Learning Better Representations Using Auxiliary Knowledge

Saed Rezayi

University of Georgia
Department of Computer Science
saedr@uga.edu

Abstract

Representation Learning is the core of Machine Learning and Artificial Intelligence as it summarizes input data points into low dimensional vectors. This low dimensional vectors should be accurate portrayals of the input data, thus it is crucial to find the most effective and robust representation possible for given input as the performance of the ML task is dependent on the resulting representations. In this summary, we discuss an approach to augment representation learning which relies on external knowledge. We briefly describe the shortcoming of the existing techniques and describe how an auxiliary knowledge source could result in obtaining improved representations.

Introduction

Neural Network-based Representation Learning has gained traction over the past few years for a variety of modalities and applications. In Natural Language Processing, for instance, RL allows us to distinguish between bank (a financial institute) and bank (the land alongside a river) or to answer questions such as how similar is "espresso" to "spaghetti" which requires an underlying knowledge to understand they are both of Italian origin.

Successful Representation Learning models require huge amounts of training data and computational resources which are very expensive to acquire. However, pretrained models are reasonably available and can be augmented and fine-tuned for specific applications. This is particularly helpful when there is a limitation in the task at hand. For instance, knowledge graph embedding methods (i.e., learning representations for the entities and relations of a knowledge graph) have several drawbacks such as ignoring contextualized information and suffering from sparsity, or language models has the shortcoming of being limited to the vocabularies of the corpus they are trained on.

In such scenarios, an auxiliary source of knowledge can improve the quality of the learned representations and help with the inherent limitations of the vanilla models. For example, (Xie et al. 2016) proposed a representation learning method for knowledge graphs via embedding entity descriptions, or (Malaviya et al. 2020) used pre-trained language models to improve the node embeddings for graph completion task. In what follows, we apply similar design idea to propose improved solutions for various NLP and knowledge graph Embedding tasks. We, furthermore, discuss robustness and explore ideas to enhance RL in presence of an adversary.

Text as External Source of Knowledge

RQ1: How RL can benefit from incorporating knowledge from an unstructured, external source of data?

Description: In (Rezayi et al. 2021b) we showed incorporating additional textual entities to a graph from an external source such as WordNet could be advantageous in obtaining more meaningful embeddings for the entities of knowledge graphs which improves the performance of the downstream task, e.g., link prediction or node classification. Previous work (Kartsaklis, Pilehvar, and Collier 2018) has partially addressed the issue of sparsity by enriching knowledge graph entities based on *"hard"* co-occurrence of words present in the entities of the knowledge graphs and external text, while we achieve *"soft"* augmentation by proposing a knowledge graph enrichment and embedding framework. Given an original knowledge graph, we first generate a rich but noisy augmented graph using external texts in semantic and structural level. To distill the relevant knowledge and suppress the introduced noise, we design a graph alignment term in a shared embedding space between the original and augmented graph. This work was published in NAACL 2021.

Knowledge Graph as External Source of Knowledge

RQ2: Can knowledge graphs be used as an external source of knowledge to assist in obtaining improved emebddings?

Description: In (Rezayi et al. 2021a), we posed the question of whether an external source of knowledge can guide the search behavior of a user, and we proposed to find similar entities to the user query with the aid of an external knowledge graph, i.e., using the following pipeline: entity linking + customized link prediction which yields to the introduction of a new entity that satisfies the user's information need. This work was published in IEEE BigData-2021.

In another research study and for database matching application in agricultural domain, where the domain knowledge is very scarce, we demonstrated that enhancing a language model with a domain-specific knowledge graph, namely FoodOn, can boost the precision@1 by 20% compared to fine-tuned language model without the help of the external knowledge graph (Rezayi et al. 2022). In this work, we fine-tuned a transformer-based language model with a large corpus of agricultural literature (46,446 papers, and more than 300 million tokens). This work was published in IJCAI-2022.

Cross Domain Clustering

RQ3: Given clustering short text is challenging (unknown data distribution, topic evolution, semantic sparsity), can we exploit a labeled dataset to model the underlying distribution of a target dataset?

