Deep Learning Project Proposal

Apurva Goel
apurvago
apurvago@buffalo.edu

Amitha Kethavath
amithake
amithake@buffalo.edu

Param Patel
paramnar
paramnar@buffalo.edu

Title

Improving Drug Discovery by Inclusion of Graph-Level Representation in Graph Neural Networks.

Short Summary

Drug discovery, design, and development are primary example of Graph Neural Networks (GNN) [Scarselli et al., 2009, Sanchez-Lengeling et al., 2021, Yang et al., 2019]. GNNs can represent the model outputs at two different levels:

- Node-level representation and,
- Graph-level representation

However, most of the previous research and models utilised in this domain of in-silico-based biochemical engineering focus solely on the node-level representation and ignore the graph-level representation. Based on one of the research studies provided by Li et al. [2017] titled **Learning Graph Level Representation for Drug Discovery**, they open up new avenues and provide insights on how graph-level representations could enhance and improve the previous findings.
In summary, we used the above-mentioned study for the classification of various chemical compounds based on their toxicity and other variables.

Objectives

- To understand and work with Graph Neural Networks Cao et al. [2016].
- To understand the graph-level representation of GNNs using Graph Convolutional Network Daigavane et al. [2021], Duvenaud et al. [2015].
- Comparison and contrast study of the differences between the node-level and graph-level representation of the output provided by the GNN.
- Compare our results with previous studies in a similar field.

Methodology

For the checkpoint, we use a 7-layer architecture. It contains 3 Graph Convolutional Layers, 2 Graph Pooling layers Weng et al. [1992], 1 dense layer and 1 output layer. However, in the future, we may increase the number of layers.
The graph convolution layers will each have a convolution layer, a ReLU as an activation layer and a batch normalisation layer Ioffe and Szegedy [2015].

Possible Future Modifications:

- Increasing the number of layers.
- Adding a Dropout Layer Srivastava et al. [2014].
- Working with different activation functions.
- Working with different optimizers.

Evaluation

For evaluation purposes, we plan to use the following metrics:

I **Quantitatively**
  (a) Confusion matrix
    - Accuracy
    - Precision
    - Recall
    - F1 Score
  (b) Area Under ROC Curve (AUC-ROC)
  (c) Log-Loss
  (d) Cross-Entropy Loss

II **Qualitatively**
  For visualization, we plan on plotting a bar graph for the confusion matrix and the confusion matrix itself. For AUC-ROC, and Loss calculation (both Log-Loss and Cross-Entropy Loss), we plan on using a distinct line graph for every aforementioned metric for every dataset mentioned in the next section. So, in brief:
  (a) Bar Graph
  (b) Line Graph
  (c) Confusion Matrix

Dataset

We use a subset of databases from **MoleculeNet** Wu et al. [2018], a public dataset collection for our purpose. They are:

1. ToxCast
2. Tox21 and
3. HIV

All the above-mentioned datasets contain various distinct qualitative data about toxins present in chemical compounds.

GitHub Link

Summer Deep Learning Final Project.

Individual Tasks Screenshots

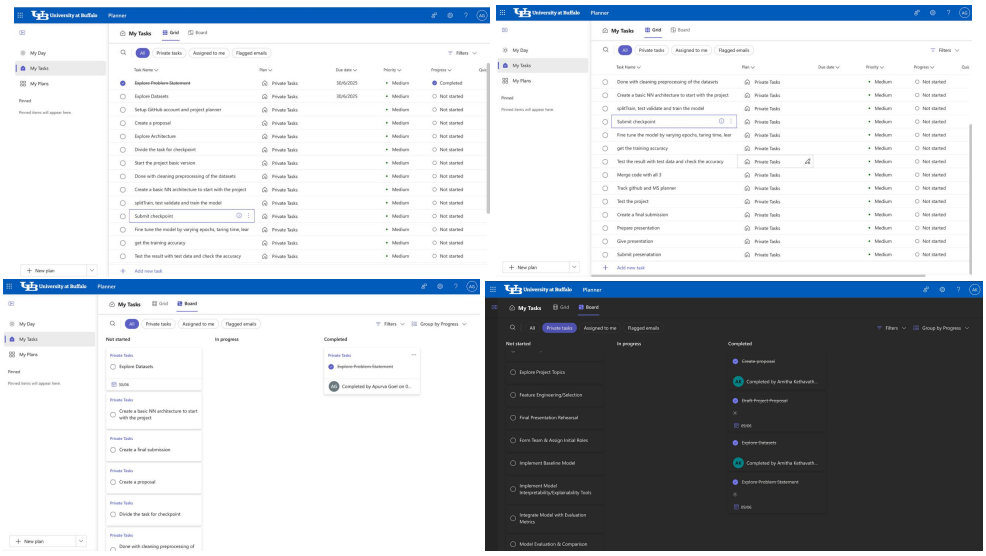

Figure 1: Screenshots of Individual Tasks

References

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alexander B. Wiltschko. A gentle introduction to graph neural networks. *Distill*, 2021. doi: 10.23915/distill.00033. https://distill.pub/2021/gnn-intro.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction, 2019. URL https://arxiv.org/abs/1904.01561.

Junying Li, Deng Cai, and Xiaofei He. Learning graph-level representation for drug discovery, 2017. URL https://arxiv.org/abs/1709.03741.

Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Feb. 2016. doi: 10.1609/aaai.v30i1.10179. URL https://ojs.aaai.org/index.php/AAAI/article/view/10179.

Ameya Daigavane, Balaraman Ravindran, and Gaurav Aggarwal. Understanding convolutions on graphs. *Distill*, 2021. doi: 10.23915/distill.00032. https://distill.pub/2021/understanding-gnns.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints, 2015. URL https://arxiv.org/abs/1509.09292.

J. Weng, N. Ahuja, and T.S. Huang. Cresceptron: a self-organizing neural network which grows adaptively. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 1, pages 576–581 vol.1, 1992. doi: 10.1109/IJCNN.1992.287150.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning, 2018. URL https://arxiv.org/abs/1703.00564.