Information Sciences 608 (2022) 1557–1571

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

GFCNet: Utilizing graph feature collection networks for coronavirus knowledge graph embeddings

Zhiwen Xie [a], Runjie Zhu [b], Jin Liu [a,*], Guangyou Zhou [c], Jimmy Xiangji Huang [d,*], Xiaohui Cui [e]

[a] School of Computer Science, Wuhan University, Wuhan 430072, China
[b] Lassonde School of Engineering, York University, Toronto, Canada
[c] School of Computer Science, Central China Normal University, Wuhan 430079, China
[d] School of Information Technology, York University, Toronto, Canada
[e] School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Article history:
Received 5 March 2021
Received in revised form 4 April 2022
Accepted 3 July 2022
Available online 14 July 2022

Keywords:
Natural Language Processing
Text Mining
Knowledge Graph
COVID-19

ABSTRACT

In response to fighting COVID-19 pandemic, researchers in machine learning and artificial intelligence have constructed some medical knowledge graphs (KG) based on existing COVID-19 datasets, however, these KGs contain a considerable amount of semantic relations which are incomplete or missing. In this paper, we focus on the task of knowledge graph embedding (KGE), which serves an important solution to infer the missing relations. In the past, there have been a collection of knowledge graph embedding models with different scoring functions to learn entity and relation embeddings published. However, these models share the same problems of rarely taking important features of KG like attribute features, other than relation triples, into account, while dealing with the heterogeneous, complex and incomplete COVID-19 medical data. To address the above issue, we propose a graph feature collection network (GFCNet) for COVID-19 KGE task, which considers both neighbor and attribute features in KGs. The extensive experiments conducted on the COVID-19 drug KG dataset show promising results and prove the effectiveness and efficiency of our proposed model. In addition, we also explain the future directions of deepening the study on COVID-19 KGE task.

1. Introduction

Pandemic has been a critical but unexpected factor threatening people's lives in human history. The outbreak of COVID-19 is an unprecedented global health crisis that not only shocks the healthcare industry but also may lead to subsequent serious economic and financial consequences. According to the real time statistics announced by Johns Hopkins Coronavirus Resource Center [1], as of April 1, 2022, there have been more than 488 million confirmed coronavirus cases worldwide, and the total death toll has exceeded 6.14 million people.

* Corresponding authors.
E-mail addresses: xiezhiwen@whu.edu.cn (Z. Xie), sherryzh@cse.yorku.ca (R. Zhu), jinliu@whu.edu.cn (J. Liu), gyzhou@mail.ccnu.edu.cn (G. Zhou), jhuang@yorku.ca (J.X. Huang), xcui@whu.edu.cn (X. Cui).

[1] Johns Hopkins University of Medicine. Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html

https://doi.org/10.1016/j.ins.2022.07.031

There have been a great amount of COVID-19 data accumulated over the past approximately nine months of time. This excites researchers and scholars from various background to devote into the fight against COVID-19 pandemic. For example, ACL 2020 opens a NLP COVID-19 Workshop which focuses on research themes of document analysis and retrieval across COVID-19 corpus, COVID-19 question answering for mental health, social media analysis of COVID-19 pandemic, etc. [46,11,3]. While many people are still processing and completing the COVID-19 medical data which purely stands for a collection of facts, we are forced to generate COVID-19 related knowledge out of these open sourced data collections by various advanced data mining or artificial intelligence (AI) approaches given the restricted time.

Knowledge Graph (KG), which serves as an interpretable and explainable base for the medical text mining methods, plays an important role in understanding the information structures and entity relations among the heterogeneous COVID-19 data. Currently, some researchers and scholars have constructed knowledge graphs based on the existing COVID-19 data collections. For instance, COKG-19[2] integrates many open knowledge graphs which allow future studies to stand on the shoulder of giant. Taking the example of the COVID-19 antiviral drug knowledge graph (DrugKG) in COKG-19, which uncovers links between drug, virus and protein, shown in Fig. 1. There are four types of entities, namely *Drug*, *Virus*, *HostProtein* and *VirusProtein*, which are represented in different colors. These entities are connected by four types of relations including *effect*, *produce*, *binding* and *interaction*, which indicate the relationships between entities. For example, the edge (*Vidarabin*, *effect*, *HHV-3*) indicates the drug *Vidarabin* has effect on the virus *HHV-3*. Other lately published COVID-19 KGs also include [63,47,40,51,7]. These knowledge graphs serve as significant fundamental blocks for COVID-19 downstream tasks, such as supporting information retrieval and extraction for COVID-19 pneumonia diagnosis, detection and treatment automatically.

Although there have been many domain specific KGs constructed, we find that the coverage of knowledge among these existing COVID-19 knowledge graphs is very limited. These knowledge graphs are new, sparse and scattered, containing a considerable amount of semantic relations which are incomplete or missing. For example, we took a deep look at the DrugKG in COVID-19 research knowledge graph, around 36.1% of the drug *effect* relations are missing, around 92.9% of the protein *binding* relations are missing, around 32.6% of the protein *interaction* relations are missing, and around 1% of the virus *produce* relations are incomplete. The missing nodes and the incompleteness of the KG strongly affect the accuracy of the analyses on COVID-19 datasets. What is even worse than that is if the analytical results generated from the above are applied to real life clinical decision support process, it might lead to hundreds or millions of life or death consequences.

Knowledge Graph Embedding (KGE), which refers to the representations of entities and relations at low-dimension in a KG, is a fundamental and significant task to infer nodes relations and to get insights about the existing vast COVID-19 medical knowledge graphs. In the past few years, researchers have been continuously proposing a number of KGE models with different scoring functions to learn entities and relations embeddings. These examples include but are not limited to TransE [9], TransD [23], TransH [50], TransR [30], DistMult [58], ComplEx [43], RotatE [42] etc. Although these models are able to deliver comparable and promising results in general domain datasets, they share the same problems of rarely taking important features of KG like attribute features other than relation triples into account, while dealing with the heterogeneous, complex and incomplete COVID-19 medical data.

In this paper, we aim to take a heterogeneous approach to automatically infer the missing semantic relations in the COVID-19 knowledge graph. To the best of our knowledge, we are the first one to perform KGE task on the open-sourced COVID-19 antiviral drug knowledge graph. Specifically, we propose a graph feature collection network (GFCNet) for KGE task on COVID-19 DrugKG. Our proposed model utilizes both neighboring and attribute information, which include the entity type and drug category, to enhance the entity representation. Different from the well defined R-GCN [39], we propose a simple and parameter-efficient R-GCN (SR-GCN) which uses GCN to fuse the neighboring and attribute information for better KGE.

The main contributions of this paper are listed as follows:

- We contribute to the global COVID-19 research community by investigating the newly built antiviral drug knowledge graph (DrugKG) while constructing a KGE task on top of it.
- We propose a graph feature collection network (GFCNet) which combines neighbor collector and attribute collector to tackle the KGE problem in DrugKG.
- We prove the effectiveness of our proposed model, and provide one of the baseline models for future study on the COVID-19 medical knowledge graphs.

The remainder of the paper is structured as follows: Section 2 provides the related work of knowledge graph embedding and the state-of-the-art approaches. Section 3 describes the framework of our proposed method. Section 4 presents the experimental details and we discuss the experiemtal results compared to a list of baseline models. In Section 5, we conclude with a summary and suggest several possible directions of future work.

[2] https://covid-19.aminer.cn/kg?lang = en

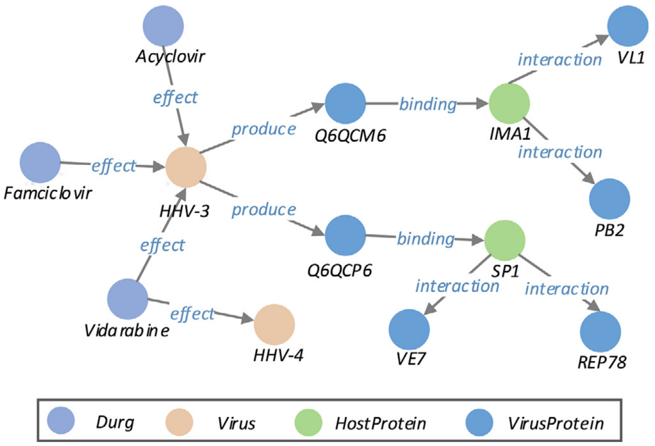


Fig. 1. An example of sub-graph in DrugKG. Different types of entities are represented in different colors: the purple nodes represent drug, the pink nodes represent virus, the green nodes represent host proteins, and the blue nodes represent virus protein.

