THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC

Data Science Program

Capstone Report - Spring 2022

Molecular Property Prediction

Jongchan Kim

supervised by
Amir Jafari

Abstract

In this paper, GNN has been introduced with self-attention mechanism to solve drug candidate screening problems which could save a huge amount of money and time. Specifically, the model is trained to predict whether a drug molecule can inhibit HIV virus replication or not. This model has showed the same performance compared to the model without GNN with a smaller number of parameters, which could be possibly useful as a efficient tool for the screening which needs to be proved in the future study.

Introduction

Recent applications of deep learning in drug discovery have shown a promising future. For example, prioritizing drug candidates with machine learning techniques can help save a huge amount of time and money. As the amount of available molecular data increases over time, developing efficient and effective ways of mining molecular data becomes a major problem for drug discovery. Deep learning(DL) has been known for good performance with an increased amount of data and computing power. To take advantage of it, there have been many deep learning applications in drug discovery including compound property and activity prediction. For example, Mayr *et al* [1] won the Tox21 challenge on a dataset of 12,000 compounds for 12 toxicity assays with multitask DNN model.

Graph is a set of nodes and edges. A molecule can be represented as a graph where nodes are atoms, and the edges are chemical bonds. Graph neural network(GNN) is a type of neural model that capture the dependence of graphs via message passing between the nodes of graphs aiming to learn a parametric mapping function that embeds nodes, subgraphs, or the entire graph into low-dimensional continuous vector spaces. [2] A embedding of each node is updated by messages that are created from embeddings of neighboring nodes. GNN has been considered an attractive modeling way for molecular property prediction. [3]

While GNN has been gaining attention from the drug discovery industry, the Attention mechanism has been getting popular in Natural-language-processing(NLP) domain. Self-Attention is an mechanism relating different positions of a single sequence in order to compute a representation of the sequence [4] and has become one of the most important concepts in the DL field, especially in natural language processing(NLP). This attention mechanism was also introduced with GNNs achieving state-of-the-art accuracies on benchmark datasets for graph classification.[5]

In this paper, GNN on top of self-attention encoder layers has been applied to one of the popular public molecular property prediction datasets called Ogbg-Molhiv to predict whether each molecule inhibits HIV replication.

Method

(1) Dataset

The ogbg-molhiv dataset contains 41,127 graphs. Each graph represents a molecule. Input nodes features are 9-dimensional, containing atomic number and chirality, as well as other additional features such as formal charge and whether the atom is in the ring. Input edge features are 3-dimensional, containing bond type, bond stereochemistry as well as an additional bond feature indicating whether the bond is conjugated. [6] Task type is a binary classification of whether each molecule inhibits HIV virus replication or not. As the dataset is unbalanced, ROC-AUC was used to evaluate the model's performance. The dataset has been already split into train/val/test sets by a method called scaffold split which is based on the scaffold of the molecules so that the train/val/test set is more structurally different.

(2) Network

First, every node feature and edge feature are embedded and used as an input for the encoder layer of the self-attention model. As each molecule could have a different number of atoms, zero-padding was used to handle that situation. The main difference between the common encoder layer of the NLP model and this model is that this model can utilize the edge information by adding an embedded edge matrix to the result of multiplication between queries and keys of a molecule before the softmax function. As a result, compared to a common NLP embedding layer where it extracts the implicit relationship between words, the embedding layer of this model is not only learning the implicit relationship between atoms but also can explicitly use additional information on actual relationships between atoms. [7][Fig.1]

On top of the embedding layer, GNN was selectively used to compare the performance. GNNs use an aggregation function to update the vector representation of each node by transforming and aggregating the vector representations of its neighbors. In this paper, the basic sum function was used to create a message using vectors of neighbors. Then the sum of the neighbor's vectors and its own vector was averaged. This whole process has been implemented through multiplication with an adjacency matrix. [Fig. 2] In the end, a linear output layer was used to predict whether each molecule inhibits HIV virus replication.

