Graph-based Neural Networks for Retrosynthesis Planning: A Comprehensive Review and Comparative Analysis

Xun Zhou
Department of Computer Science
Shanghai Jiao Tong University
X21
chosen@sjtu.edu.cn

Abstract

In this project, we implemented LocalRetro method for single-step retrosynthesis prediction, and an MLP based neural network for Molecule evaluation. LocalRetro demonstrates promising results, but MLP performs poorly in our experiments.

1 Introduction

The pursuit of designing molecules and materials with desired properties is a crucial goal in chemistry and materials science. Cheminformatics leverages data to uncover the intricate relationship between molecular structures and their properties, enabling the discovery of novel functional molecules. Machine learning has revolutionized this field by rapidly predicting molecular properties based on their structures and even reverse-engineering molecules from desired functionalities. However, synthesis planning remains a crucial step to bridge the gap between theoretical designs and practical implementation. The retrosynthesis problem inherently poses challenges as it involves a one-to-many mapping, which means that there can be multiple possible reaction pathways to synthesize a target compound. This aspect adds complexity and makes the task more difficult.

2 Related Works

GNNs have emerged as a powerful tool for solving this problem by leveraging the graph structure of molecules. In a graph representation of a molecule, atoms are represented as nodes, and bonds between atoms are represented as edges connecting the nodes. GNNs can effectively capture the structural information and relationships between atoms in a molecule graph.

One popular GNN model used in Retrosynthetic Reaction Prediction is the Graph Convolutional Network (GCN). The GCN applies convolutional operations on the graph to propagate and update information across nodes. The key idea behind GCN is to aggregate and combine information from neighboring nodes to compute new node representations. The propagation rule can be formulated as follows:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)}\right) \tag{1}$$

In the above formula, $h_i^{(l+1)}$ represents the updated representation of node i at layer $l+1, h_j^{(l)}$ represents the representation of neighboring node j at layer $l, \mathcal{N}(i)$ represents the neighbors of node $i, W^{(l)}$ represents the weight matrix at layer l, c_{ij} is a normalization constant, and σ is a non-linear activation function.

To predict retrosynthetic reactions, GNN models are trained on a large dataset of known reactions. The model learns to encode the structure of the reactant molecules and predict the product molecules. During inference, the GNN model takes a target molecule as input and predicts a set of precursor molecules along with the corresponding reaction steps.

By effectively capturing the graph structure of molecules and learning from a large number of known reactions, GNN-based models have shown promising results in Retrosynthetic Reaction Prediction. They have the potential to accelerate the process of drug discovery and synthesis planning by providing guidance on feasible reaction pathways for target molecules.

Please note that the formula provided above represents a simplified version of the propagation rule used in GCN. Different variations of GNNs may have slightly different formulations, but the underlying idea of aggregating information from neighboring nodes remains consistent.

3 Methods

3.1 Task 1: LocalRetro

We employ the LocalRetro algorithm to solve the single-step retrosynthesis prediction problem. We leverage the fact that reactant molecules retain a significant portion of their substructures and fragments during and after a reaction. This allows us to concentrate on the changes in the molecular structure, specifically the alterations in atoms or bonds, to facilitate retrosynthesis. Rather than starting from scratch to identify suitable reactants, our focus is on deducing the specific local changes that took place to generate the given product. These changes can involve the formation or breaking of bonds and/or the addition or removal of atoms. By analyzing and understanding these structural transformations, we can effectively navigate the retrosynthetic process.

We leverage the domain-specific knowledge extracted by the original authors, such as reaction templates, to classify reactions. This approach significantly improves the classification performance by reducing the complexity of the task. Instead of dealing with thousands of distinct templates, we can now categorize reactions into 657 classes, thanks to the utilization of these extracted templates.

Next, the new model concentrate more on local environments, predict the correct local reaction template at each atom and bond that leads to the given product. We represent the molecule as a heterogeneous graph $G = (V, E)$ with V(vertices) denoting atoms and E(edges) denoting bonds. To encode the surrounding environmental information for each atom,a message passing neural network (MPNN) is used to described to update each atom feature. We denote the message passing function by $MPNN(\cdot)$, with the atomic features of atom a as v_a, the atomic features of its neighboring atom b as v_b, and the features of their connecting bond as e_{ab}. The atomic features of atom a are updated by the MPNN via:

$$v'_a = MPNN(v_a, \{v_b\}, \{e_{ab}\}) \tag{2}$$

$\{\}$ denotes a set of neighboring atoms around a given atom. After the atomic features are updated, the bond feature is represented by concatenating two atomic features $(v'_a||v'_b)$ and goes through a fully connected layer.

$$e'_{ab} = w(v'_a||v'_b) + c \tag{3}$$

Where w is the weights and c is the bias of the fully connected layer.

