IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.8, August 2020 E-ISSN: 2347-2693

CompNet : A novel Knowledge Graph Embedding Technique for Link Prediction

Kohsheen Tiku[1*], Jayshree Maloo[2], R. Indra[3]

[1,2,3]Dept. of Information Science, BMS College of Engineering, Visvesvaraya Technological University, Bangalore, India

[*]*Corresponding Author: kohsheen.t @gmail.com, Tel.: +91 8867902892*

DOI: https://doi.org/10.26438/ijcse/v8i8.14 | Available online at: www.ijcseonline.org

Received: 12/May/2020, Accepted: 18/May/2020, Published: 31/Aug/2020

Abstract — A Knowledge Graph(KG) is a graph that contains information about anything and everything in the world, these graphs can be represented by plain-simple nodes and links. Knowledge Graph Embedding refers to attaining valuable information about every node present in the graph, essentially it could be defined as representing a node as a low-dimensional vector. Knowledge graph Embedding techniques ha**ve** become an increasingly popular research **topics.** Despite all the efforts invested in creation and maintenance of knowledge graphs only contain a part of what it contains is true and it also still consists of missing facts. Prediction of these missing facts in scientific term can be known as Link Prediction. Several recent works use deep learning approaches to generate richer and more expressive embedding. However, observations show that the following methods are very expensive computationally. This paper proposes a novel approach for generating Graph Embeddings which can be used to perform the task of Link Prediction i.e. interpreting missing data. 'CompNet' uses a modified version of the complEx and Cluster-Graph Convolution Network (Cluster-GCN) algorithms. Laying out certain constraints on the ComplEx algorithm i.e. discussed in detail through this paper, can improve the time complexity drastically. 'CompNet' is tested on the large-scale dataset Freebase, wherein it out-performs the traditional approaches (translation based approaches and semantic search approaches) in terms of Mean Reciprocal Rank (MRR) and also utilizes low storage space.

Keywords—Convolutional Neural Networks, ComplEx, Knowledge Graphs, Link Prediction, Graph Embedding

I. INTRODUCTION

Knowledge bases organize and store factual information, allowing for a variety of applications including responding to questions (Yao and Van Durme 2014; Bao et al . 2014; Seyler, Yahya, and Berberich 2015; Bordes et al . 2015; Dong et al . 2015) and retrieval of information (Kotov and Zhai 2012; Dalton, Dietz, Allan 2014; Xiong and Callan 2015b; 2015). Even the largest bases of knowledge (e.g. DBPedia, Wikidata, or Yago) are incomplete despite the enormous effort invested in their maintenance, and the lack of coverage damages downstream applications.

Knowledge Graph expresses data in the form of a triplet, consisting of (head, tail, relation). The head and tail, both represent nodes/entities whereas the relation represents the edge between the two entities. For instance, (Budapest, capital_of, Hugary) is a triplet constructed following the above principles.

Even though there exist knowledge graphs such as WikiData, the truth remains that it still is incomplete, and a lot of information is yet to be discovered from the former. The only viable method which can be applied in such a case is a link prediction algorithm, which allows one to understand more depth of relationship present between any two entities and predict future links. Link Prediction involves predicting missing entities or relations in an incomplete knowledge graph.

Link prediction can be applied when there exists a vector representation for each node/relation. This vector representation is merely a feature vector of that particular node/entity. Considering that knowledge graphs are large scale graphs and are sparse in nature, it is necessary to ensure that these representations are low dimensional and precise, this method is referred to as Knowledge Graph Embedding generation.

This paper proposes a new technique called CompNet. The novel CompNet algorithm is suitable for extracting vector representations by assuming that the representation space is Complex and also utilizes the power of convolutional network on clusters to improve the vector representations.
The remaining part of this paper is organized as follows: Section II contain the related work that focuses particularly on two subdivisions of knowledge representation learning, i.e. Scoring Function(ComplEx) and Encoding model (Cluster GCN[1][2]). Section III contains the architecture and essential steps of that have been deployed in this paper to obtain accuracy similar to previous algorithms for large scale graphs taking into consideration the computational complexity, explaining the methodology with flow chart and Section IV describes results of the CompNet method in comparison with the traditional approaches.