2. Related Work

2.1. Knowledge Discovery for COVID-19

Medical knowledge is useful to improve the performance of downstream tasks [55]. Therefore, knowledge discovery gains extensive attention in recent years. Knowledge discovery refers to the process of extracting relevant information and finding useful knowledge from the large amount of structured or unstructured data. Natural language processing and text mining for knowledge discovery in medical domain are even more difficult tasks to work on by nature.

Therefore, most of the researchers dedicate to COVID-19 literature studies by building information retrieval tools and knowledge bases that could potentially ease future research and enhance explainability. Shen et al. [40] develop an end-to-end recommendation system of academic research papers to match them with potential use cases of COVID-19 studies. Wise et al. [51] construct the COVID-19 Knowledge Graph(CKG) to understand and present complex relations between COVID-19 scientific articles in the graph, with latent schema and enriched entity information generated by Amazon Web Services (AWS) technologies. Wang et al. [49] design the EVIDENCEMINER system to facilitate researchers and scholars needs for mining textual evidence from COVID-19 literature corpus. Soni et al. [41] conduct empirical studies on information retrieval results from two commercial search engines, Google and Amazon, for COVID-19, and compare them to the more academic prototypes evaluated by the TREC–COVID track [37]. Zhao et al. [66] combine the knowledge graph embedding and logic rules to refine the COVID-19 knowledge graph.

Other COVID-19 knowledge discovery research comprises of diagnosis and detection studies. Most of the existing studies utilize CT scans and X-rays to practise COVID-19 positive cases diagnoses and detections. For example, COVID-Net [31] is an open-sourced COVID-19 cases diagnosis and detection platform for chest X-ray images, and [14] develops an automated chest CT scan analysis tool for COVID-19 diagnosis and detection with 2D and 3D CNN algorithms. Other related work include but are not limited to [48,33,1]. However, these work mainly focus on understanding and analysis of X-ray and CT images, which is out of the research scope of this specific study.

2.2. KG Embedding

With the development of big data, the data sets with graph structure are ubiquitous, i.e., ranging from social networks to the World Wide Web and knowledge graphs (KGs) [61,62]. Among these graphs, KGs are heterogenous networks which carry rather richer information and semantic meanings. Learning the embedding for KGs is useful to capture the rich information hidden in the KGs. KGE aims to embed entities and relations of KGs into continuous vector spaces. The purpose of embedding is to inherit the original structure of knowledge graph entities, and to simplify the process of manipulating them in future. As the number and scale of KGs grow rapidly, especially in the medical domain, KGE becomes more important in tasks of KG analyses and semantic data modeling.

Translation-based models are considered to be one of the major approaches to the KGE problem. Bordes et al. proposed the most representative translational distance method named TransE [9]. Later on, various improvements have been made to boost the model performance accuracy [23,50,30]. However, the shallow structures of these models restrict the expressiveness of this approach.

Bilinear or Semantic-based models are another major approach to the KGE problems. This group of models use matrix decomposition to learn KGE. Specifically, they match the entities and relations' latent semantics that are contained in the vector space representations and computes scoring functions based on similarities. [36,43,25,5,58] are examples that fall into the semantic-based models category. Similar to the translation-based models, the expressiveness of bilinear models are also limited unless the embedding size is increased. This could potentially lead to a considerable increase in parameters and a fundamental confine on scalability.

The third major category of KGE tasks are rotate-based models. Representative models of the rotate-based approaches comprise examples such as RotatE [42] and QuatE [65]. RotatE [42] is named after its concept of projecting each relation of source entity to target entity with a rotation in a complex vector space. In essence, it is a translational model specializes in modeling and inferring various relations, such as symmetric, inverse or composition information between nodes. Whereas for QuatE [65], unlike RotatE that involves only rotation at one single plane, it involves geometric rotation at two planes and in essence is a semantic-based matching model. The QuatE utilizes quaternion representations and relational rotational quaternions to exercise semantic matching between entities of heads and tails.

Aside from the three major categories listed, other KGE approaches also include adopting multi-layer CNN-based, GNN-based models and etc. ConvE [12], ConvKB [35], RSN [15] and R-GCN [39] are all representative examples of this CNN-based approach. For example, ConvE [12] uses multi-layer convolutional neural network (CNN) to do link prediction and to capture more expressive features. Other research studies such as InteractE [44] and ReInceptionE [56] also follow this direction to learn more expressive features. CNN-based models deliver promising results but lack of modelling the structural information. GNN-based models, such as R-GCN [39], KBGAT [34] and BiGAT [57], leverage graph convolutional network (GCN) to aggregate neighborhood features. ManifoldE [53] and MAKR [16] embed the entities and relations using manifold-based embedding. Xiao et al. [54] proposed a CNN-based model to integrate text features and structural features. These models generally show good performance on predicting link task and classifying entities task.

In general, all the papers mentioned above focus relatively more on models instead of specific datasets. Although these approaches achieve promising results in their settings, they only work generally good for the KGs with more neighbors and more complete nodes and relations. Given our COVID-19 KG dataset, where the data points are in scarcity and missing links are one of the major issues, these models are not able to perform well. Meanwhile, the existing models also lack of considering the structure and attributes of the KG, which is not suitable for our COVID-19 dataset either.

Beyond the application of COVID-19, our approach can also be applied to a wide range of down stream tasks, such as drug discovery [2,8,64], contact tracing [60,17], biomedical and genomics information retrieval [21,22,59] and detection of coronavirus-themed mobile malware [20]. In this paper, we focus on the link prediction task on COVID-19 KGs and we will leave the application on these tasks in the future work.

2.3. Differences from Existing Methods

Among the above mentioned various models, the major contribution is that we propose a graph feature collection network (GFCNet) for KGE on COVID-19 DrugKG, aiming at effectively capturing both the neighboring and attribute information. In fact, we note some existing models (e.g., R-GCNs [39], KR-EAR [29], MARINE [13]) also use the neighboring and attribute information for KGE. Here we highlight the novelty and differences in the following two ways: (1) R-GCN need to learn different parameter matrices for each relation, resulting the number of parameters increasing. Different from R-GCN, we propose a simple and parameter-efficient R-GCN (SR-GCN). The proposed SR-GCN uses a simple relation-specific diagonal matrix for each relation instead of a full matrix, which is parameter efficient. (2) KR-EAR [29] and MARINE [13] use attribute information to enrich the representation of entities, while they are not able to capture neighborhood information. We propose a graph feature collection network (GFCNet) which combines neighbor collector and attribute collector to tackle the KGE problem in DrugKG. The neighbor collector is used to aggregate neighborhood features and the attribute collector is used to learn attribute information. We ensemble these two kind of features to obtained the final entity representations.

3. Our Approach

In this section, we present our graph feature collection network (GFCNet) for KGE task in details. We first describe the notations and problem definition used in our paper. Then we elaborate on the model architecture, following by the loss function and training. Lastly, we give a complexity analysis of our model.

3.1. Notations and Problem Definition

In this subsection, we first present some KG and KGE mathematical notations that would be applied in the rest of the study. For the ease of descriptions, this study uses lower case letters to represent vectors, upper case letters to represent matrices, and italic lower case letters to represent subscript notions, such as i, j, r. The notions and definitions used in the following sections are shown in Table 1.

Typically, a KG is denoted as $\triangle$, which stands for a collection of triples in the form $(h, r, t), h \in \mathcal{E}, t \in \mathcal{E}$, and $r \in \mathcal{R}$, where $\mathcal{E}$ is the vocabulary collection of the entity and $\mathcal{R}$ is the set of pre-defined entity relations. KGE aims to embed knowledge which inherits certain properties into low-dimensional continuous vector space. In the vector space, each node (entity) is presented as a point, and each relation between the nodes is presented as an operation on the embeddings of entities.

Table 1
The Notation and Definition of KG and KGE.

