그래프 신경망 기반 가변 자동 인코더로 분자 생성에 관한 연구

에드워드 카야디[1], 송미화[1],
[1]세명대학교 스마트IT학부
email: edw,chydi@gmail.com, mhsong@semyung.ac.kr

A study on Generating Molecules with Variational Auto-encoders based on Graph Neural Networks

Edward Dwijayanto Cahyadi[1], Mi-Hwa Song[1]
[1]School of Smart IT, Semyung University

요 약

Extracting informative representation of molecules using graph neural networks(GNNs) is crucial in AI-driven drug discovery. Recently, the graph research community has been trying to replicate the success of self supervised in natural language processing, with several successes claimed. However, we find the benefit brought by self-supervised learning on applying varitional auto-encoders can be potentially effective on molecular data.

1. Introduction

Due to their effectiveness, Graph Neural Networks(GNNs) have been adopted to model a wide range of structured data, such as social networks and road graphs. Among those applications, molecule modelling is probably one of the most important, as it is the foundation of the biomedical area. However, since biomedical labelling is usually time-consuming and expensive, task-specific labels are highly inadequate in this domain, posing a significant challenge to this field. They were recently inspired by the remarkable success of the self-supervised pre-train-finetune paradigm to molecule modelling with GNN, hoping to boost the performance of various molecular tasks by pre-training the model on the enormous unlabeled data.

2. Related Research

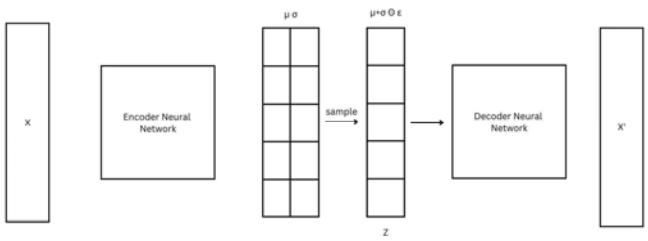
We have surveyed how to apply self-supervised learning on graph neural networks. We found three possible methods. The first one is auto-encoder approaches like based on reconstructing the input. With this method, we can reconstruct the graphs adjacency matrix or parts of the node feature matrix. The second method is to do multiple pre-calculated tasks based on the graph representation (ex., node degrees, substructures, path length). The task of the model is to predict those descriptors as precisely as possible. Moreover, the third one is contrastive approaches. This self-supervised framework pushes the representations of similar data points together and pulls different data points apart from each other.

3. Research method

This paper will explain how to apply the self-supervised learning GNN (variational graph autoencoder) method to generate new molecules. This method was proposed in a paper called variational graph auto-encoders. Generally, the architecture consists of an encoder and decoder, which are both neural networks. In the middle of

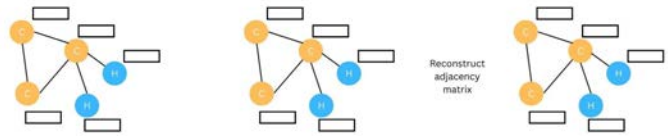
the encoder, There is a bottleneck in the form of a latent vector. This bottleneck forces the model to compress the input representation into the vector to recover the input from the compressed representation, So the model learns how to compress information into something from which it can recover.

The latent representation is a single deterministic vector like a point estimate. Variational autoencoders add some variation to the vector. That is, each of the z vectors should follow a distribution instead of being a single number. This allows us to sample from the distribution to generate new data, so the latent representation is now space and not a single point anymore. Oppositely in plane autoencoders, the latent vector is not organized, which means we have no control over how the model compresses the information. This is a problem because if we want to decode random data points, we cannot ensure that the points make sense or will be decoded into something helpful; therefore, in the classical variational autoencoder, each of the values follows a normal distribution with a specific mean μ and a variance σ.



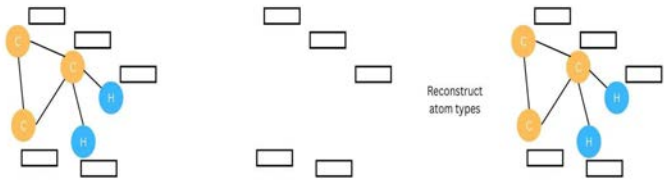
(Figure 1) Variational autoencoders architecture.

Then we proceed to get the μ and σ. All it needs to do is output the μ and σ by the encoder. With this, we can build normal distributions and then sample from them. It gives us a single vector just like before, but now it comes from each distribution. If the latent representation would only consist of two values, the latent space would have a dimension of two. First, we need to reconstruct the adjacency matrix to apply this concept to the graph. This even means we must reconstruct the bond types in the latent representation for molecules. We can use the node embeddings and the atom types to recover the adjacency information.



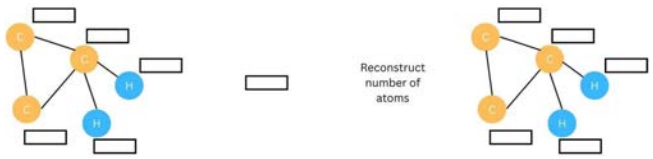
(그림 2) Reconstruct the adjacency matrix

In the decoder, the goal is that the model should be able to sample the atom types automatically. So the second type is that we need to reconstruct the atom types of each node in the molecule.



(그림 3) Reconstruct the atom types

Next, we need to specify how many atoms it has in the output molecule. So the last one is to reconstruct the number of atoms.



(그림 4) Reconstruct the number of atoms

4. Conclusion

Graph Neural network is a powerful deep learning method for processing graph data. By doing self-supervised learning on GNN with a graph variational auto-encoder, it can potentially generate molecules based on the data set. There are three main steps to generate the result. First, reconstruct the adjacency matrix. The second is to reconstruct atom types, and the third is to reconstruct the number of atoms. However, we have explained the process but are still unable to

experiment because of the lack of resources. We hope that in the future, we can get the data and visualize this idea by doing an experimental project.

Reference

[1] Yuning You "Graph Contrastive Learning with Augmentation" arXiv, 2020

[2] Ruoxi Sun "Does GNN pretraining help molecular representation" arXiv, 2022

[3] Yaochen Xie "Self-Supervised Learning of Graph Neural Networks: A Unified Review" arXiv, 2022

[4] Thomas N. Kipf "Variational Graph Auto-Encoders" Cornell University, 2022

[5] Fan Yun Sun "InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization" arXiv, 2019

[6] Petar Veličković "Deep Graph Infomax" arXiv, 2018