评测指标均有较高的准确率。这是因为人工构造的特征依赖性大且对数据不敏感，耗时耗力，导致最终的谣言检测效果不理想。针对目前流行的谣言检测模型，本文提出的引入知识表示的图卷积网络谣言检测方法在三种数据集上皆有明细提升效果。例如，在 Twitter16 数据集关于希拉里的谣言性质属于假谣言事件。针对希拉里假谣言事件所构成的三元组 (h,r,t)，采取预训练好的词嵌入模型和知识图谱嵌入模型获取文本表示，接着基于本文所构建的图卷积模型进行特征提取，在 F_1 值上达到了 0.828。其中，在 Weibo 数据集上真谣言事件的准确率达到了 96.1%，在 Twitter15 和 Twitter16 数据集上真谣言事件对比现有表现优越的基准模型在 F_1 值分别提升了 3.1% 和 3.3%。

b) CIAM 模型在规模更大的 Weibo 数据集上表现不佳，Weibo 数据集包含大量转发评论和跟帖，在线性传播结构中的谣言检测性能较差。RVNN 模型认为传播图中的所有节点贡献度相同，在传播过程中对新的传播节点敏感，导致模型无法精确提取更精细的文本特征，分类效果不理想。

c) 对比 Bi-PG 模型不难发现，该模型仅仅提取内容本身特征，缺乏外部知识的导入，容易引起词之间的歧义发生，容易导致正负类谣言分类的准确率变低。尽管 PPC 模型在社交媒体谣言传播中使用了线性结构进行高维特征提取，然而其忽略了谣言信息传播过程的异构图结构。MKN 方法从知识图谱中检索外部知识，但是在训练节点表达时过于依赖节点之间的属性，因此没有达到最佳效果。相反，本文提出的模型在各种情况下皆有稳定准确的效果，展示了其稳定性和鲁棒性。

3.4.2　消融实验结果及分析

为了验证引入知识图谱、图卷积结构对于谣言检测模型的有效性，本文针对 Weibo 数据集进行了消融实验。消融实验结果如表 5 所示。

表 5　针对 Weibo 数据集的消融实验结果

Tab. 5　Ablation results of Weibo

类型	*precision*	*recall*	F_1
base	0.828	0.809	0.818
KG	0.882	0.831	0.856
GCN	0.897	0.842	0.869
base + GCN	0.928	0.935	0.931
KG + UDGCN	0.949	0.939	0.944
KG + GCN	0.961	0.942	0.952

在表 5 中，base 模型表示使用简单的词向量文本表示方法并使用多头注意力机制提取文本特征；KG（knowledge graph）表示在知识图谱嵌入的基础上结合了多头注意力机制提取文本特征；GCN（graph convolutional network）表示有向传播图卷积网络；UDGCN（undirectional convolutional network）为无向传播图卷积网络。

从对比实验和消融实验的结果可以看出，本文使用的知识表示方法在词向量的基础之上引入知识图谱作为先验知识表示，对实体以及文本内容起到更好的表示作用，有利于提取更

精细的谣言文本特征，提升模型的泛化能力。

为了验证使用图卷积网络进行谣言检测的有效性，本文分别构建了无向和有向传播图来提取谣言传播的特征。谣言传播是一个有向的过程，因此采用有向传播图提取谣言特征能提升模型的效果进而提升谣言检测准确率。从表 5 可以得知，仅使用图卷积网络比 base 模型在准确率有着 7.7% 的提升，在 F_1 值有着高达 5.9% 的提升。这表明在谣言传播异构图中使用图卷积网络可以更好地进行特征融合，有利于获取特征表示，增强模型的谣言检测效果。

4 结束语

本文提出了准确率更高更有效的知识表示和图卷积网络谣言检测模型。引入知识图谱作为先验知识来获得更好的实体表示，同时融合图卷积网络以提取谣言传播图的特征，以提升谣言检测的精确率。在三种数据集上将本文方法与六种基线模型相比较，并在三个评测指标上均有明显提升，同时具有良好的鲁棒性。未来，笔者将致力于图像、音视频等多模态信息的谣言检测方法研究，从多模态的角度进行谣言检测研究可以更加增强模型的普适性和有效性，在社交网络、数据资产管理和舆情分析等领域将有广阔的前景。

参考文献:

[1] Wang W Y. Liar, liar pants on fire: a new benchmark dataset for fake news detection [C]//Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 422-426.