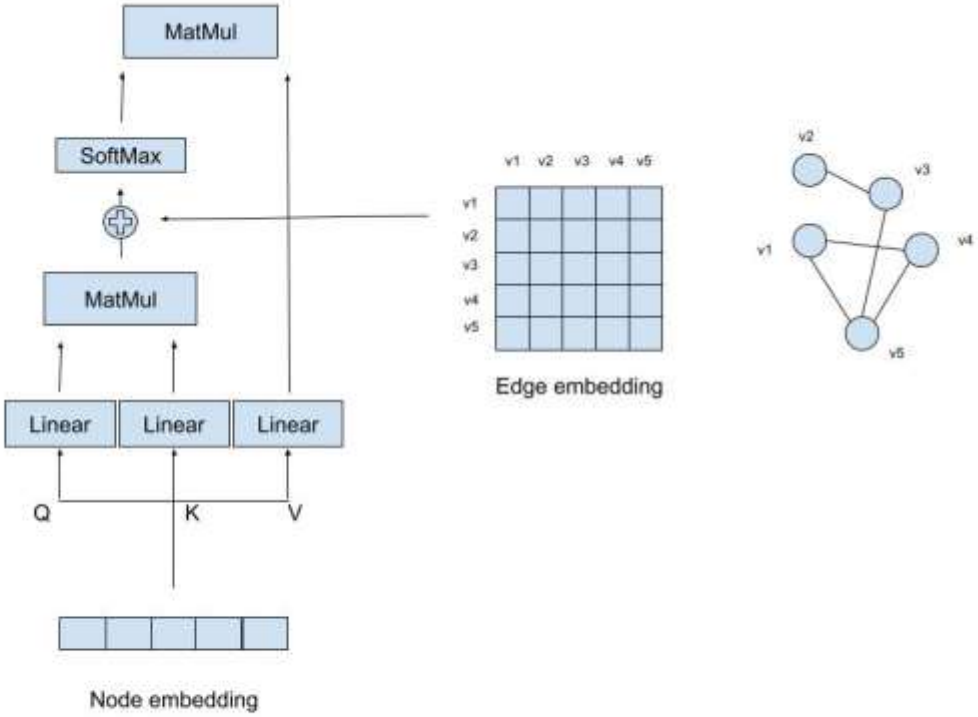


Fig.1 Graphomer

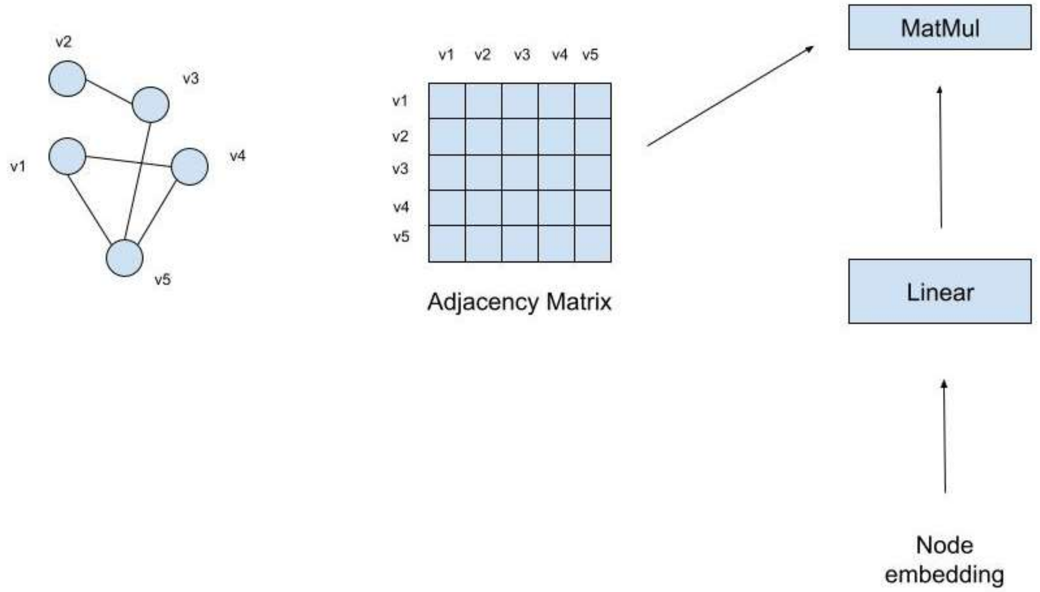


Fig. 2. GNN layer

(3) Training

A deep learning model with 2 self-attention encoding layers was used as a baseline model to be compared with 1 self-attention encoding layer plus the GNN layer. [Fig. 3] Models were trained until validation loss is no longer decreasing.