To incorporate the global dependency and nonlocality of reactions, the model employ a global attention mechanism to update the atomic features, denoted as x_a, and bond features, denoted as x_{ab}. The update process can be expressed as:

$$x'_a = GRA(x_a, \{v'_a\}, \{e'_{ab}\}) \tag{4}$$

$$x'_{ab} = GRA(e'_{ab}, \{v'_a\}_{a \in V}, \{e'_{ab}\}_{ab \in E}) \tag{5}$$

where GRA($\cdot$) represents the application of the global attention mechanism using the current atomic features x_a, the set of all updated atomic features is $\{v'_a\}$, and the set of all updated bond features is $\{e'_{ab}\}$.

Next comes the typical step of training classifiers using the extracted features.Separate classifiers for atom reaction templates and bond reaction templates are used to compute their respective outputs. These classifiers utilize the updated representations of atoms and bonds.Given the nature of the problem, we used cross-entropy loss to measure the performance of the model.

3.2 Task 2: MLP

We utilize a multilayer perceptron (MLP) to learn molecule evaluation in a supervised manner. Due to the significant difference in the value distribution between the training and test domains, I attempted to apply z-score normalization to both the values in their respective domains. Subsequently, I fed the normalized data into an MLP for supervised learning.

We used the unpacked Morgan FingerPrint as the input and apply layer normalization at each layer of the MLP. The ReLU activation function is used. We also utilize the Mean Squared Error (MSE) loss function.

Table 1: The training and test datasets exhibit different data distributions.

Model	min	max	mean	std
Train set	0	44.67	0.46	1.27
Test set	0	45.2	1.93	3.1

4 Expriment and Discuss

4.1 Task 1

Given the limited time available, We trained the model for only 3 epochs. Despite the incomplete convergence, the model has shown impressive performance.

Table 2: predict accuracy in task 1

k	1	3	5	10	50
accuracy	0.345	0.516	0.566	0.6090	0.627

In this model, we incorporate more prior knowledge by emphasizing the correlation between chemical properties and neighboring atoms. We preserve more connections between neighboring atoms and chemical bonds to capture local dependencies. Additionally, we leverage attention mechanisms to prevent important long-range correlations from being overlooked. We believe that this is one of the key factors contributing to the success of this model compared to other graph-based models. By effectively incorporating prior knowledge and leveraging attention, we can capture both local and global information, leading to improved performance in handling chemical structures and properties.

4.2 Task 2

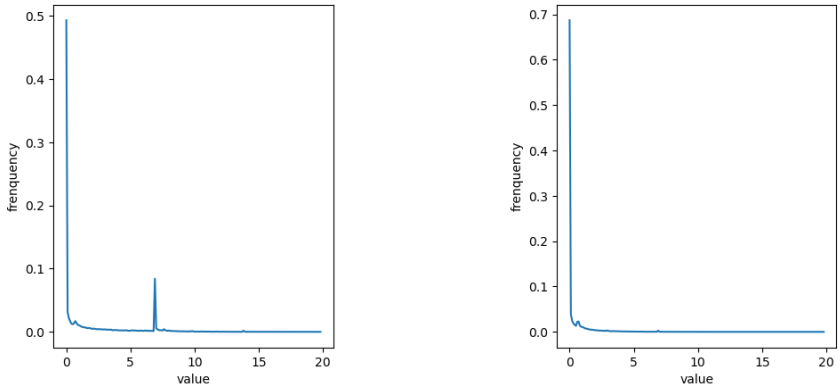

Figure 1: The training(right) and testing(left) datasets exhibit different data distributions.

4.3 Training and evaluation

The performance of the model used in Task 2 is unsatisfactory, as there is minimal improvement in the loss on the test set. We believe there could be two possible reasons for this.

Firstly, there is a significant disparity in the data distribution between the training and test sets. The training set is more concentrated around zero, while the test set exhibits higher variance, with many data points distributed around seven. To improve the performance, it may be necessary to explore more transfer learning techniques or expand the dataset to better capture the variability in the test set.

Secondly, the Morgan FingerPrint used as input may be too sparse for effective processing by the multilayer perceptron (MLP). MLPs may not be well-suited for handling such sparse data. Considering alternative methods like graph neural networks (GNNs) may yield better results, as GNNs are specifically designed to process graph-structured data and can better capture the structural information inherent in the Morgan FingerPrint.

5 Contribution

Xun Zhou 517082910012 100%

code: https://github.com/chosen66/Retrosynthetic-Planning