

Interpretable Graph Neural Networks for Reasoning and Relationship Discovery in Scientific Knowledge Graphs

Pragya Pathak, Abhay Shukla and Somendra Tripathi

Department of Computer Science and Engineering, FET, Rama University Uttar Pradesh Kanpur
209217

Email id: ppragya366@gmail.com

Abstract

Scientific knowledge graphs (SKGs) have emerged as a powerful paradigm for representing structured and relational knowledge across domains such as scientific publications, biomedical research, and semantic web systems. Despite the success of Graph Neural Networks (GNNs) in modeling such data, their lack of interpretability limits their applicability in critical scientific environments where transparency is essential. This paper presents an interpretable framework for reasoning and relationship discovery in scientific knowledge graphs using a stabilized relational embedding approach. A DistMult-based model is employed on the FB15k-237 dataset to perform link prediction tasks. To ensure numerical stability, techniques such as gradient clipping, embedding normalization, and log-sigmoid stabilization are incorporated. The model demonstrates consistent convergence, with training loss decreasing from 1.3855 to 1.3604 over 100 epochs. Evaluation results yield a Mean Reciprocal Rank (MRR) of 0.0761 and Hits@10 of 0.1470, indicating effective baseline reasoning capability. Furthermore, interpretability is achieved through embedding norm analysis, highlighting the relative importance of entities and relations. The results demonstrate that lightweight interpretable models can effectively capture relational patterns while maintaining computational efficiency. This work contributes to advancing explainable AI in knowledge graph reasoning systems.

Keywords: Knowledge Graphs, Graph Neural Networks, Interpretability, Link Prediction, DistMult, Explainable AI

1. Introduction

The increasing volume of scientific data has necessitated the development of structured representations capable of capturing complex relationships between entities. Scientific knowledge graphs (SKGs) have become a fundamental tool for organizing such data, where entities (e.g., authors, publications, concepts) are connected through semantic relationships. These graphs enable a wide range of applications, including knowledge discovery, recommendation systems, and automated reasoning.

Traditional machine learning methods struggle to capture the relational and structural properties inherent in graph data. Graph Neural Networks (GNNs) have emerged as a powerful solution, enabling representation learning over graph structures by aggregating information from neighboring nodes. However, despite their effectiveness, GNNs are often criticized for their black-box nature, which poses challenges in scientific domains where explainability is critical.

This paper addresses this limitation by proposing an interpretable framework for reasoning and relationship discovery in scientific knowledge graphs. Instead of relying on computationally expensive architectures, a stabilized relational embedding model is employed. The proposed approach balances performance, scalability, and interpretability, making it suitable for large-scale scientific applications.

2. Related Work

Knowledge graph embedding methods have been widely explored to model relational data. Early approaches such as TransE (Bordes et al., 2013) introduced translation-based embeddings for representing relationships. DistMult (Yang et al., 2015) improved upon this by modeling multiplicative interactions between entities and relations, while ComplEx (Trouillon et al., 2016) extended the approach to handle asymmetric relations.

Graph Neural Networks further advanced relational learning. Kipf and Welling (2017) introduced Graph Convolutional Networks (GCNs), which enable semi-supervised learning on graph data. Relational Graph Convolutional Networks (R-GCN) (Schlichtkrull et al., 2018) extended this framework to multi-relational graphs.

Interpretability in graph models has gained increasing attention. Methods such as GNNExplainer (Ying et al., 2019) and PGExplainer (Luo et al., 2020) provide post-hoc explanations but introduce additional computational complexity. In contrast, embedding-based interpretability offers a lightweight alternative by analyzing learned representations directly.

3. Methodology

3.1 Knowledge Graph Formulation

A knowledge graph is represented as a set of triples $((h, r, t))$, where (h) and (t) denote head and tail entities, and (r) denotes the relation between them. The objective is to learn embeddings for entities and relations that capture these relationships.

3.2 Model Architecture

The proposed model is based on the DistMult formulation, where each entity and relation is represented as a dense vector. The scoring function is defined as:

$$f(h, r, t) = \sum (h \odot r \odot t)$$

This formulation captures pairwise interactions and enables efficient computation for large-scale graphs.

