ON THE EVALUATION OF HETEROGENEOUS LINK PREDICTION METHODS ON AN RNA-CENTERED KNOWLEDGE GRAPH

MAYANK PRATAP SINGH - 14383A
10/12/2024

1. Institution where the Thesis work was carried out

This Thesis was conducted at the AnacletoLab, University of Milan, Department of Computer Science, under the supervision of Prof. Marco Mesiti and Dr. Emanuele Cavalleri.

2. Initial context

Several embedding techniques have been proposed in the last few years for creating a vectorial representation of heterogeneous graphs that rely on the use of graph neural networks and random walks. The embedding representation is then exploited for predicting the plausibility of new links that can be included in the heterogeneous graph. Purpose of this Thesis is to compare different embedding techniques as well as GNN-based techniques in the context of RNA-KG ([7]), a recently proposed RNA- centered knowledge graph with more than 12.7 million edges representing different kinds of relationships in which RNA molecules can be involved.

3. Objectives of the work

The objective of this Thesis is to evaluate the efficiency of graph representation learning (GRL [5]) techniques, including both non-neural and neural models, for tasks such as link prediction and node classification.

4. Description of work done

The Thesis explores GRL methods tailored for heterogeneous graphs applied on RNA-KG. We highlighted the ability of heterogeneous GRL techniques to enhance the interpretability and predictive power of RNA-KG.

Various methods can be employed to generate node embeddings, which encode the graph structure and biological semantics of RNA-KG. We consider DistMult and Complex, as well as Graph Neural Networks (GNNs) based methods. Among them, we considered GCN and RGCN, which are designed to handle the heterogeneous nature of a KG. Traditional machine learning methods, i.e. Decision Trees

and Random Forests, are used as classifiers to predict the likelihood of relationships between pairs of nodes based on their embeddings.

5. Technologies involved

We tested both traditional methods (DistMult [4] and ComplEx [2]) and neural models (GCN [3] and RGCN [1]), implemented in GRAPE [6] and pytorch [8] libraries. Random Forest classifier and Decision Tree were employed to evaluate the embeddings generated by these models. t-SNE representations were used to interpret node and edge embeddings. All experiments were conducted on a T4 GPU support.

6. Skills and results achieved

The results presented in the Thesis explore the performance of two non-neural models, DistMult and ComplEx, and two neural models, GCN and RGCN, applied to RNA-KG. These models were evaluated on their ability to perform link prediction and node classification.

The results showed that RGCN outperformed all other models, demonstrating superior accuracy in both link prediction and node classification tasks. Its ability to learn distinct representations for different types of relationships allowed it to uncover patterns within RNA-KG that non-neural models could not capture. The key difference in performance between non-neural and neural models lies in their capacity to represent relational diversity and structural complexity. While non-neural models like DistMult and ComplEx are computationally efficient and work well for simpler tasks, they fall short in scenarios requiring the modeling of multi-relational dependencies. Neural models, particularly RGCN, leverage deep learning to dynamically learn representations that adapt to the complexity of the graph, providing a significant edge in performance. For example, in link prediction tasks, RGCN effectively captured multi-hop relationships, which were beyond the capabilities of non-neural models. Similarly, in node classification, R-GCN's ability to integrate relational context allowed it to achieve higher precision and recall compared to simpler models.

7. Bibliography

1. Schlichtkrull, Michael et al. “Modeling relational data with graph convolutional networks.” 2017, arXiv:1703.06103.
2. Trouillon, T et al. “Complex embeddings for simple link prediction. In Proceedings of the International Conference on Machine Learning (ICML).” 2016.
3. Wu, Felix et al. “Simplifying graph convolutional networks.” 2019. *arXiv:1902.07153*.
4. Yang, B et al. “Embedding entities and relations for learning and inference in knowledge bases.” 2014.
5. Hamilton, William L. “Graph representation learning. Morgan & Claypool Publishers.” 2020.
6. Cappelletti et al. “A software resource for large graph processing and analysis.” *Nature Computational Science*, vol. 3, no. 7, 2023, https://doi.org/10.1038/s43588-023-00466-7. Accessed 21 11 2024.
7. Cavalleri, E et al. “An ontology-based knowledge graph for representing interactions involving RNA molecules.” *Sci. Data*, vol. 11, 2024, p. 906.
8. Jason Ansel et al. “PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation.” 2024, https://pytorch.org/assets/pytorch2-2.pdf.