Name	Notation	Definition
Knowledge Graph	$\triangle$	A set of triplets in the form (h, r, t)
Entity Collection	$\mathcal{E}$	Vocabulary collection of the entity
Entity Relations	$\mathcal{R}$	The set of pre-defined entity relations
Entity Attributes	$\mathcal{A}_{e_i}$	A set of attributes for entity e_i
Entity Embedding	$\mathbf{E}^l$	The entity embedding in the l-th layer
Vector Dimension	d	The dimension of entity, attribute and relation vectors
Activation Function	σ	The activation function
Entity Vectors	$\mathbf{h}, \mathbf{t}$	The entity vectors for entity h and t
Relation Vector	$\mathbf{r}$	The vector for relation r
Attribute Vector	$\mathbf{a}_k$	The vector for the attribute $a_k \in \mathcal{A}_{e_i}$
Hidden Neighbor Vector	$\mathbf{e}_{ih}^{l+1}$	The hidden feature vector for entity e_i obtained by Neighbor Collector
Hidden Attribute Vector	$\mathbf{e}_{ia}^{l+1}$	The hidden feature vector for entity e_i obtained by Attribute Collector
Scoring Function	$f(h, r, t)$	The scoring function for the triple (h, r, t)
Loss Function	$\mathcal{L}$	The loss function of the model

3.2. Model Architecture

In this study, in response to the COVID-19 pandemic, we investigate a new antiviral drug knowledge graph (DrugKG) and construct a knowledge graph embedding (KGE) task on the DrugKG. In the DrugKG, some useful features are available for the KGE task, including relation triples (e.g., KG structure features) and attribute triples (e.g., entity attribute features). Most previous study only use relation triples to learn knowledge graph embeddings for a KG [9,58,43]. Some methods based on graph neural network (GNN) try to leverage more information from neighborhood features [39,34,6]. However, existing methods rarely consider some other important features in KG, such as attribute features. In this paper, we propose a graph feature collection network (GFCNet) for knowledge graph embedding in DrugKG.

As shown in Fig. 2, the proposed GFCNet consists of three components, namely neighbor collector, attribute collector and feature ensemble. The neighbor collector is used to learn local neighborhood features around the given entity. The attribute collector is designed to make full use of the attributes for each entity. Given an entity and its neighbors and attributes, we firstly represent them as vectors through embedding layers (e.g., entity ebmedding layer, attribute embedding layer and relation embedding layer). Then, we apply neighbor collector and attribute collector to gather both neighborhood and attribute features. Finally, these two kind of features are integrated together by using a feature ensemble module.

3.2.1. Neighbor Collector

Fig. 2 gives an example of the neighbor collector. The drug entity *Vidarabine* will aggregate information from its neighbors *HHV-3* and *HHV-4*. Recently, some graph neural networks (GNNs) have been successful applied to model graph data, such as GCN [27]. However, traditional GNNs are designed to perform on unlabeled graphs which has no label in the edge. Different with unlabeled graph, KGs are graphs with different relations, which makes traditional GNNs not suitable to model KG data. To address this issue, R-GCN [39] uses relation-specific transformation matrices to aggregate neighbors linking by different relations, which is formulated as:

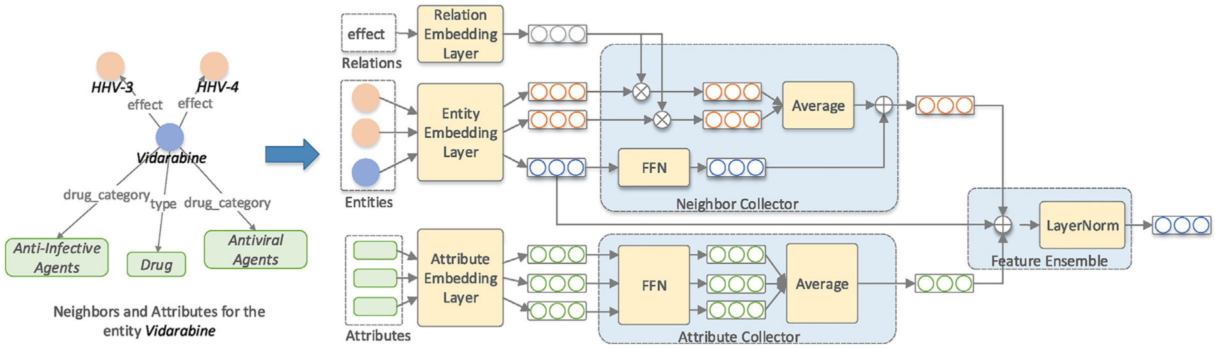


Fig. 2. The structure of the proposed graph feature collection network (GFCNet). The proposed model consists of three components: neighbor collector, attribute collector and feature ensemble. The embedding layers are used to convert the relations, entities and attributes to low-dimensional vectors. The FFN denotes a feed-forward network.

$$\mathbf{e}_i^{l+1} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}\mathbf{W}_r^l\mathbf{e}_j^l + \mathbf{W}_0^l\mathbf{e}_i^l\right) \tag{1}$$

where $\mathcal{N}_i^r$ denotes a set of indices for the neighbors of entity e_i under relation $r, c_{i,r} = |\mathcal{N}_i^r|$ is a normalization factor, $\mathbf{W}_r^l \in R^{d\times d}$ is the relation-specific transformation parameter matrices for relation r, σ is an activation function, $\mathbf{W}_0^l$ is the parameter matrices of a one layer feed-forward network (FFN).

R-GCN is able to model relational graph data and has proved that explicitly model neighborhood information is important for KGE task. However, R-GCN need to learn different parameter matrices for each relation, resulting the number of parameters increasing.

To make full use of the neighborhood information meanwhile reduce the number of parameters, we use a simple and parameter-efficient R-GCN (SR-GCN) to aggregate neighboring nodes for each entity. Specifically, instead of using different relation-specific transformation parameter matrices which need $O(|\mathcal{R}| \times d \times d)$ parameters for all relations (where d is the dimension of the embedding), we use a simple relation-specific diagonal matrix as transformation matrix, which is inspired by [58]. Formally, in our SR-GCN, the entity embeddings can be updated as:

$$\mathbf{e}_{ih}^{l+1} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}\mathrm{diag}(\mathbf{r}^l)\mathbf{e}_j^l + \mathbf{W}_0^l\mathbf{e}_i^l\right) \tag{2}$$

where $\mathbf{r}^l \in R^d$ is the relation-specific vector for relation r, and $\mathrm{diag}(\mathbf{r}^l)$ is a diagonal matrix for $\mathbf{r}^l$.

By applying SR-GCN, we can efficiently model the KGs with different relations using $O(|\mathcal{R}| \times d)$ transformation parameters for all the relations, which is much smaller than R-GCN.

3.2.2. Attribute Collector

In the DrugKG, each entity has some attributes which are important for learning the representation for the entities. For example, the drug entity *Vidarabine* has some attribute triples (*Vidarabine*, *type*, *Drug*), (*Vidarabine*, *drug_category*, *Anti-infective Agents*), (*Vidarabine*, *drug_category*, *Antiviral Agents*) and etc. These attributes contain some useful information of the entity, e.g., the type of the entity is *Drug*, and the drug is used as anti-infective and antiviral treatment. We believe these attribute information can be utilized to enrich the entity embedding and improve the performance of KGE task. However, previous studies for KGE rarely consider attribute information especially on the medical knowledge graph. To this end, we design an attribute collector module to take full advantage of the attribute information.

As shown in Fig. 2, we first represent each attribute value as a vector by using the attribute embedding layer. Thus, we can obtain some attribute features for the centre entity e_i, which can be denoted as $\{\mathbf{a}_k | a_k \in \mathcal{A}_{e_i}\}$. Then, the attribute information is obtained by aggregating these attribute embeddings:

$$\mathbf{e}_{ia}^{l+1} = \frac{1}{|\mathcal{A}_{e_i}|}\sum_{a_k\in\mathcal{A}_{\rceil_\rangle}}\mathbf{W}_a^l\mathbf{a}_k \tag{3}$$

where $\mathbf{W}_a^l$ is the parameters of a FFN layer for attribute features in the l-th layer, $|\mathcal{A}_{e_i}|$ is the number of attributes for e_i.