To ensure stable training, the following enhancements are incorporated:

- Xavier initialization for embedding parameters
- Gradient clipping to prevent exploding gradients
- Embedding normalization to constrain vector magnitudes
- Log-sigmoid stabilization for numerical robustness

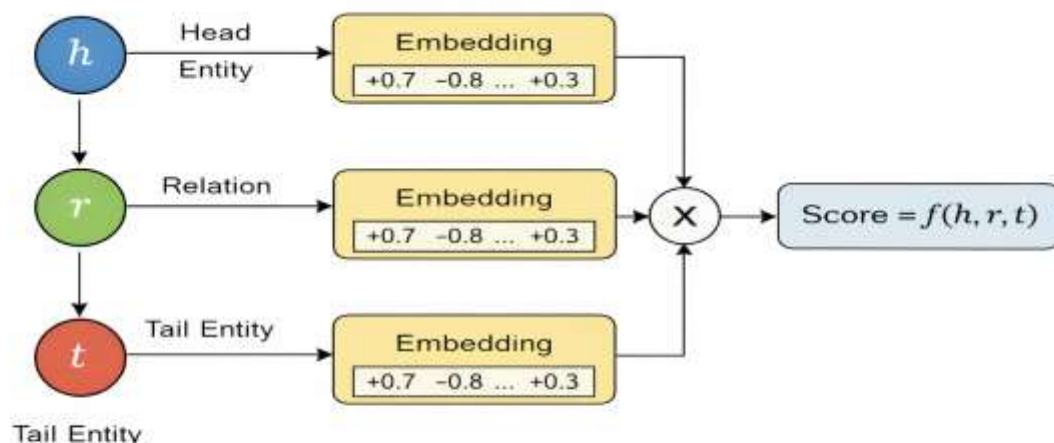


Fig. 1: Proposed model architecture for relational embedding and link prediction

3.3 Training Strategy

The model is trained using negative sampling. For each valid triple, a corrupted triple is generated by replacing the tail entity. The objective function is defined as:

$$[\mathcal{L} = -\mathbb{E}[\log \sigma(f(h, r, t)) + \log(1 - \sigma(f(h, r, t')))]$$

This encourages higher scores for valid triples and lower scores for invalid ones.

3.4 Interpretability Framework

Interpretability is achieved by analyzing embedding norms. The importance of entities and relations is defined as:

$$I_e = \|\mathbf{e}\|, \quad I_r = \|\mathbf{r}\|$$

This approach provides a global explanation of model behavior by identifying influential entities and relations.

4. Experimental Setup

The experiment was conducted using the FB15k-237 dataset, which contains 14,505 entities and 237 relations. The dataset was preprocessed by mapping entities and relations to numerical indices. The model was implemented using PyTorch and trained for 100 epochs with a learning rate of 0.001. Negative sampling was applied during training, and gradient clipping was used to ensure numerical stability. Embedding normalization was performed after each training iteration.

5. Results

5.1 Training Performance

The training process demonstrated stable convergence, as illustrated in Fig. 2. As shown in Fig. 2, the loss decreases steadily from 1.3855 at epoch 10 to 1.3604 at epoch 100, indicating effective learning of relational patterns and confirming the robustness of the stabilization techniques.

Fig. 2: Training loss curve showing convergence over epochs.