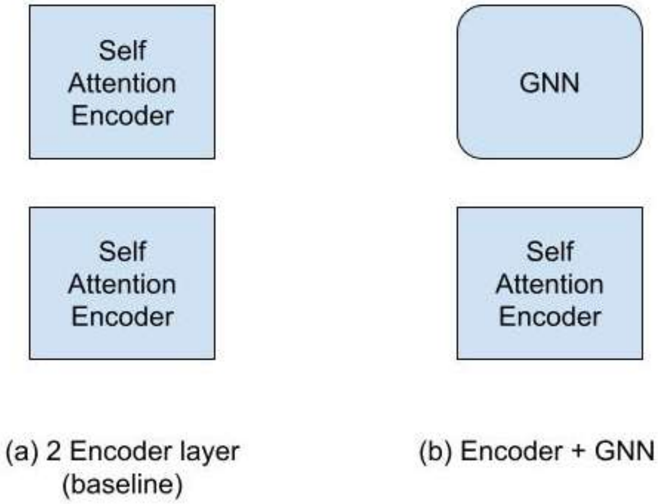


Fig. 3. 2 self-attention encoder layers as a baseline and 1 self-attention encoder layer plus GNN

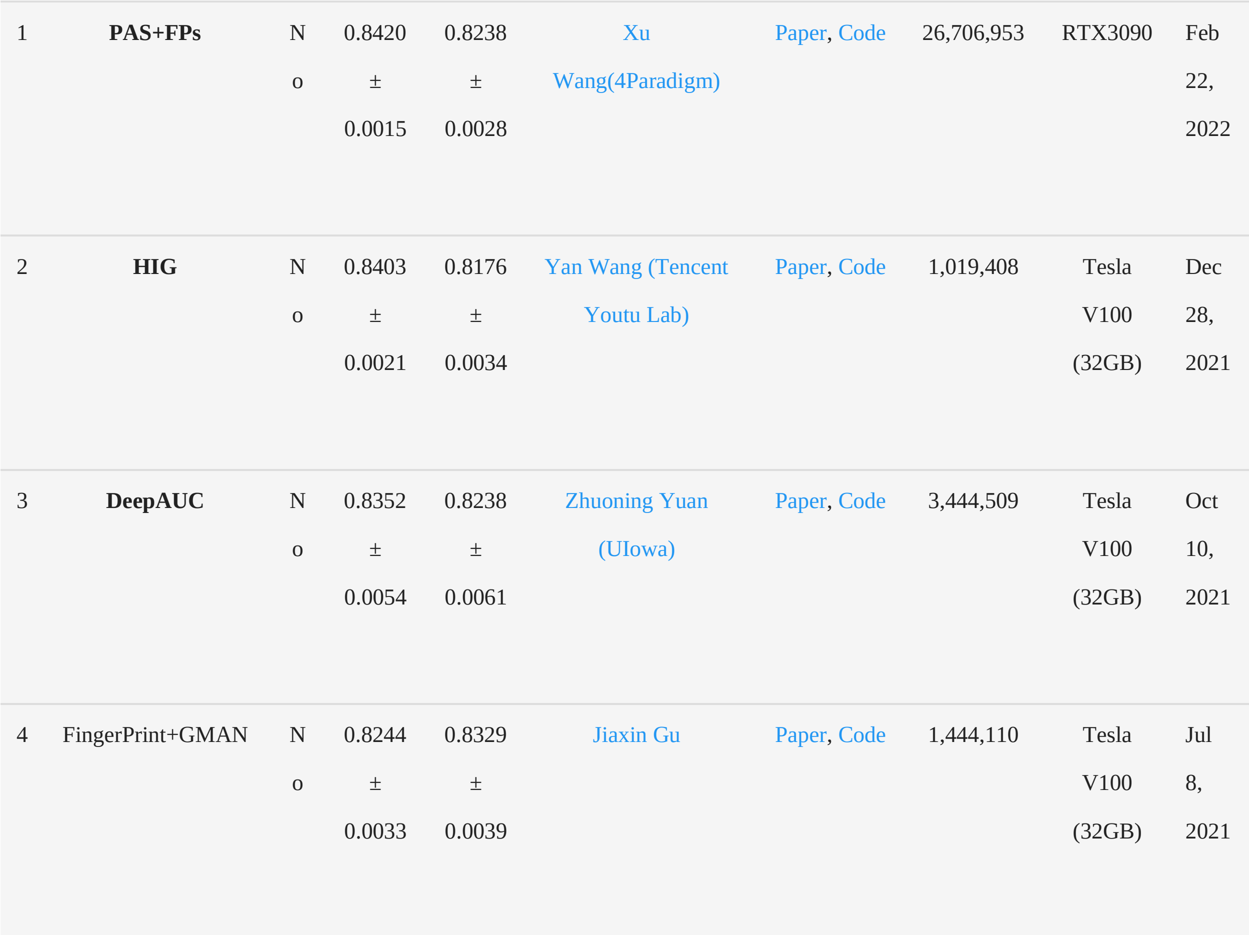
Result

[Table. 1] Performance of models

	2 Encoder layers	1 Encoder + GNN
Test - ROCAUC	0.824	0.824
Valid - ROCAUC	0.837	0.839
Num of Parameters	532,418	456,074
Epochs	20	75

Although the model with only one encoder layer has a smaller number of parameters, it was still able to show the same performance on the test set. However, it took much bigger epochs to train. [Table. 1] As of April 23th in 2022, this test ROCAUC result would take fifth place in the public leaderboard. [Table. 2] In the public dashboard, even though the top 4 achieved better performance on the test set, this model outperformed these models in the validation set.

[Table. 2]



1	**PAS+FPs**	No	0.8420 ± 0.0015	0.8238 ± 0.0028	Xu Wang(4Paradigm)	Paper, Code	26,706,953	RTX3090	Feb 22, 2022
2	**HIG**	No	0.8403 ± 0.0021	0.8176 ± 0.0034	Yan Wang (Tencent Youtu Lab)	Paper, Code	1,019,408	Tesla V100 (32GB)	Dec 28, 2021
3	**DeepAUC**	No	0.8352 ± 0.0054	0.8238 ± 0.0061	Zhuoning Yuan (UIowa)	Paper, Code	3,444,509	Tesla V100 (32GB)	Oct 10, 2021
4	FingerPrint+GMAN	No	0.8244 ± 0.0033	0.8329 ± 0.0039	Jiaxin Gu	Paper, Code	1,444,110	Tesla V100 (32GB)	Jul 8, 2021

Discussion