3.2.3. Feature Ensemble

The neighborhood and attribute features are obtained by the neighbor collector and attribute collector. In our study, we apply a sum operation to combine these two features. To make the model training more stable, we also use a residual connection [19] followed by a layer normalization [4]. The output entity embedding for entity e_i is computed as:

$$\mathbf{e}_i^{l+1} = \mathrm{LayerNorm}(\mathbf{e}_{ih}^{l+1} + \mathbf{e}_{ia}^{l+1} + \mathbf{e}_i^l) \tag{4}$$

where LayerNorm denotes the layer normalization [4].

We can stack L layers to propagate the neighborhood and attribute features as defined in Eq. 2, Eq. 3 and Eq. 4. Thus we can obtain the final entity embedding $\mathbf{E}^L = \{\mathbf{e}_i^L | e_i \in \mathcal{E}\}$ which consists of all the entity embeddings.

3.3. Loss Function and Training

In order to optimize the parameters in our GFCNet, we propose to use a simple and efficient function to compute the score for a triple [58]. Different with the existing scoring function, we compute the score based on the entity embedding $\mathbf{E}^L$ obtained by our model, which is formulated as:

$$f(h, r, t) = \mathbf{h}^T\mathrm{diag}(\mathbf{r})\mathbf{t} \tag{5}$$

where $\mathbf{h}, \mathbf{t} \in \mathbf{E}^L$ are the entity embeddings for head and tail entity, $\mathbf{r}$ is the embedding for relation r.

The objective function is a very important factor for KGE task which has a great affect on the performance [24,32,38]. In this paper, we use the efficient cross-entropy loss function over the distribution of all the entities to optimize our model. For a given triple (h, r, t), we compute the probability of the tail entity as:

$$P(t|h, r) = \frac{\exp(f(h, r, t))}{\sum\limits_{t^{-}\in\mathcal{E}} \exp f(h, r, t^{-})} \tag{6}$$

where $t^{-} \in \mathcal{E}$ is a candidate tail entity for the triple. Then, the loss function is defined as:

$$\mathcal{L} = -\log(P(t|h, r)) \tag{7}$$

Note that we use reciprocal relations [25,28] in our model, which introduce an inverse triple (t, r^{-1}, h) for (h, r, t). Thus, when predicting the head entity for the triple (h, r, t), we can predict the probability of its inverse form $P(h|t, r^{-1})$. Table 1.

3.4. Complexity Analysis

In addition to performance, efficiency is also important for the KGE task, especially for the COVID-19 KGs whose scale grow rapidly. Hence, it is essential to develop efficient model to address the KGE problem. Table 2 shows the parameter efficiency of our model and some strong baseline models. Compared to the related GNN-based model (e.g., R-GCN) with the same layers and embedding size, our model use fewer parameters since we use a diagonal matrix as relation-specific transformation parameters. Note that the entity and relation embeddings contributes a lot to the parameters since they are depended on the size of KGs which are always very large in practice. Therefore, our model is more parameter efficient than the models with double or more entity and relation embeddings, such as RotatE and QuatE. Specifically, when $d = 100, L = 1, |\mathcal{E}| = 7817, |\mathcal{R}| = 4$, the number of parameters of RotatE is $1.56M$, but the parameters of our model is only $0.802M$, which is much fewer than RotatE. Thus, our model GFCNet is efficient and can be easily to adapted to many real world applications.

4. Experiments

In this section, our proposed model is evaluated on the DrugKG dataset and compared with other existing KGE models, including five categories of popular KGE models. Then, we further implement some variants of the GFCNet by removing different components and conduct ablation study by comparing our GFCNet with these model variants. The ablation study proves that each components in our GFCNet plays an important role in the task of KGE.

4.1. Datasets

In response to the COVID-19 pandemic, we conduct experiments on a new antiviral drug knowledge graph (DrugKG) released in COVID-19 Research KG[3]. The DrugKG is constructed based on the relationship between antiviral drugs, viruses, virus-related proteins, the host and host proteins in the DrugBank [52] [4] database. As shown in Fig. 1, there are four kind of entities in DrugKG, namely *Drug*, *Virus*, *VirusProtein* and *HostProtein*, and four types of relations between them, namely *effect*, *produce*, *binding* and *interaction*. The relation *effect* is a relation between entity *Drug* and *Virus*, which indicates that the antiviral drugs have a certain effect on the virus. The relation *produce* is a relation between entity *Virus* and *VirusProtein*, which is used to express the relationship between the virus and the protein it expresses. The relations *binding* and *interaction* are the interaction relationships between entity *VirusProtein* and *HostProtein*. In DrugKG, each entity also has some important attributes, such as *type* and *drug_category*. Following previous study for KGE [9], we split the triples in DrugKG into training, validation, and testing sets. The statistic of the datasets are summarized in Table 3.

4.2. Evaluation

For the KGE model, a common evaluation method is to build a link prediction task which aims to predict missing triples in the KG, namely, predict missing tail entity t given $(h, r, ?)$ or predict missing head entity h given $(?, r, t)$. Specifically, we rank all the entities in KG to predict the most probable entities. To evaluate the performance of the model, we use three popular metrics: MRR (mean reciprocal rank), MR (mean rank) and the Hits@N (the correct percentage in the top N ranks, where $N = 1, 3, 10$).

However, directly ranking all the entities can mislead the computation of the metrics for testing set when some corrupted triples are correct ones, such as the triples existed in training set [9]. In this situation, the correct triples in training set may rank above the current test triple (h, r, t). Thus, even if the testing triple (h, r, t) is correct, it may rank behind other correct triples, making the MRR, MR and Hits@N metrics not exact. To avoid this problem, we evaluate the model using a filtered

[3] http://www.openkg.cn/dataset/covid-19-research

[4] https://go.drugbank.com/

Table 2
Parameter efficiency of different models.

Model	Parameter efficiency
TransE	$\lvert\mathcal{E}\rvert d + \lvert\mathcal{R}\rvert d$
DistMult	$\lvert\mathcal{E}\rvert d + \lvert\mathcal{R}\rvert d$
Rescal	$\lvert\mathcal{E}\rvert d + \lvert\mathcal{R}\rvert d^2$
RotatE	$2\lvert\mathcal{E}\rvert d + \lvert\mathcal{R}\rvert d$
QuatE	$4\lvert\mathcal{E}\rvert d + 4\lvert\mathcal{R}\rvert d$
TuckER	$\lvert\mathcal{E}\rvert d + \lvert\mathcal{R}\rvert d + d^3$
R-GCN	$\lvert\mathcal{E}\rvert d + \lvert\mathcal{R}\rvert d + L\lvert\mathcal{R}\rvert d^2 + Ld^2$
GFCNet	$\lvert\mathcal{E}\rvert d + \lvert\mathcal{R}\rvert d + L\lvert\mathcal{R}\rvert d + 2Ld^2$

Table 3
The statistic of the DrugKG dataset.

Relations	Training	Validation	Testing
effect	47	0	6
produce	709	28	84
binding	8790	458	1028
interaction	13999	723	1636
Total	23545	1209	2754

setting [9,12], namely filter out all the correctly triples occurred in training, validation and testing sets but current triple (h,r,t) itself. Specifically, let $rank_h$ denotes the filtered rank of head entity h and $rank_t$ denotes the rank of tail entity t, then the MR is computed as: $\mathrm{MR} = \frac{1}{2|\mathcal{T}|}\sum_{(h,r,t)\in\mathcal{T}} rank_h + rank_t$; the MRR is computed as: $\mathrm{MRR} = \frac{1}{2|\mathcal{T}|}\sum_{(h,r,t)\in\mathcal{T}} \frac{1}{rank_h} + \frac{1}{rank_t}$; and the Hits@$N$ is computed as Hits@$N = \frac{1}{2|\mathcal{T}|}\sum_{(h,r,t)\in\mathcal{T}} I[rank_h \leqslant N] + I[rank_t \leqslant N]$, where $I[*]$ is 1 if the condition is true, and 0 oterwise.

4.3. Experimental Setups

All the experiments are conducted on a Linux server with 128G memory and RTX 2080Ti GPUs. We implement our model using pytorch [5], which is a popular deep learning framework.