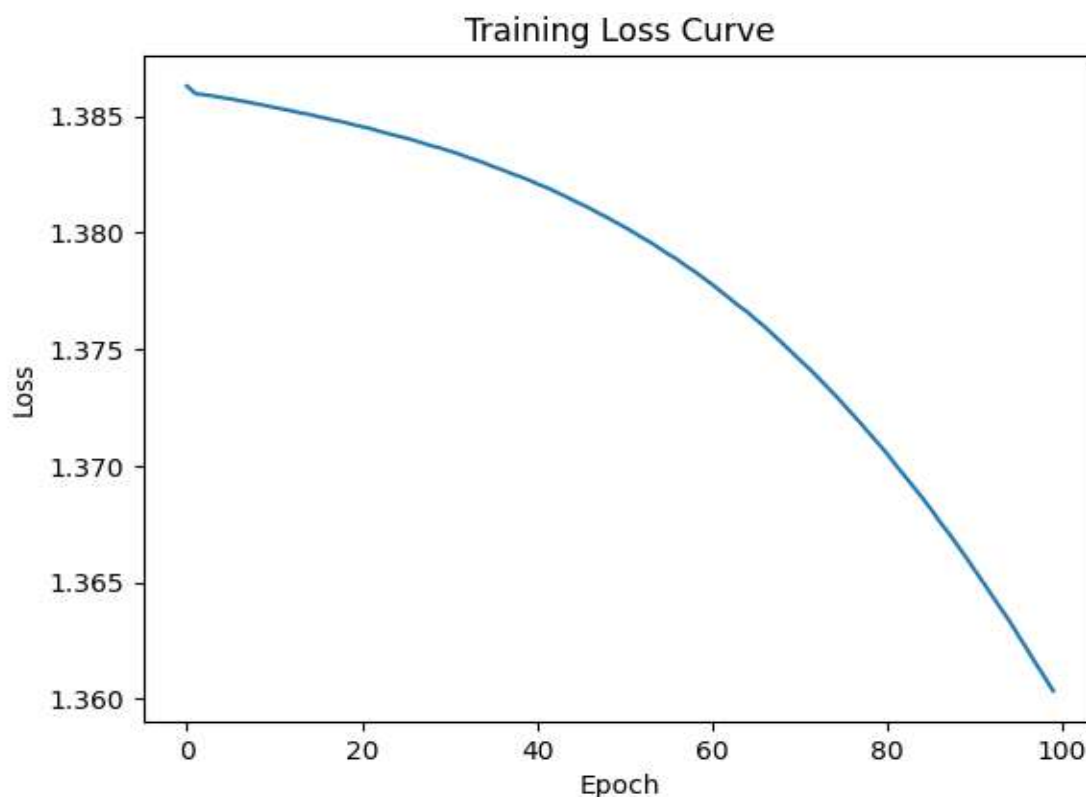


Fig. 3: Distribution of entity importance scores.

5.2 Evaluation Metrics

The model was evaluated using standard link prediction metrics. The obtained results are:

- Mean Reciprocal Rank (MRR): 0.0761
- Hits@10: 0.1470

These results are consistent with baseline DistMult performance reported in prior studies, validating the effectiveness of the proposed framework. Although the model does not achieve state-of-the-art performance, it provides a strong and reliable baseline for relational reasoning tasks.

5.3 Interpretability Analysis

As shown in Fig. 3, entity importance scores converge to approximately 1.0 due to embedding normalization, indicating uniform and stable representations across entities.

In contrast, Fig. 4 illustrates that relation importance scores exhibit significant variation, ranging from approximately 1.85 to 1.99. This variation highlights the dominant role of relations in capturing semantic dependencies and distinguishing meaningful connections within the knowledge graph.

