Applied Machine Learning Methodology Report

Moaz Mohamed
The University of Adelaide
Adelaide SA 5005
a1779177@student.adelaide.edu.au

1. Introduction

There has been an increase of the usage of machine learning technique in medicine. from helping doctors predicting the severity of diseases [7], and using machine learning for protein folding problems [5]. in this project machine learning will be used to re-purpose existing drugs to treat one of neglected tropical diseases visceral Leishmaniasis. Which is the most deadly species of the Leishmaniasis parasite. Drugs for Neglected Diseases initiative (DNDi) headquartered in Geneva, Switzerland has released the protien targets for Leishmaniasis parasite. an attempted will be made to use message passing neural network to predict the binding affinity [1] of existing drugs at the market with the provided target protein from DNDi.

2. Chemical compounds representation

The usage of machine learning in ligand based screening require having the lest amount reduction in information when representing a chemical compound. a popular representation method called Simplified molecular-input line-entry system (SMILES), its a one line notation encoding of a molecular structure. another representation is graphs. edges can have weights associated with how strong the bond is and each node can have a feature vector that can encode specific attributes for each node as each node represent a chemical element. whereas in SMILES only the elements and the bonds are encoded hence graphs is a better tool for representing chemical compounds [6].

3. Deep Learning on Graphs

3.1. Graph Convolution Neural Network

there are challenges on performing convolution operation on graphs due to the complex nature of graphs and the fact that with the different representation of graphs we can have different adjacency matrix hence the requirement of locality and aggregation of convolution isn't met [4].

Different strategy is needed to apply different deep learning frameworks on graphs. In 2017 on " Neural Message Passing for Quantum Chemistry" [2] outlines and summarized different machine learning methods to apply convolution or commonly what is known in graphs machine learning community as massage passing. the key idea is to use encoder like structure as having lower the dimentionality of a node. that is done by aggregating massages from the target node neighbors then applying a learn-able weight matrix and bias to encode the graph features into a lower dimension. by summing the feature vectors that are connected to the targeted node and multiplying it with a learn-able and differentiable wright matrix. its important to note that the summation operation is order invariant and the notion of depth instead of adding more layers as in typical neural network instead of Graph neural network we are borrowing information from more nodes or going deeper in the graph network to compute the summation and finally after applying the weight matrix and the activation function. we are left with a node embedding

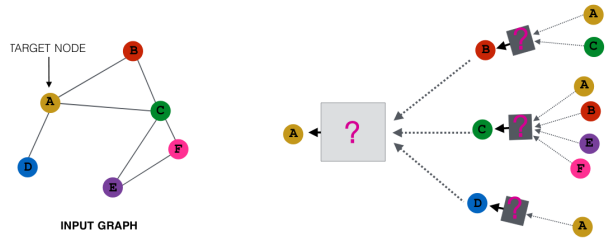

Figure 1. graph computational model

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1}\right) \quad (1)$$

3.2. Aggregation variants

In GraphSage paper [3]. instead of adding both the weight and the bias. the authors concatenated both terms and let the Back-propagation decide which term is more important W_k or B_k, also the aggregation method isn't limited to finding the mean of the neighbors feature vectors. a pool-

ing or applying a LSTM can be utilized and in the GraphSage paper using different aggregation resulted in better performance.

$$\mathbf{h}_v^k = \sigma\left(\left[\mathbf{W}_k \cdot \mathrm{AGG}\left(\left\{\mathbf{h}_u^{k-1}, \forall u \in N(v)\right\}\right), \mathbf{B}_k \mathbf{h}_v^{k-1}\right]\right) \tag{2}$$

$$\mathrm{AGG} = \gamma\left(\left\{\mathbf{Q}\mathbf{h}_u^{k-1}, \forall u \in N(v)\right\}\right) \tag{3}$$

$$\mathrm{AGG} = \mathrm{LSTM}\left(\left[\mathbf{h}_u^{k-1}, \forall u \in \pi(N(v))\right]\right) \tag{4}$$

4. Preparing data and loss function

RDKit will be used to turn the smiles notation into graph representation and to generate the feature vectors for the edge. DeepChem will be used to provide atom feature vector. Since binding affinity is what is required for this project as a continuous prediction as a regression task the mean squared error will be used as a loss function.

References

[1] Evan N. Feinberg, Debnil Sur, Zhenqin Wu, Brooke E. Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S. Pande. PotentialNet for Molecular Property Prediction. *ACS Central Science*, 4(11):1520–1530, 2018.

[2] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *34th International Conference on Machine Learning, ICML 2017*, 3:2053–2070, 2017.

[3] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 2017-December(Nips):1025–1035, 2017.

[4] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–14, 2017.

[5] Frank Noé, Gianni De Fabritiis, and Cecilia Clementi. Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology*, 60:77–84, 2020.

[6] A. A. Toropov, A. P. Toropova, S. E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski. Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines. *Chemometrics and Intelligent Laboratory Systems*, 109(1):94–100, 2011.

[7] Haochen Yao, Nan Zhang, Ruochi Zhang, Meiyu Duan, Tianqi Xie, Jiahui Pan, Ejun Peng, Juanjuan Huang, Yingli Zhang, Xiaoming Xu, Hong Xu, Fengfeng Zhou, and Guoqing Wang. Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests. *Frontiers in Cell and Developmental Biology*, 8(July):1–10, 2020.