In our study, we choose the Adam [26] as the optimizer to train our model. In order to select the hyper-parameters for our model, we search the hyper-parameters using grid search method and select hyper-parameters according to the Hits@10 on validation set. We select the embedding size of KG embeddings from $\{50, 100, 200\}$, the weight decay from $\{1e-4, 1e-5, 1e-6\}$, the batch size from $\{128, 256\}$, the dropout rate from $\{0.2, 0.3, 0.4, 0.5\}$, the learning rate from $\{0.002, 0.001, 0.0005, 0.0002\}$, the hidden layer number L from $\{1, 2\}$. We finally set the embedding size to 100, the weight decay to $1e-5$, the batch size is 128, the dropout rate to 0.5, and the learning rate to 0.001, the hidden layer number L to. The activation function σ is Relu. We initialize the parameters of our model using Kaiming initialization [18] which is a robust initialization method designed for the rectifier nonlinearities.

Most of the KGE models use negative sampling method to sample some negative triples to train the model [9,38], which randomly construct some negative triples by perturbing the head or tail entity in the correct triple (h,r,t). However, the performance of these models always depend on the quantity and quality of the negative samples. Anther training method termed as 1vsAll [28,38] is to take all possible triples by replacing head or tail entities with all other entities. In our study, we apply the 1vsAll method to train our model due to its simplicity and efficiency for training the model [6].

4.4. Baselines of Comparison

To investigate the KGE performance of our model, we compare with some popular baselines, which can be divided into five categories: translation-based models, bilinear models, roate-based models, CNN-based models and GNN-based models.

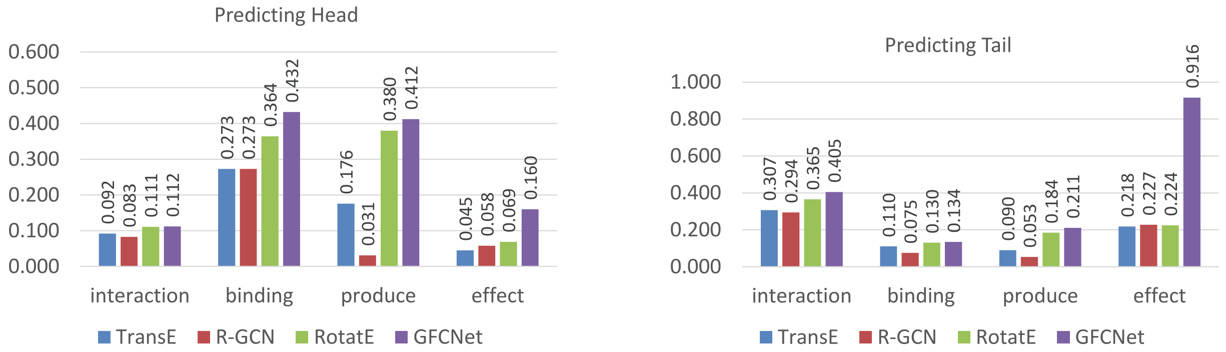
Translation-based models view the relation r as a translation operation from head entity h to tail entity t, which assume that the head entity vector plus the relation vector should be close to the tail entity vector. Translation-based methods include TransE [9], TransD [23], TransH [50], TransR [30] and KR-EAR [29]. Bilinear models, also called semantic matching models, use matrix decomposition to learn knowledge graph embeddings, including Rescal [36], DistMult [58], ComplEx [43], SimplE [25], TuckER [5] and MARINE [13]. Rotate-based models view the relation as rotation operation in complex

[5] https://pytorch.org/

[6] In order to facilitate the study of KGE for COVID-19, and to make our work reproduceable, We will release the data and source code after the paper is accepted.

Table 4
Link prediction results on DrugKG. The best results are in **bold** and the second best results are in underline.

Models		MRR	MR	Hits@10	Hits@3	Hits@1
Translation-based models	TransE [9]	0.196	734.81	0.367	0.226	0.108
	TransD [23]	0.147	750.71	0.332	0.181	0.052
	TransH [50]	0.153	765.69	0.330	0.188	0.061
	TransR [30]	0.130	792.03	0.282	0.147	0.056
	KR-EAR [29]	0.184	768	0.346	0.205	0.102
Bilinear models	Rescal [36]	0.104	880.45	0.202	0.103	0.055
	DistMult [58]	0.169	796.35	0.302	0.180	0.104
	ComplEx [43]	0.171	1004.97	0.313	0.184	0.104
	SimplE [25]	0.172	788.11	0.308	0.179	0.106
	TuckER [5]	0.224	1242.00	0.368	0.249	0.150
	MARINE [13]	0.177	1126	0.338	0.186	0.115
Rotate-based models	RotatE [42]	0.243	820.40	0.408	0.273	0.160
	QuatE [65]	0.198	777.81	0.351	0.220	0.123
CNN-based models	ConvE [12]	0.193	970.09	0.331	0.214	0.123
	ConvKB [35]	0.069	816.49	0.205	0.090	0.000
GNN-based models	R-GCN [39]	0.181	1341.61	0.294	0.196	0.124
	KBGAT [34]	0.092	761.00	0.198	0.090	0.040
	GFCNet (ours)	**0.270**	**630.12**	**0.432**	**0.299**	**0.188**



(a) Predicting Head

(b) Predicting Tail

Fig. 3. MRR results on different relations.

space, including RotatE [42] and QuatE [65]. CNN-based models, such as ConvE [12] and ConvKB [35], use CNNs to capture more expressive features. GNN-based models, such as R-GCN [39] and KBGAT [34], apply graph convolutional network (GCN) [27] and graph attention network (GAT) [45] to gather neighborhood features.

The results of these baseline models are obtained by downloading and running the available source codes [7,8,9,10,11] on the DrugKG datasets.

4.5. Experimental Results

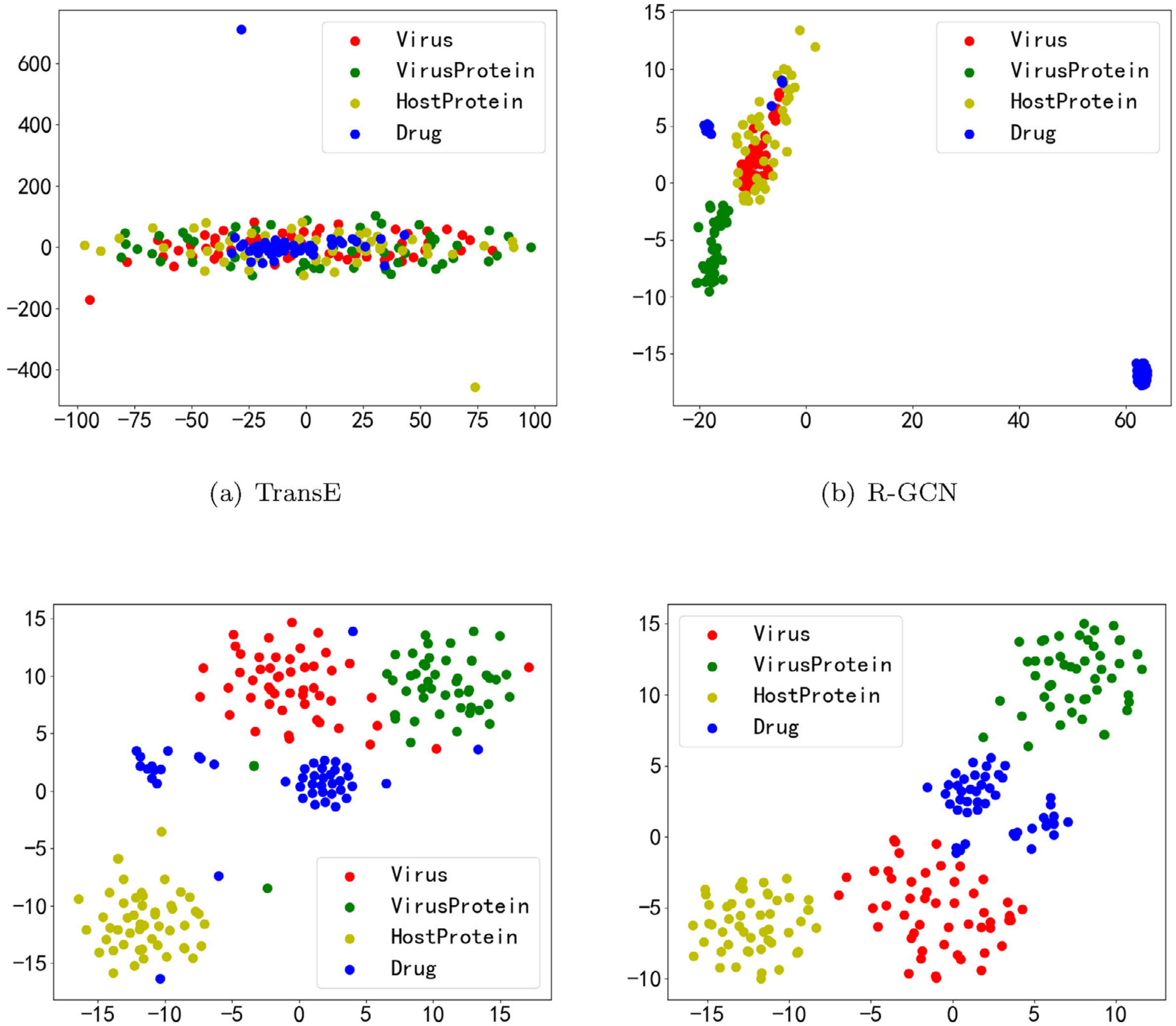
Table 4 shows the link prediction results of some strong baselines and our model on DrugKG. Among the translation-based models, the baseline model TransE [9] achieves good performance on the DrugKG which surpasses various of its extension models, such as TransD [23], TransH [50], TransR [30]. And TransE also outperforms some bilinear models, such as [58], ComplEx [43] and SimplE [25]. This indicates that the simple TransE model is more suitable to model the DrugKG than other complex models. Among all the bilinear models, TuckER [5] performer much better than other bilinear models, which also outperforms the strong baseline TransE model. The rotate-based model RotatE achieves the second best results on MRR, Hits@10, Hits@3 and Hits@1.