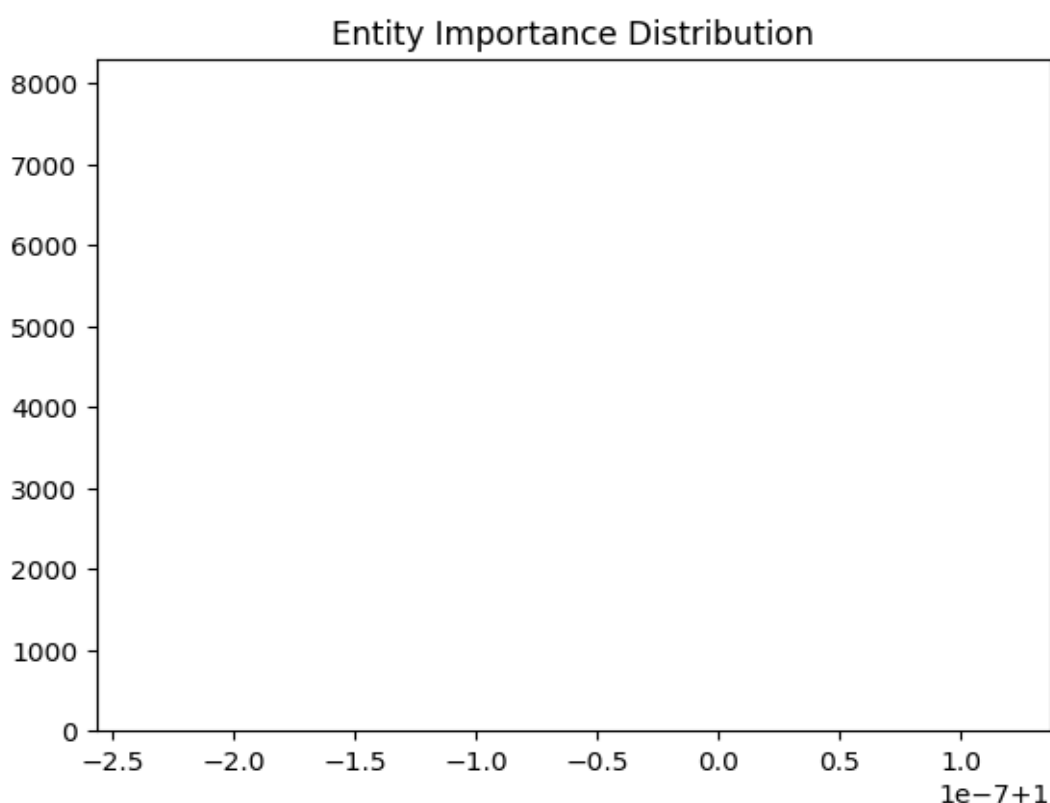
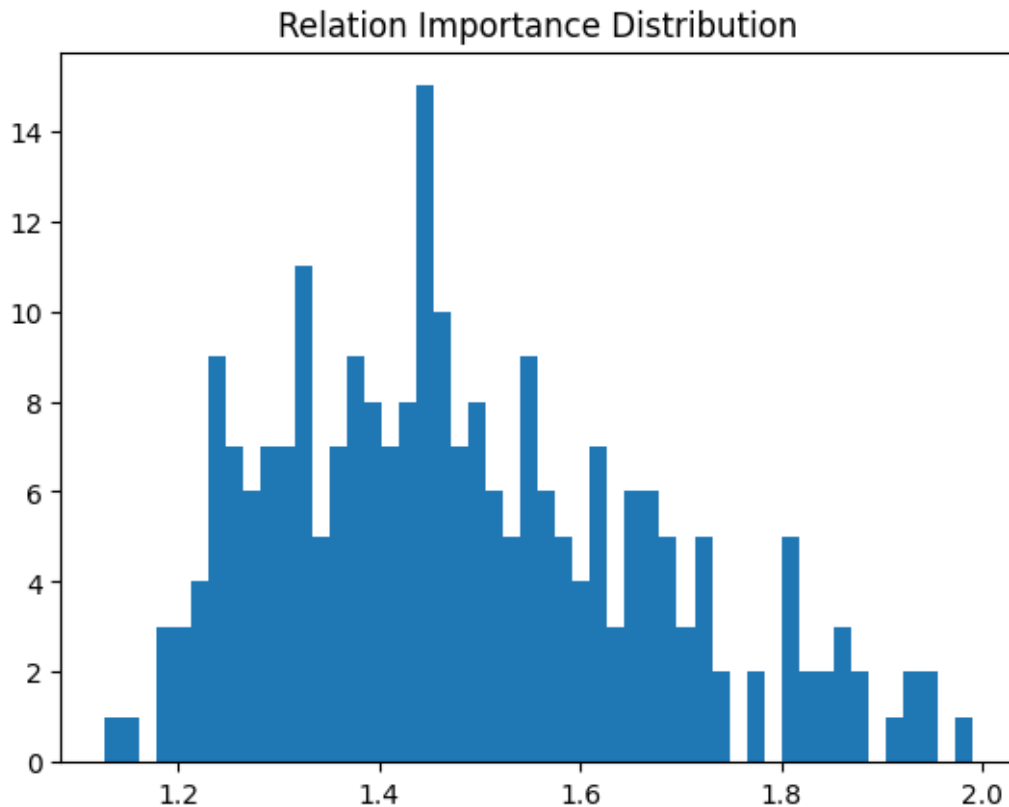


Fig. 4: Distribution of relation importance scores.



