IterDE: An Iterative Knowledge Distillation Framework for Knowledge Graph Embeddings

Jiajun Liu, Peng Wang*, Ziyu Shang, Chenxiao Wu

School of Computer Science and Engineering, Southeast University
{jiajliu, pwang, ziyus1999, chenxiaowu}@seu.edu.cn

Abstract

Knowledge distillation for knowledge graph embedding (KGE) aims to reduce the KGE model size to address the challenges of storage limitations and knowledge reasoning efficiency. However, current work still suffers from performance drops when compressing a high-dimensional original KGE model to a low-dimensional distillation KGE model. Moreover, most work focuses on the reduction of inference time but ignores the time-consuming training process of distilling KGE models. In this paper, we propose IterDE, a novel knowledge distillation framework for KGEs. First, IterDE introduces an iterative distillation way and enables a KGE model to alternately be a student model and a teacher model during the iterative distillation process. Consequently, knowledge can be transferred in a smooth manner between high-dimensional teacher models and low-dimensional student models, while preserving good KGE performances. Furthermore, in order to optimize the training process, we consider that different optimization objects between hard label loss and soft label loss can affect the efficiency of training, and then we propose a soft-label weighting dynamic adjustment mechanism that can balance the inconsistency of optimization direction between hard and soft label loss by gradually increasing the weighting of soft label loss. Our experimental results demonstrate that IterDE achieves a new state-of-the-art distillation performance for KGEs compared to strong baselines on the link prediction task. Significantly, IterDE can reduce the training time by 50% on average. Finally, more exploratory experiments show that the soft-label weighting dynamic adjustment mechanism and more fine-grained iterations can improve distillation performance.

Introduction

Knowledge graphs (KGs) describe concepts and facts in graph models (Dong et al. 2014), where knowledge is stored as triples. With the increasing sizes of KGs such as Wikipedia (Bizer et al. 2009) and Yago (Suchanek, Kasneci, and Weikum 2007), efficient knowledge graph embedding (KGE), which embeds triples into a continuous vector space, plays a pivotal role in downstream applications such as question answering (Bordes, Weston, and Usunier 2014), recommendation system (Zhang et al. 2016) and knowledge graph completion (Lin et al. 2015). Most KGE models such as TransE (Bordes et al. 2013), ComplEx (Trouillon et al. 2016), SimplE (Kazemi and Poole 2018), RotatE (Sun et al. 2019) have shown better performances with higher embedding dimensions and larger model sizes, however, that also leads to slower inference efficiency for practical applications. Specifically, the 512-dimensional KGE models have 7-15 times more embedding layer parameters and 2-6 times more inference time than the 32-dimensional KGE models (Zhu et al. 2022). Therefore, it is a nontrivial problem to compress KGEs from high-dimensional teacher models to low-dimensional student models while maintaining excellent performance. In realistic applications, KGE models are often required to simultaneously keep high performance and fast inference speed. For example, financial investors need to get accurate and fast market decision aids from financial KGs via edge devices. In this scenario, KGE models can be compressed by knowledge distillation and then deployed to edge devices to help financial investors make faster and more accurate decisions.

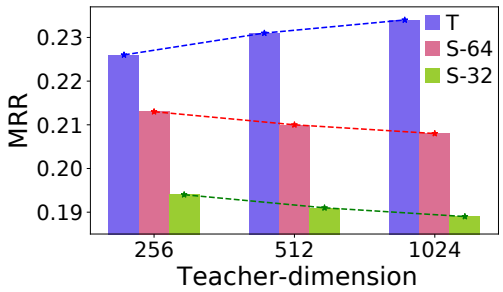

Figure 1: The phenomenon of *good teachers cannot always teach good students in KGEs based on KD*: with the increase of dimension (from 256 to 1024) of teacher T, the performances of students (S-64 denotes 64 dimensions and S-32 denotes 32 dimensions) drop. Results are obtained on WN18RR dataset with TransE.

Knowledge distillation (KD) is a popular technique for model compression, where a larger model serves as a teacher model, and a smaller model as a student model tries to mimic the output of the teacher model (Hinton et al. 2014). Recently, although several compression methods for KGEs

*Corresponding author.