CS249 Project Proposal(Spring 2022)

Yikai Zhu [1]

Abstract

Knowledge graph embedding(KGE) is a techhnique for learning continuous embeddings for entities and relations in the knowledge graph. Despite the widespread use of KGE models, little is known about the security vulnerabilities that might disrupt their intended behaviour. In this project we will study data poisoning attacks against KGE models for link prediction.

1. Introduction

Knowledge graphs have become a critical resource for a large collection of real world applications, such as information extraction, question answering and recommendation system. In knowledge graphs, knowledge facts are usually stored as *(head entity, relation, tail entity)* triples. Although such triples can effectively record abundant knowledge, their underlying symbolic nature makes them difficult to be directly fed to many machine learning models. Hence, knowledge graph embedding, which projects the symbolic entities and relations into continuous vector space, has quickly gained significant attention. Lots of embedding models have be proposed and they have been proved to preserve the inherent characteristics of entities and relations while enabling the use of these knowledge facts for a large variety of downstream tasks such as link prediction, question answering and recommendation.

Despite the increasing success and popularity of Knowledge graph embedding models, their robustness has not been fully analyzed. In fact, many knowledge graphs are built upon unreliable or even public data sources. For instance, the well known **Freebase** harvests its data from various sources including individual, user-submitted wiki contributions. The openness of such data unfortunately would make KGE models vulnerable to malicious attacks. (Pujara et al., 2017) first find that most popular KGE models are sensitive to sparsity and noise. (Bhardwaj et al., 2021) proposed a effective adversarial deletion strategy which can successfully degrade the predictive performance of KGE models.

However, (Bhardwaj et al., 2021) and several other attack methods can not degrade the performance with adversarial additions. Most methods have the same effect as random deletion. While in reality, it's more possible to add malicious knowledge to the knowledge graph instead of deleting existing knowledge. Therefore, we want to find whether there exist a more efficient addition to degrade the performance of KGE models.

In this paper, we want to systemically investigate the sensitivity of KGE model, through test the performance of different KGE models with different adversarial attack strategies. Due to unique characteristics characteristics of knowledge graph and it embedding models, existing adversarial attack methods on graph data cannot be directly applied to attack KGE models. After a brief survey, we found (Bhardwaj et al., 2021) is now the state-of-art adversarial attack method and can be applied to any KGE models. We also find the model predictions are likely influenced by a group of training triples (Basu et al., 2020) while current method focus on single triple. We will investigate these mehods for KGE models in this project, too.

References

Basu, S., You, X., and Feizi, S. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pp. 715–724. PMLR, 2020.

Bhardwaj, P., Kelleher, J., Costabello, L., and O'Sullivan, D. Adversarial attacks on knowledge graph embeddings via instance attribution methods. *arXiv preprint arXiv:2111.03120*, 2021.

Pujara, J., Augustine, E., and Getoor, L. Sparsity and noise: Where knowledge graph embeddings fall short. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 1751–1756, 2017.