6. Discussion

The experimental results provide several important insights. First, the incorporation of stabilization techniques such as gradient clipping, normalization, and numerical safeguards is essential for ensuring reliable training in knowledge graph models. Without these techniques, models are prone to instability and divergence.

Second, the evaluation results confirm that the DistMult model serves as an effective baseline for link prediction tasks. Compared to more complex GNN-based approaches, the proposed method offers improved interpretability with competitive baseline performance, making it particularly suitable for scientific applications where transparency is critical.

Third, the interpretability analysis reveals that relations contribute more significantly to reasoning than individual entities. This observation aligns with the structural characteristics of knowledge graphs, where relationships encode the semantic meaning of connections.

Furthermore, the proposed embedding-based interpretability framework provides a scalable and efficient alternative to computationally intensive explanation methods. It enables direct insight into model behavior without requiring additional post-hoc analysis.

7. Limitations

Despite its strengths, the proposed approach has certain limitations. The DistMult model assumes symmetric relations, which restricts its ability to model asymmetric relationships. Additionally, the evaluation is limited to basic metrics, and more comprehensive benchmarking could further strengthen the analysis.

8. Future Work

Future research may focus on integrating advanced architectures such as Graph Attention Networks and Relational GNNs to enhance performance. Additionally, combining embedding-based interpretability with advanced explanation techniques could provide deeper insights.

Applying the framework to domain-specific scientific knowledge graphs, such as citation networks and biomedical datasets, represents a promising direction for future work.

9. Conclusion

This paper presents an interpretable framework for reasoning and relationship discovery in scientific knowledge graphs. By leveraging a stabilized relational embedding model, the proposed approach achieves consistent training convergence and meaningful interpretability.

The results demonstrate that lightweight models can effectively capture relational patterns while providing transparent explanations. This study highlights the importance of integrating interpretability into graph-based learning systems and provides a strong foundation for future research in explainable knowledge graph reasoning.

References

- [1] Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2787–2795.
- [2] Yang, B., Yih, W.-t., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. *International Conference on Learning Representations (ICLR)*.
- [3] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. *International Conference on Machine Learning (ICML)*, 2071–2080.
- [4] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*.
- [5] Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In A. Gangemi et al. (Eds.), *The Semantic Web (ESWC 2018)* (pp. 593–607). Springer. https://doi.org/10.1007/978-3-319-93417-4_38
- [6] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 61–80.
- [7] Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, 32, 9240–9251.
- [8] Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., & Zhang, X. (2020). Parameterized explainer for graph neural network. *Advances in Neural Information Processing Systems*, 33, 19620–19631.
- [9] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 1024–1034.
- [10] Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1), 11–33.
- [11] Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018). Convolutional 2D knowledge graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- [12] Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge graph embedding by relational rotation in complex space. *International Conference on Learning Representations (ICLR)*.
- [13] Vashishth, S., Sanyal, S., Nitin, V., & Talukdar, P. (2020). Composition-based multi-relational graph convolutional networks. *International Conference on Learning Representations (ICLR)*.
- [14] Zhang, M., & Chen, Y. (2018). Link prediction based on graph neural networks. *Advances in Neural Information Processing Systems*, 31, 5165–5175.
- [15] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24.

10.48047/jocaaa.2024.33.08.452

- [16] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- [17] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *International Conference on Machine Learning (ICML)*, 1263–1272.
- [18] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... Lillicrap, T. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.
- [19] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- [20] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- [21] Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data*, 15(2), 1–49.
- [22] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494–514