[7] https://github.com/thunlp/OpenKE
[8] https://github.com/cheungdaven/QuatE
[9] https://github.com/TimDettmers/ConvE
[10] https://github.com/ntumslab/MARINE
[11] https://github.com/thunlp/KR-EAR

Compared to the baseline models, the proposed GFCNet achieves the best performance across all the metrics. Our model outperforms the strong baselines such as TransE [9], RotatE [42] and TuckER [5]. KR-EAR [29] and MARINE [13] are two representative models which incorporate attribute features for KGE. The empirical results also show that our GFCNet is still outperforms than KR-EAR and MARINE. The advantages behind these comparison are that our model is able to aggregate heterogenous graph features including neighborhood features and attribute features, which are essential for learning good representations for the entities in knowledge graph. These experimental results show the effectiveness of our GFCNet.

Table 5
Ablation Study.

Model	MRR	MR	Hits@10	Hits@3	Hits@1
GFCNet w/o A	0.265	674.51	0.426	0.291	0.185
GFCNet w/o N	0.264	**508.39**	0.424	0.292	0.184
GFCNet w/o A&N	0.253	1052.89	0.402	0.283	0.175
GFCNet-RGCN	0.267	654.34	0.429	0.294	0.187
GFCNet	**0.270**	630.12	**0.432**	**0.299**	**0.188**



(a) TransE

(b) R-GCN

(c) RotatE

(d) GFCNet

Fig. 4. The visualization of the two-dimensional PCA projection of the entity embeddings for different models.

4.6. Results on Different Relations

Fig. 3 illustrates the detail MRR results on different relations of the DrugKG. Our GFCNet achieves the best results on all the four relations. When predicting head entity h given $(?, r, t)$, we obtain high performance gains over the state-of-the-art RotatE model especially on relations *binding*, *produce* and *effect*. When predicting tail entity t given $(h, r, ?)$, our model also outperforms RotatE by large margins (e.g., 0.04 MRR on relation *interaction*, 0.692 MRR on relation *effect*). Compared with previous state-of-the-art models, our model is able to capture both neigbhorhood and attribute features in DrugKG, which enable our model to learn better embeddings for entities. The experimental results demonstrate that our GFCNet can achieve good performance on different relations.

4.7. Ablation Study

Recently, some studies have shown that the training strategies may have a great impact on the performance, making it difficult to analyse whether the performance gains are obtained from a model architecture. Therefore, an ablation study is required to performed to evaluate the effect of different parts of the model. In this section, we conduct ablation experiments under the same experimental setting to explore the importance of different components in our model. The results of the ablation study are shown in Table 5. "GFCNet w/o A" denotes the model without using attribute collector, "GFCNet w/o N" denotes the model without using neighbor collector. "GFCNet w/o A&N" denotes the model without using the neighbor and attribute collectors. "GFCNet-RGCN" denotes the model obtained by replacing the SR-GCN with the traditional R-GCN.

From Table 5, we can see that the GFCNet with both neighbor collector and attribute collector achieves the best performance over all metrics except MR. Without using the attribute collector or neighbor collector, the models perform worse than GFCNet, which indicates both the neighborhood information and the attribute information contribute to the performance of the model. Compared the SR-GCN with the traditional R-GCN (e.g., GFCNet vs. GFCNet-RGCN), the proposed GFCNet also performs better even if the SR-GCN using fewer parameters, which verify the effectiveness of the SR-GCN.

4.8. Analysis

4.8.1. Visualization

In order to analyse the representation ability of our model, we show the two-dimensional PCA projection of the entity embeddings for different models, which is shown in Fig. 4. From Fig. 4 we can see that different types of entity embeddings learned by the baseline TransE are confused in the two-dimensional vector space, which indicates that TransE can not well model different types of entities. The R-GCN model is able to learn similar representations for different types of entities. But it still can not differentiate between *Virus* and *HostProtein* entities, because both of them have relationships with *VirusProtein*. Compare with these two baselines, the state-of-the-art model RotatE and our proposed GFCNet are able to distinguish all these four types of entities in DrugKG, which can learn similar embeddings for the entities with the same type and learn different embeddings for the entities with different types. This is mainly because that different types of entities have different attributes. By aggregate attribute features, our model can easily capture the difference between different types of entities.

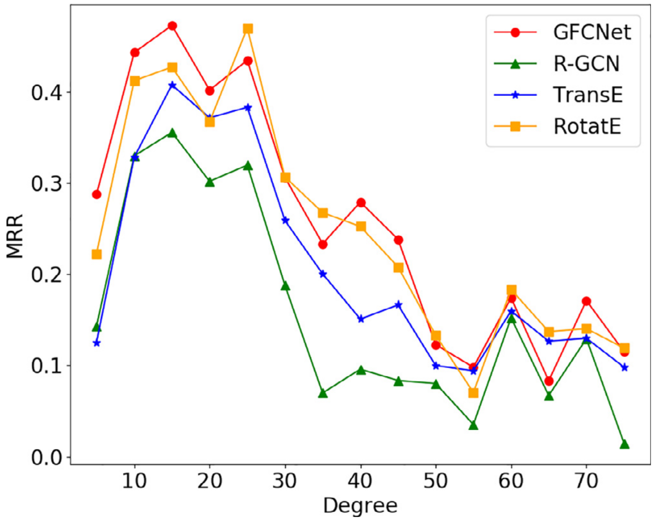


Fig. 5. The MRR results for entities with different degree.

Table 6
Some example Predictions on test set using our model. **Bold** indicates the true tails in the DrugKG.