II. RELATED WORK

Recent years have witnessed growing interest in learning new features without expert knowledge and in the case of generating Knowledge graph embeddings. The problem of link prediction can be formalized as a point-wise learning to rank problem, where the objective is to learn a scoring function $\psi: E \times R \times E \rightarrow R$, given an input triple x = (h, r, t), its score $\psi(x)$ R is proportional to the probability that the fact encoded by x being true. The traditional approach to generating embedding revolves around a Scoring function which can be roughly classified into Translation based models and Semantic Search Based Models. Translation Based Models generate embedding based on the distance between any two nodes. nodes. TransE [3],TransR [4], TransH [5] are some instances of Translation based Models. On the other hand, Semantic based models use a scoring function dependent on similarity. RESCAL [6], DistMul[7], complEx[8] are examples of Semantic based Model.

Table 1. Time Complexity of Existing Methods

SL No.	Existing Methods		
	Name	***Space Complexity***	***Time Complexity***
1.	TransE	O(nd+md)	O(d)
2.	TransR	O(nd+mdk)	O(dk)
3.	RESCAL	O(nd+md^2)	O(d2)
4	TransH	O(nd+md)	O(d)
5.	DistMul	O(nd+md)	O(d2)
6.	ComplEx	O(nd+md)	O(d)

ComplEx represents each entity e ∈ E(Entity) as a complex-valued vector e ∈ c^d , and each relation r ∈ R(Relation) a complex-valued vector r ∈ c^d, where 'd' is the dimensionality of the embedding space. Each x ∈ c^d, consists of a real vector component Re(x) and an imaginary vector component Im(x), i.e., x = Re(x) + i*Im(x). For any given triple $(e_i, R_k, e_i) \in$ E × R × E, the scoring function of complEx model involves multi linear dot product of all the three, i.e. head, tail and relation.

Various –Knowledge acquisition techniques that help to complete an existing Knowledge graph such as GCN(Graph Convolutional Network), GAN , RL. Google has recently developed an algorithm known as Cluster-GCN[9], Which has essentially reduced computational costs, it increases exponentially with the number of GCN layers or a wide space requirement to hold the entire graph and store each node in memory.

III. METHODOLOGY

A. Overview of the Methodology

The motivation behind this paper is to make the algorithm more suitable for large scale graphs, and to enhance the perforce of complEx algorithm. By discussing the bottle neck of the complex algorithm it is clear that the time complexity of this algorithm is the least among the Sematic Scoring functions and is very efficient. An improvement on the time complexity and the accuracy has been discussed in this paper.

To ensure that this algorithm is suited for larger graphs, we have mixed the ideology behind the Cluster GCN algorithm with the complex method. The basic aim of the Cluster-GCN is to split the graph into clusters for reducing time-complexity.

Firstly, understanding that complEx captures symmetric and anti-symmetric relations, we followed the following steps:

1. These relations would be more valuable if were derived from similar or related entities rather than the unrelated ones.
2. This can be achieved when similar entities (closely related) are grouped into clusters. For instance, one cluster can consist actor names, movies, TV Shows, their place of birth, but all these entities should be closely related.
3. Graph clustering methods such as Metis [10 in order to better capture the clustering and group structure of the graph, the goal is to create the partitions over the vertices in the graph so that ties between clusters are much more than cluster ties.
4. When the clustering process is over, each cluster is treated as a single network, and the embeddings for each entity and relation is generated simultaneously for each of these clusters.

