Efficient Graph Representation Framework for Chemical Molecule Similarity Tasks

CS4991 Capstone Report, 2024

Jiaji Ma
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
yjk8jd@virginia.edu

ABSTRACT

Graph machine learning faces the challenge of efficiently representing graphs, as most machine learning algorithms require vector-encoded data. To address this, I propose a novel two-stage framework that combines Graph Isomorphism Networks (GINs) with Siamese autoencoders for the representation of chemical molecule graphs. The first stage involves constructing and training a GIN model using node and edge attributes of chemical graphs to generate high-dimensional Graph Embeddings that capture structural characteristics and predict molecular properties. In the second stage, these embeddings are optimized through a Siamese autoencoder, which reduces their dimensionality while preserving structural information, facilitating tasks such as approximate nearest neighbor search. This framework demonstrates effectiveness in accurately predicting molecular similarity and preserving graph structure in low-dimensional embeddings. Future work will focus on refining this framework to enhance its efficiency and accuracy in representing chemical molecule graphs, with potential applications in computational chemistry and drug discovery, including further testing and evaluation to address any identified limitations.

1. INTRODUCTION

Graphs have proven to be highly effective structures for representing and analyzing real-world data in numerous domains, such as social networks and biological networks. The effectiveness of graphs in these domains can be attributed to their ability to encode both structural and semantic information. By representing entities as nodes and relationships as edges, graphs provide a powerful framework for capturing the underlying patterns, dependencies, and inter-dependencies within complex systems. This flexibility and ability to capture and model arbitrary relationships between arbitrary entities allow graphs to go beyond the limitations of traditional data structures, providing a more comprehensive and holistic understanding of complex systems and real-world data.

2. RELATED WORKS

The integration of graph-based methodologies with machine learning techniques for graph representation learning has gained significant attention (Bengio, et al., 2013). There have been methods developed for graph isomorphism testing without mapping functions (Shervashidze, et al., 2011) and thus cannot be applied for general similarity learning. For similarity learning models to be useful for querying in downstream tasks, the graphs must be encoded into vector representations. Graph embeddings enable the transformation of graph data into vector

representations, thereby bridging the gap between graph structures and traditional vector-based machine learning models (Wu, et al., 2020) (Cai, et al., 2018). Using graph embeddings, a wide array of established machine learning techniques can be employed to tackle diverse tasks, extending beyond the limited applications of graph data in its original form.

This enhanced flexibility allows for more comprehensive analysis, prediction, and decision-making in real-world scenarios. With the emergence of deep learning on graph data, Graph Neural Networks (GNNs) have become a powerful tool to encode graphs into embedding vectors (Hamilton, et al., 2017) (Velickovic, et al., 2017) (Xu, et al., 2018). Compared to traditional graph embedding methods, GNNs address tasks in an end-to-end manner (Ma, et al., 2021) and can better leverage graph feature for specific learning tasks. GNN models have been proposed to solve problems in multiple domains, such as brain networks (Ma, et al., 2019) and computer security (Li, et al., 2019).

In the field of biomedicine, due to the nature of the graph data as chemical molecules, there have been cheminformatics tools for mapping the chemical space long before machine learning, called molecular fingerprinting. Specifically, Morgan fingerprinting (Morgan, 1965) is one of the most widely used featurization methods for chemical molecules. The algorithm iteratively encodes circular substructures of a molecule as identifiers, hashes them, and folds them to bit positions to generate a bit string. Since fingerprinting methods are optimized specifically for chemical molecules, fingerprinting has achieved good performance when used as input representations in deep neural networks (Unterthiner, et al., 2014), but in many cases fingerprinting methods are not able to offer ideal performance due to the length of the resulting embedding vectors.

3. PROJECT DESIGN

The overall pipeline in the proposed framework consists of two stages. In the first stage, the structural information of molecule graphs is utilized to train a Graph Isomorphism Network (GIN) model for predicting individual graph attributes. Then the embeddings of the molecules for each graph attribute are calculated. In the second stage, the output vectors from the first stage are used to train Siamese autoencoders, which preserves the information in the vectors and the relative similarity between vectors. Then the corresponding low-dimensional Graph Embedding vectors are obtained from the high-dimensional embedding vectors from the previous stage.

Each chemical molecule can be represented as an undirected weighted graph denoted as G = (V, E). The set of nodes in the graph, V, corresponds to the atoms present in the molecule. The set of edges, E, represents the bonds between the atoms, capturing the connectivity information of the molecular structure. The graph includes intrinsic properties associated with the molecule itself in nodes, edges, and the overall graph, providing valuable data for analysis. Specifically, each node (atom) in the graph is associated with 11 attributes that characterize various atomic properties. Similarly, each edge (bond) carries 4 attributes that describe bond-specific properties. At the graph level, 19 attributes are provided that represent different molecular properties. These descriptors encompass a diverse set of features and provide relevant information that captures global aspects of the chemical molecules.