Input:$(h, r, ?)$	Predicted Tails
(*Famciclovir*, *effect*,?)	**HHV-3**, **HSV-1**, *Influenza A virus*, *HBV genotype C*, *HHV-5*, *HIV-1*
(*Vidarabine*, *effect*,?)	**HSV-1**, **HHV-4**, *HSV-2*, **HHV-3**, *Walleye dermal sarcoma virus*, *HHV-5*
(*Ganciclovir*, *effect*,?)	*HIV-1*, **HSV-1**, *HHV-5*, *Vaccinia virus WR*, *Human gammaherpesvirus 4*
(*PUF60*, *interaction*,?)	**NCAP**, **VE2**, **NRAM**, **POLG**, *M1*, *REP78*
(*Influenza A virus*, *produce*,?)	**NS1**, **NEP**, *Q30NB4*, *E3*, *POLG*, *NEF*

4.8.2. Performance for Entities with Different Degree

In this subsection, we further analyse how the degree of entities has an affect on the performance. As shown in Fig. 5, we compare our model with three baseline models TransE, RotatE and R-GCN. With the degree increasing, the MRR results first improve rapidly, while after a threshold, the performance drops a lot. All the four models satisfy this phenomenon simultaneously, including the graph-based method (e.g., R-GCN and GFCNet) and the other models (e.g., TransE and RotatE). Previous studies [10] observe similar phenomenon based on experimental results by graph-based methods without comparing with methods without aggregating neighbors, and they think this phenomenon is because that too few neighbors can support limited neigbhor information and too many neigbhors will make the model hard to optimize. However, our experiment on DrugKG shows that the models without gathering neighborhood information also follow the same tendency. Therefore, we think the main reason for this phenomenon is that: (1) the entities with low degree have only a few triples to train, which make it difficult to learn better representations; (2) the entities with too many neighbors always suffers from the many-to-many relation pattern, which is difficult for all the KGE models, including graph-based models, rotate-based models and translation-based models.

4.8.3. Case Study of the Link Prediction

Table 6 gives some examples predicted by our GFCNet model on the testing set. Given a head entity *h* and relation *r*, we predict possible tail entities by ranking all the entities in DrugKG and show some results ranking at the top. In these examples, the correct tail entities in the DrugKG rank in the top results. For example, given the query (*Famciclovir*, *effect*,?), the correct tail entity *HHV-3* ranks in the first position and another correct tail entity *HSV-1* ranks in the second position. For other examples, such as (*Vidarabine*, *effect*,?), (*PUF60*, *interaction*,?) and (*Influenza A virus*, *produce*,?), we can also obtain similar observation. These examples show the good performance of our model in an intuitive way.

5. Conclusion and Future Work

In this paper, we dive into the COVID-19 KGE task and take a heterogeneous approach to automatically infer the missing semantic relations in the COVID-19 knowledge graph. To the best of our knowledge, we are the first one to perform KGE task on the open-sourced COVID-19 antiviral drug knowledge graph (DrugKG).

In order to tackle the problems of existing models which rarely take important features of KG like neighboring and attribute features, other than relation triples, into account, we propose a novel graph feature collection network (GFCNet) that utilizes different attribute and neighbor information to enhance the entity representation, in a simpler and more parameter-efficient way than R-GCN. The extensive experiments carried out on the DrugKG prove the effectiveness of our proposed model, however, there are still a lot of future work need to be continued based on our proposed method.

- First, we look forward to seeing more high quality open-sourced COVID-19 KGs to be available and applicable to our research in the near future. Due to the limited time and the urgency of work on COVID-19, the access to high quality COVID-19 KGs is not yet ready. Therefore, we only applied one dataset that is available to the proposed model. Hopefully as the global research community cooperate further and deeper in the COVID-19 NLP and text mining domain, we would be able to apply our GFCNet model to other datasets with more varieties, such as drug discovery [2,8,64], contact tracing [60,17] and detection of coronavirus-themed mobile malware [20].
- Second, due to the length of our paper, our proposed model is just a starting point of solving the COVID-19 KGE problem. We have planned further advancement to our method, which include but are not limited to incorporating dynamic data concepts to serve the characteristic and the need for real time data for pandemic development. This improvement of our model also requires more high quality data to support model advancements.
- Third, given the COVID-19 is an unprecedented and new virus, we need to practice further research on how to apply our models safely and appropriately as a foundation block to support various real world medical downstream applications, including clinical decision support and other epidemiological research.

CRediT authorship contribution statement

Zhiwen Xie: Writing - original draft, Data curation, Software, Validation. **Runjie Zhu:** Writing - original draft, Writing - review & editing. **Jin Liu:** Writing - review & editing, Supervision. **Guangyou Zhou:** Writing - review & editing, Methodology,

Supervision. **Jimmy Xiangji Huang:** Writing - review & editing, Methodology, Supervision. **Xiaohui Cui:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 61972290 and 61972173, the National Key R&D Program of China under Grant 2018YFC1604000, the Funamental Research Funds for the Central Universities (No. CCNU22QN015). This work was also supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, an NSERC CREATE award in ADERSIM[12], the York Research Chairs (YRC) program and an ORF-RE (Ontario Research Fund-Research Excellence) award in BRAIN Alliance[13]. The authors thank the anonymous reviewers and associate editor for their insightful comments.

References

[1] Abbas, A., Abdelsamea, M.M., Gaber, M.M., 2020. Classification of COVID-19 in chest x-ray images using detrac deep convolutional neural network. CoRR abs/2003.13815.

[2] K. Abbas, A. Abbasi, D. Shi, N. Ling, L. Yu, B. Chen, S. Cai, Q. Hasan, Application of network link prediction in drug discovery, BMC Bioinform. 22 (2021) 187.

[3] Aggarwal, J., Rabinovich, E., Stevenson, S., 2020. Exploration of gender differences in COVID-19 discourse on reddit, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.

[4] Ba, L.J., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. CoRR abs/1607.06450.

[5] Balazevic, I., Allen, C., Hospedales, T.M., 2019. Tucker: Tensor factorization for knowledge graph completion, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, pp. 5184–5193.

[6] Bansal, T., Juan, D., Ravi, S., McCallum, A., 2019. A2N: attending to neighbors for knowledge graph inference, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 4387–4392.

[7] Bellomarini, L., Benedetti, M., Gentili, A., Laurendi, R., Magnanimi, D., Muci, A., Sallinger, E., 2020. COVID-19 and company knowledge graphs: Assessing golden powers and economic impact of selective lockdown via AI reasoning. CoRR abs/2004.10119.

[8] Bonner, S., Barrett, I.P., Ye, C., Swiers, R., Engkvist, O., Hamilton, W.L., 2021. Understanding the performance of knowledge graph embeddings in drug discovery. CoRR abs/2105.10488.

[9] Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data, in: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pp. 2787–2795.

[10] Cai, L., Yan, B., Mai, G., Janowicz, K., Zhu, R., 2019. Transgcn: Coupling transformation assumptions with graph convolutional networks for link prediction, in: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19–21, 2019, pp. 131–138.

[11] Das, D., Katyal, Y., Dubey, J.V.S., Singh, A.D., Agarwal, K., Bhaduri, S., Ranjan, R.K., 2020. Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-bert embeddings, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.

[12] Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S., 2018. Convolutional 2d knowledge graph embeddings, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pp. 1811–1818.

[13] Feng, M., Hsu, C., Li, C., Yeh, M., Lin, S., 2019. MARINE: multi-relational network embeddings with relational proximity and node attributes, in: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13–17, 2019, pp. 470–479.

[14] Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P.D., Zhang, H., Ji, W., Bernheim, A., Siegel, E., 2020. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis. CoRR abs/2003.05037.

[15] L. Guo, Z. Sun, W. Hu, Learning to exploit long-term relational dependencies in knowledge graphs, in: Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, USA, 2019, pp. 2505–2514.

[16] Y. Han, G. Chen, Z. Li, Z. Geng, F. Li, B. Ma, An asymmetric knowledge representation learning in manifold space, Inf. Sci. 531 (2020) 1–12.

[17] M. Hatamian, S. Wairimu, N. Momen, L. Fritsch, A privacy and security analysis of early-deployed COVID-19 contact tracing android apps, Empir. Softw. Eng. 26 (2021) 36.

[18] He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, pp. 1026–1034.

[19] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp. 770–778.

[20] He, R., Wang, H., Xia, P., Wang, L., Li, Y., Wu, L., Zhou, Y., Luo, X., Guo, Y., Xu, G., 2020. Beyond the virus: A first look at coronavirus-themed mobile malware. CoRR abs/2005.14619.

[21] Huang, X., Hu, Q., 2009. A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval, in: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (Eds.), Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19–23, 2009, ACM. pp. 307–314.