[2] Qazvinian V, Rosengren E, Radev D R, *et al*. Rumor has it: identifying misinformation in microblogs [C]//Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2011: 1589-1599.

[3] Cao Juan, Guo Junbo, Li Xirong, *et al*. Automatic rumor detection on microblogs: a survey [EB/OL]. (2018-07-10). https://arxiv.org/abs/1807.03505.

[4] 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 International Conference on Multimedia. New York: ACM Press, 2017: 795-816.

[5] Bao Peng, Shen Huawei, Jin Xiaolong, *et al*. Modeling and predicting popularity dynamics of microblogs using self-excited Hawkes processed [C]//Proc of International Conference on World Wide Web. New York: ACM Press, 2015: 9-10.

[6] Zhao Zhe, Resnick P, Mei Qiaozhu. Enquiring minds: early detection of rumors in social media from enquiry posts [C]//Proc of International Conference on World Wide Web. New York: ACM Press, 2015: 1395-1405.

[7] Yang Fan, Yu Xiaohui, Liu Yang, *et al*. Automatic detection of rumor on Sina Weibo [C]//Proc of ACM SIGKDD Workshop on Mining Data Semantics. New York: ACM Press, 2012: 1-7.

[8] Devlin J, Chang Mingwei, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proc of the 31st North American Conference on Chinese Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171-4186.

[9] Wang Jin, Wang Zhongyuan, Zhang Dawei, *et al*. Combining knowledge with deep convolutional neural networks for short text classification [C]//Proc of the 26th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 2915-2921.

[10] Annervaz K M, Somnath B R C, Ambedkar D. Learning beyond datasets: knowledge graph augmented neural networks for natural language processing [C]//Proc of the 30th North American Conference on Chinese Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 313-322.

[11] Heike A, Hinrich S. Global normalization of convolutional neural networks for joint entity and relation classification [C]//Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2017: 1723-1729.

[12] Bian Tian, Xiao Xi, Xu Tingyang, *et al*. Rumor detection on social media with bi-directional graph convolutional networks [C]//Proc of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 549-556.

[13] Liu Yang, Wu Y B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks [C]//Proc of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 354-361.

[14] Rong Yu, Huang Wenbing, Xu Tingyang, *et al*. DropEdge: towards deep graph convolutional networks on node classification [EB/OL]. (2020-03-12). https://arxiv.org/abs/1907.10903v4.

[15] Lu Yiju, Li Chengte. GCAN: graph-aware co-attention networks for explainable fake news detection on social media [C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 505-514.

[16] 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 International Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1751-1754.

[17] Liu Xiaomo, Nourbakhsh A, Li Quanzhi, *et al*. Real-time rumor debunking on Twitter [C]//Proc of the 24th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1867-1870.

[18] Ma Jing, Gao Wei, Mitra P, *et al*. Detecting rumors from microblogs with recurrent neural networks [C]//Proc of the 25th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 3818-3824.

[19] 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: Association for Computational Linguistics, 2019: 1173-1179.

[20] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter [C]//Proc of the 20th International Conference on World Wide Web. New York: ACM Press, 2011: 675-684.

[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: Association for Computational Linguistics, 2018: 1980-1989.

[22] 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 International Conference on Multimedia. New York: ACM Press, 2019: 1942-1951.

[23] Yuan Chunyuan, Ma Qianwen, Zhou Wei, *et al*. Jointly embedding the local and global relations of heterogeneous graph for rumor detection [C]//Proc of IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2019: 796-805.

[24] Huang Qi, Yu Junshuai, Wu Jia, *et al*. Heterogeneous graph attention networks for early detection of rumors on Twitter [C]//Proc of International Conference on World Wide Web. New York: ACM Press, 2019: 114-122.

[25] Wu Lianwei, Rao Yuan, Zhao Yongqiang, *et al*. DTCA: decision tree-based co-attention networks for explainable claim verification [C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 1024-1035.

[26] 米源，唐恒亮. 基于图卷积网络的谣言鉴别研究 [J]. 计算机工程与应用，2021，57(13)：161-167. (Mi Yuan, Tang Hengliang. Rumor identification research based on graph convolutional network [J]. Computer Engineering and Applications, 2021, 57(13): 161-167.)