Secondly, we applied two constraints on the scoring function of complEx algorithm has proven to increase the time complexity of the algorithm as stated in Simple Constraints on ComplEx [11]. The constrains are described as follows:

1. The first constraint being –

Knowledge Graphs are very large and sparse in nature. Hence, storing negative properties on an entity is uneconomical. We require entities to have non-negative (and bounded) vectoral representations. By, negative examples we mean 'Cars are not animals', such sort of information is not required. Nothing can be predicted from such information and it only leads to the increase of computation time. Positive examples of the same entity could be 'Cars have four wheels'. Hence, it is wise negative. This step will lead us to more predictive and interpretable embeddings.

In order to effectively evaluate different entities on the same scale, we also require representations of entities to remain within the $[0,1]^d$, as approximately Boolean embeddings (Kruszewski et al., 2015)[12], i.e.,

$$0 \leq \text{Re(e) Im(e)} \leq 1,$$

where 'e' is the representation for entity e, with its real and imaginary components denoted by Re(e) and Im(e) ; 0 and 1 are d-dimensional vectors with all their entries being 0 or 1.

2. The second constraint being –

By stating that a person born in a country is very likely, but not necessarily, to have a nationality of that country. Each such relation pair is associated with a weight to indicate the confidence level of entailment. A larger weight stands for a higher level of confidence. We denote by $r_p y \rightarrow r_q$ the approximate entailment between relations r_p and r_q, with confidence level γ. For any two entities e_i and e_i, if $(e_{iJ} r_p, e_i)$ is a true fact with a high score $\varphi(e_i, r_p, e_i)$, then the triple (e_i, r_p, e_i) with an even higher score should also be predicted as a true fact by the embedding model.

$$\lambda\big(Re(r_p)\, Re(r_q)\big) \leq a\ ;$$
$$\lambda\big(Im(r_p)\, Im(r_q)\big)^2 \leq \beta$$

B. Implementation Details

Further we evaluate the performance of this model on the FB15K[13] and WN18[13] datasets. FB15K is a subcategory of Freebase, a compiled Knowledge base of factual information, whereas WN18 is a subcategory of Wordnet, a database with lexical interpersonal relations.
The entire implementation is done by reusing the code of OpenKE[13] for ComplEx and incorporating our changes. We have used TensorFlow for implementation purpose.

Results and Discussion

There are four different cases of Link Prediction that we have implemented. The four cases can be classified into the following: *Predicting Head, Predicting Tail, Predicting Relation, predicting if the triplet is valid or not.*

Evaluation Metrics

1. *Rank:* For instance, the system is required to replace the question mark with Moscow given the triple (XYZ, born - in - city,?).. If the system is able to achieve the answer to the question mark in the first attempt it's rank is 1. If in the second, the rank is 2 and so on.
2. *HITS@* K: For each instance list the rate of the correct entities appearing in the top k entries. That number may exceed 1 if there is more than one true entity in the average k-truncated list.
3. *MRR:* Mean Reciprocal Rank is the mean of all reciprocal ranks for the cases in which the true answer appears over the test set (1/rank).

We have measured the performance of various methods on the two datasets – FB15K and WN18 and have presented the results in comparison to our approach in the following tables.

Table 2: Link Prediction Results on FB15K

Method	MRR		HITS@K		
	Filter	Raw	1	3	10
TransE	0.380	0.221	0.231	0.472	0.641
DistMul	0.654	0.242	0.546	0.733	0.824
ComplEx	0.692	0.242	0.599	0.759	0.840
Simple Constrains on complEx	0.803	0.244	0.761	0.831	0.874
CompNet(this work)	0.862	0.247	0.84	0.865	0.931

Table 3: Link Prediction Results on WN18

Method	MRR		HITS@K		
	Filter	Raw	1	3	10
TransE	0.454	0.335	0.089	0.823	0.934
DistMul	0.822	0.532	0.728	0.914	0.936
ComplEx	0.941	0.587	0.947	0.945	0.936
Simple Constrains on complEx	0.943	0.594	0.940	0.945	0.948
CompNet(this work)	0.957	0.598	0.953	0.961	0.969

IV. CONCLUSION AND FUTURE SCOPE

This paper investigates the potential of using a technique derived from Cluster-GCN in another popular approach of complex. The use of such clusters minimises the consideration of unrelated entities and also parallel processing on these clusters improves the time complexity. Since, these clusters are much smaller than the actual graph, computation is simpler after that METIS algorithm has been applied. Using some constraints on the scoring function of ComplEx algorithm increases the accuracy of link prediction as well.