[22] X. Huang, M. Zhong, L. Si, York university at TREC 2005: Genomics track, in: E.M. Voorhees, L.P. Buckland (Eds.), Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, National Institute of Standards and Technology (NIST), 2005.

[23] Ji, G., He, S., Xu, L., Liu, K., Zhao, J., 2015. Knowledge graph embedding via dynamic mapping matrix, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26–31, 2015, Beijing, China, Volume 1: Long Papers, pp. 687–696.

[12] http://www.yorku.ca/adersim

[13] http://brainalliance.ca

[24] Kadlec, R., Bajgar, O., Kleindienst, J., 2017. Knowledge base completion: Baselines strike back, in: Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017, pp. 69–74.

[25] S.M. Kazemi, D. Poole, Simple embedding for link prediction in knowledge graphs, in: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3–8 December 2018, Montréal, Canada, 2018, pp. 4289–4300.

[26] Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.

[27] Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings.

[28] Lacroix, T., Usunier, N., Obozinski, G., 2018. Canonical tensor decomposition for knowledge base completion, in: Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, pp. 2869–2878.

[29] Lin, Y., Liu, Z., Sun, M., 2016. Knowledge representation learning with entities, attributes and relations, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016, pp. 2866–2872.

[30] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for knowledge graph completion, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas, USA, 2015, pp. 2181–2187.

[31] Linda Wang, Zhong Qiu Lin, A.W., 2020. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. arXiv.

[32] Mohamed, S.K., Novácek, V., Vandenbussche, P., Muñoz, E., 2019. Loss functions in knowledge graph embedding models, in: Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) Co-located with the 16th Extended Semantic Web Conference 2019 (ESWC 2019), Portoroz, Slovenia, June 2, 2019, pp. 1–10.

[33] Narin, A., Kaya, C., Pamuk, Z., 2020. Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks. CoRR abs/2003.10849.

[34] Nathani, D., Chauhan, J., Sharma, C., Kaul, M., 2019. Learning attention-based embeddings for relation prediction in knowledge graphs, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 4710–4723.

[35] Nguyen, D.Q., Nguyen, T.D., Nguyen, D.Q., Phung, D.Q., 2018. A novel embedding model for knowledge base completion based on convolutional neural network, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 2 (Short Papers), pp. 327–333.

[36] M. Nickel, V. Tresp, H. Kriegel, A three-way model for collective learning on multi-relational data, in: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, June 28 - July 2011, 2, Bellevue, Washington, USA, 2011, pp. 809–816.

[37] K. Roberts, S.B. Tasmeer Alam, D. Demner-Fushman, K. Lo, I. Soboroff, E. Voorhees, L.L. Wang, W.R. Hersh, Trec-covid: Rationale and structure of an information retrieval shared task for covid-19, J. Am. Med. Inform. (2020), https://doi.org/10.1093/jamia/ocaa091.

[38] D. Ruffinelli, S. Broscheit, R. Gemulla, You CAN teach an old dog new tricks! on training knowledge graph embeddings, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, April 26–30, 2020. Ethiopia, 2020.

[39] Schlichtkrull, M.S., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M., 2018. Modeling relational data with graph convolutional networks, in: The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings, pp. 593–607.

[40] Shen, I., Zhang, L., Lian, J., Wu, C., González-Fierro, M., Argyriou, A., Wu, T., 2020. In search for a cure: Recommendation with knowledge graph on CORD-19, in: KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23–27, 2020, pp. 3519–3520.

[41] Soni, S., Roberts, K., 2020. An evaluation of two commercial deep learning-based information retrieval systems for COVID-19 literature. CoRR abs/2007.03106.

[42] Z. Sun, Z. Deng, J. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 2019.

[43] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G., 2016. Complex embeddings for simple link prediction, in: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016, pp. 2071–2080.

[44] Vashishth, S., Sanyal, S., Nitin, V., Agrawal, N., Talukdar, P., 2020. Interacte: Improving convolution-based knowledge graph embeddings by increasing feature interactions, in: Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)., pp. 3009–3016.

[45] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2017. Graph attention networks. CoRR abs/1710.10903.

[46] Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D., Weld, D.S., Etzioni, O., Kohlmeier, S., 2020a. CORD-19: the covid-19 open research dataset. CoRR abs/2004.10706.

[47] Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., Liu, W., Chauhan, A., Guan, Y., Li, B., Li, R., Song, X., Ji, H., Han, J., Chang, S., Pustejovsky, J., Rah, J., Liem, D., Elsayed, A., Palmer, M., Voss, C.R., Schneider, C., Onyshkevych, B.A., 2020b. COVID-19 literature knowledge graph construction and drug repurposing report generation. CoRR abs/2007.00576.

[48] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., Xu, B., 2020c. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). medrXiv doi: 10.1101/2020.02.14.20023028.

[49] Wang, X., Liu, W., Chauhan, A., Guan, Y., Han, J., 2020d. Automatic textual evidence mining in COVID-19 literature. CoRR abs/2004.12563.

[50] Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014. Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, Québec City, Québec, Canada, pp. 1112–1119.

[51] Wise, C., Ioannidis, V.N., Calvo, M.R., Song, X., Price, G., Kulkarni, N., Brand, R., Bhatia, P., Karypis, G., 2020. COVID-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature. CoRR abs/2007.12731.

[52] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, M. Wilson, Drugbank 5.0: a major update to the drugbank database for 2018, Nuclc Acids Res. 46 (2017).

[53] Xiao, H., Huang, M., Zhu, X., 2016. From one point to a manifold: Knowledge graph embedding for precise link prediction, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016, pp. 1315–1321.

[54] Y. Xiao, R. Li, X. Lu, Y. Liu, Link prediction based on feature representation and fusion, Inf. Sci. 548 (2021) 1–17.

[55] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, S. Yu, A survey on incorporating domain knowledge into deep learning for medical image analysis, Medical Image Anal. 69 (2021) 101985.

[56] Xie, Z., Zhou, G., Liu, J., Huang, X., 2020. Reinceptione: Relation-aware inception network with joint local-global structural information for knowledge graph embedding, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 5–10, 2020, pp. 5929–5939.

[57] Z. Xie, R. Zhu, J. Liu, G. Zhou, J.X. Huang, Hierarchical neighbor propagation with bidirectional graph attention network for relation prediction. IEEE ACM Trans, Audio Speech Lang. Process. 29 (2021) 1762–1773.

[58] Yang, B., Yih, W., He, X., Gao, J., Deng, L., 2015. Embedding entities and relations for learning and inference in knowledge bases, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.

[59] X. Yin, J.X. Huang, Z. Li, X. Zhou, A survival modeling approach to biomedical search result diversification using wikipedia, IEEE Trans. Knowl. Data Eng. 25 (2013) 1201–1212.

[60] Yu, P., Tan, C., Fu, H., 2022. Epidemic source detection in contact tracing networks: Epidemic centrality in graphs and message-passing algorithms. CoRR abs/2201.06751.

[61] S. Yu, G. Gu, A. Barnawi, S. Guo, I. Stojmenovic, Malware propagation in large-scale networks, IEEE Trans. Knowl. Data Eng. 27 (2015) 170–179.

[62] S. Yu, M. Liu, W. Dou, X. Liu, S. Zhou, Networking for big data: A survey, IEEE Commun. Surv. Tutorials 19 (2017) 531–549.

[63] X. Zeng, X. Song, T. Ma, X. Pan, Y. Zhou, Y. Hou, Z. Zhang, G. Karypis, F. Cheng, Repurpose open data to discover therapeutics for COVID-19 using deep learning, J. Proteome Res. (2020), https://doi.org/10.1021/acs.jproteome.0c00316.

[64] R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, H. Kilicoglu, Drug repurposing for COVID-19 via knowledge graph completion, J. Biomed. Informatics 115 (2021) 103696.

[65] S. Zhang, Y. Tay, L. Yao, Q. Liu, Quaternion knowledge graph embeddings, in: B.C. Vancouver (Ed.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8–14 December 2019, Vancouver, BC, Canada, 2019, pp. 2731–2741.

[66] Zhao, S., Qin, B., Liu, T., Wang, F., 2020. Biomedical knowledge graph refinement with embedding and logic rules. CoRR abs/2012.01031.