This paper has shown that using GNN on top of the self-attention encoder layers could work as a more efficient method for a graph dataset by providing the same performance with a lower number of parameters. However, this idea needs to be further proved with future studies.

Also, as a smaller number of parameters is used, this model architecture has one of the possible advantages in minimizing overfitting problems due to reduced model complexity. Nonetheless, considering the drug candidate screening process doesn’t necessarily prefer fast decision-making over accuracy, it may need to be further studied to see if this architecture could possibly improve the performance.

Reference

[1] Mayr, Andreas, et al. "DeepTox: toxicity prediction using deep learning." *Frontiers in Environmental Science* 3 (2016): 80.

[2] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[3] Jiang, Dejun, et al. "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models." *Journal of cheminformatics* 13.1 (2021): 1-23.

[4] Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." *AI Open* 1 (2020): 57-81.

[5] Nguyen, Dai Quoc, Tu Dinh Nguyen, and Dinh Phung. "Universal graph transformer self-attention networks." *arXiv preprint arXiv:1909.11855* (2019).

[6] Hu, Weihua, et al. "Open graph benchmark: Datasets for machine learning on graphs." *Advances in neural information processing systems* 33 (2020): 22118-22133.

[7] Ying, Chengxuan, et al. "Do Transformers Really Perform Badly for Graph Representation?." *Advances in Neural Information Processing Systems* 34 (2021).