ACKNOWLEDGMENT

We would like to thank the principal of BMS College of Engineering as well as the Dept of Information Science, for helping and guiding us through the entire process.

REFERENCES

[1] Schlichtkrull, Michael & Kipf, Thomas & Bloem, Peter & Berg, Rianne & Titov, Ivan & Welling, Max. Modeling Relational Data with Graph Convolutional Networks. 10.1007/978-3-319-93417-4_38, **2018.**

[2] Dai, Shaozhi & Liang, Yanchun & Liu, Shuyan & Wang, Ying & Shao, Wenle & Lin, Xixun & Feng, Xiaoyue. Learning Entity and Relation Embeddings with Entity Description for Knowledge Graph Completion. 10.2991/icaita-18.2018.49,**2018**.

[3] Antoine Bordes, Nicolas Usunier, Alberto GarciaDuran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multirelational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pages **2787**– **2795**, **2013.**

[4] Feng, Jianlin. Knowledge Graph Embedding by Translating on Hyperplanes, **2014**.

[5] TransH: Knowledge Graph Embedding by Translating on Hyperplanes**.** Zhen Wang, Jianwen Zhang, Jianlin Feng, Zheng Chen. *AAAI* , **2014**.
[6] Nickel, Maximilian & Tresp, Volker & Kriegel, Hans-Peter. A Three-Way Model for Collective Learning on Multi-Relational Data.. Proceedings of the 28th International Conference on Machine Learning, ICML 2011. 809-816, **2011.**
[7] Yang, Bishan & Yih, Wen-tau & He, Xiaodong & Gao, Jianfeng & Deng, li. (2014). Embedding Entities and Relations for Learning and Inference in Knowledge Bases. Thomas L., "*A Scheme to Eliminate Redundant Rebroadcast and Reduce Transmission Delay Using Binary Exponential Algorithm in Ad-Hoc Wireless Networks*", International Journal of Computer Sciences and Engineering, Vol.**3,** Issue.**8**, pp.**1**-**6**, **2017**.
[8] ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume **48** Pages **2071**–**2080**, June **2016**
[9] KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Pages **257**–**266** , July **2019**
[10] George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. 20, 1, **359**–**392**, **1998.**
[11] Improving Knowledge Graph Embedding Using Simple Constraints, Boyang Ding and Quan Wang and Bin Wang and Li Guo, year **2018**.
[12] German Kruszewski, Denis Paperno, and Marco Baroni. 2015. Deriving Boolean structures from distributional vectors. Transactions of the Association for Computational Linguistics 3:**375**–**388**.
[13] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multirelational data," in NIPS, pp. **2787**–**2795**. **2013**.

AUTHORS PROFILE

Ms. Kohsheen Tiku is currently pursuing her Bachelor of Engineering from BMS College of Engineering, Bangalore during the period 2016 -2020. She is a member of IEEE & ACM since 2017.

Ms. Jayshree Maloo is currently pursuing her Bachelor of Engineering from BMS College of Engineering, Bangalore during the period 2016 -2020.

Mrs. Indra R is has completed her masters degree in Computer Science from PES University in the yeat 2014 and has completed her Bachelor of Engineering from BMS College of Engineering, Bangalore in the year 2011. She also posses a Diploma in Computer Science from Oxford Polytechnic, Technical Education in Karnataka, 2007. She is also a member of IEEE and ACM.