Description: Because of the inherent challenges of clustering (unknown number of clusters and unknown data distribution) we propose a cross domain clustering (XDC) framework, which leverages adversarial learning to train an adaptive clustering model across text domains. We propose to jointly exploit a labeled source domain and an unlabeled target domain during model training. Owing to domain adversarial learning, the distribution shift across source and target domains should be mitigated. Additionally, we implement linkage-based clustering approach which is agnostic of the number of clusters. Here the external source of knowledge is a labeled dataset. For this work a student abstract was already published in AAAI-2022 and the full version was accepted in SDM-2023 (Rezayi et al. 2023).

External Knowledge to Enhance Robustness

RQ4: Knowledge Graph Embedding methods are prone to removal and addition attacks that degrades the performance of the link prediction task, can we utilize a form of external knowledge to mitigate the effect of the attack and improve the robustness?

Description: An adversarial attack against knowledge graph embedding aims at identifying the training instances that are most influential to the model's predictions on test instances. Existing works in this area are limited (Bhardwaj et al. 2021; Betz, Meilicke, and Stuckenschmidt 2022), and even more limited is the design of a defense mechanism to alleviate the effect of adversarial attacks against knowledge graph embedding methods.

Research Proposal: The simplest form of enhancing the embeddings of knowledge graph entities is using an external knowledge source as follows:

$$\hat{\mathbf{e}}_i = \mathbf{e}_i + \mathbf{w}_i\mathbf{M}$$

In the above equation $\mathbf{e}_i$ is the entity representation and $\mathbf{w}_i$ is the word embedding of the text of the entity, and $\mathbf{M}$ is a transformation matrix to project the word embedding space into the graph embedding space. The issue with this approach is that it is not targeted toward the attacked entities. One improvement in this approach is to focusing on the neighborhood of the target facts. Furthermore we can improve the structure of the knowledge graph in the data space by incorporating relevant entities from external source. This has been proven to be effective (Rezayi et al. 2021b). For this work literature review has been done and the baseline has been established. We will consider submitting this work to NeurIPS-2023 or AAAI-2024 once the experimental results have been achieved,

Timeline

Following table illustrates the timeline of my PhD completion after my internship which concludes by the end of year 2022.

	Jan	Feb	Mar	Apr	May
Implementation	x	x			
Evaluation		x	x		
Submission			x	x	
Thesis Writing		x	x	x	x

Table 1: PhD timeline for the year 2023.

References

Betz, P.; Meilicke, C.; and Stuckenschmidt, H. 2022. Adversarial explanations for knowledge graph embeddings. IJCAI.

Bhardwaj, P.; Kelleher, J.; Costabello, L.; and O'Sullivan, D. 2021. Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods. In *EMNLP*, 8225–8239.

Kartsaklis, D.; Pilehvar, M. T.; and Collier, N. 2018. Mapping Text to Knowledge Graph Entities using Multi-Sense LSTMs. In *EMNLP*, 1959–1970.

Malaviya, C.; Bhagavatula, C.; Bosselut, A.; and Choi, Y. 2020. Commonsense Knowledge Base Completion with Structural and Semantic Context. In *AAAI*.

Rezayi, S.; Lipka, N.; Vinay, V.; Rossi, R. A.; Dernoncourt, F.; King, T. H.; and Li, S. 2021a. A Framework for Knowledge-Derived Query Suggestions. In *IEEE Big Data*, 510–518. IEEE.

Rezayi, S.; Liu, Z.; Wu, Z.; Dhakal, C.; Ge, B.; Zhen, C.; Liu, T.; and Li, S. 2022. AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition. In *IJCAI*, 5150–5156.

Rezayi, S.; Zhao, H.; Kim, S.; Rossi, R.; Lipka, N.; and Li, S. 2021b. Edge: Enriching Knowledge Graph Embeddings with External Text. In *NAACL*, 2767–2776.

Rezayi, S.; Zhao, H.; Zhu, R.; and Li, S. 2023. XDC: Adversarial Adaptive Cross Domain short text clustering. Forthcoming.

Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M. 2016. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, volume 30.