The proposed GIN model has the following stages: (1) Node embedding, which encodes

the features and structural properties of each node into a vector; (2) Graph embedding, during which graph level representations are obtained from each set of node embeddings using global pooling, then graph level embeddings from each level of abstraction is concatenated to form a single high-dimensional embedding vector; and (3) the graph attribute computation stage, which reduces the high-dimensional vectors into one final attribute value, which is compared with the ground-truth value to update model parameters.

After obtaining the complete, high-dimensional Graph Embedding vectors from the GIN model, optimizing them to be more suitable to use for downstream tasks is an important issue. Since superficially high-dimensional and complex phenomena can be dominated by a small number of simple variables in most situations, this can be done using some learned projection method that maps the vectors in high-dimensional feature space to low-dimensional feature space. Siamese autoencoders were used as the dimensionality reduction technique, to maximize preservation of information in the original high-dimensional embedding vectors while reducing the dimensionality of the vectors, which optimizes them for tasks requiring vector input.

4. ANTICIPATED RESULTS

The baseline approach will be molecular fingerprinting, in which a kernel is applied that extracts feature from the molecule. The features are hashed and then used to calculate a bit vector. Specifically, one of the most widely used methods, the Morgan fingerprint, will be used. It is optimized to compare the similarity between molecules, by considering the neighborhood of each atom and perceives the presence of specific circular substructures around each atom in a molecule, which are predictive of the biological activities. It is one of the best performing molecular fingerprinting methods for target prediction tasks. By default, it produces vectors of length 2048.

For both the baseline method and after each stage of the framework, three different benchmarks will be run to evaluate different aspects of the performance of the models. First, the Graph Embedding vectors will be used as input to a simple Forward Feeding Neural Network that is trained to predict the original graph attribute; this measures how well the structural information (which was used to obtain the embeddings) is preserved. Second, Uniform Manifold Approximation and Projection (UMAP) will be used to project the Graph Embedding vectors onto a 2-dimensional space. This allows for examination of the degree to which the distribution and vector distances reflect the ground truth value for attribute similarities between the graphs. Finally, the Graph Embedding Vectors will be used as input to perform Approximate Nearest Neighbor Search. This measures the downstream task performance from multiple aspects for each set of Graph Embedding vectors.

5. CONCLUSION

This study contributes to existing knowledge about graph representation and provides valuable insights into the complex nature of chemical molecule data and proposes an efficient approach to obtain Graph Embeddings in vector format optimized for chemical molecule similarity tasks. Compared to traditional methods in the chemistry field, such as molecular fingerprinting, my approach can greatly reduce the dimensionality of the embeddings, making them more computationally efficient. This also helps to improve the interpretability and scalability of the embeddings for downstream tasks.

6. FUTURE WORK

The current study focuses on static molecule structure graphs. Future work can benefit from spatial-temporal chemical molecule data that better reflect how molecules transform over time. For more realistic molecule behavior, molecule isomerization can be modeled using dynamic graphs in the form of time series data. Moreover, this framework is limited to individual graph attributes, multi-tasking models can be explored in the future. The proposed framework and techniques can also be explored in various other domains that require analysis and prediction of complex graph-structured data.

REFERENCES

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798-1828.

Cai, H., Zheng, V. W., & Chang, K. C. C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE transactions on knowledge and data engineering*, 30(9), 1616-1637.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Li, Y., Gu, C., Dullien, T., Vinyals, O., & Kohli, P. (2019). Graph matching networks for learning the similarity of graph structured objects. In *International conference on machine learning* (pp. 3835-3845). PMLR.

Ma, G., Ahmed, N. K., Willke, T. L., Sengupta, D., Cole, M. W., Turk-Browne, N. B., & Yu, P. S. (2019). Deep graph similarity learning for brain data analysis. In Proceedings of the 28th ACM *International Conference on Information and Knowledge Management* (pp. 2743-2751).

Ma, G., Ahmed, N. K., Willke, T. L., & Yu, P. S. (2021). Deep graph similarity learning: A survey. *Data Mining and Knowledge Discovery*, 35, 688-725.

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2), 107-113.

Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K., & Borgwardt, K. M. (2011). Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9).

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., & Hochreiter, S. (2014). Deep learning as an opportunity in virtual screening. In *Proceedings of the deep learning workshop at NIPS* (Vol. 27, pp. 1-9).

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *stat*, 1050(20), 10-48550.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1), 4-24.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K. I., & Jegelka, S. (2018, July). Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning* (pp. 5453-5462). PMLR.