Final Project Proposal

Jihun Jeung (20211155)

jihun@gm.gist.ac.kr

Life Data Mining Lab, School of Life Science, GIST

Motivation for GNN

Graph Neural Network (GNN) is a neural model that leverages the structure and properties of graphs. Compared to convolution network (CNN) which requires the rigid grid, GNN can be applied to flexible graph structure. A typical example of GNN application is molecular structure. A molecule can be projected as a graph, whose nodes are their atom and edges are their bonds. AlphaFold2 could increase its performance by adopting GNN in *Evoformer* block. [1] Another promising result is antibiotic discovery in 2020. A directed message passing neural network that is trained with molecular properties from publicly opened database predicts Halicin as an effective antibiotics.[2] GNN is a promising model in cheminformatics and bioinformatics.

Molecular properties such as toxicity in biological systems are mainly determined by 3-dimensional conformation. Free energy is the key quantity to describe molecular dynamics and conformation. Although calculating free energy is always a crucial issue in chemistry and biology, molecular dynamics (MD) simulation, which calculates the lowest free energy conformation and predicts its 3-dimensional structure, requires heavy computational cost. Can GNN become a breakthrough for structural chemistry and biology?

Problem Statement

The QM9 dataset from the “MoleculeNet: A Benchmark for Molecular Machine Learning” paper contains about 130,000 molecules with 19 targets, such as dipole moment, HOMO, LUMO, internal energy, free energy, and so on. Each molecule includes complete 3-dimensional spatial information for the single low energy conformation. [3]

Each molecule can be interpreted as into a graph whose node is an atom and whose edge is a bond between atoms. Each atom will be embedded into 11-dimensional vector: 5-dimensional one hot encoding (hydrogen, carbon, nitrogen, oxygen, fluorine) and 5-dimensional chemical properties (atomic number, aromatic, sp, sp2, sp3, number of hydrogen atoms). The target label is free energy.

[1] Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[2] https://www.cell.com/cell/pdf/S0092-8674(20)30102-1.pdf

[3] https://pytorch-geometric.readthedocs.io/en/latest/modules/datasets.html#torch_geometric.datasets.QM9

Method

PyTorch Geometry (PyG) [4]is a library built upon PyTorch to train GNN for a wide range of applications. The dataset would be split into training set (80%), validation set (10%), test set (10%). Adam or SGD will be used as an optimizer (the optimizer with best performance will be selected). Mean squared error (MSELOSS) will be used as loss function. The model will make graph-level prediction using all the node embedding in the graph.

Dataset --> node embedding --> [GNN layer] X N --> pooling --> graph-level prediction

The type of GNN layer and the number of GNN layers will be selected among GCNConv, SAGEConv, and so on in terms of performance. To reduce overfitting, dropout and skip connection will be considered. For graph-level prediction, global mean pooling, global max pooling, global sum pooling will be compared in order to select the best one.

[4] https://github.com/pyg-team/pytorch_geometric