Assignment-3

GMLFA (AI60007) - Autumn,2024 - IIT Kharagpur

Release Date: [18/09/2024]
Submission Date: [17/10/2024]
Total Marks: 14

General Instructions:

- All graded questions are compulsory to solve, and non-graded questions are optional.
- ***Negative marking*** will be there as per our ***plagiarism policy*** given in the course webpage.
- You can use any language for coding questions, but ***'python'*** is preferred.
- Frameworks like Pytorch and Tensorflow are encouraged to construct deeper neural network architectures.
- Any required help will be provided to you in the code notebook regarding data or any specific library.

Submission Instructions:

Following are the Deliverables and submission instructions for the assignment:

1. **Code Notebook (.ipynb):** A notebook containing all the code, including the implementation and execution of experiments. Notebook Format: <***group_number>_assignment3.ipynb,*** replace <***group_number***> with your assigned group number.
2. **Dataset Path:** Make sure you keep the dataset files in the following path: ***'/content/<dataset>/'***
3. Ensure that the notebook runs smoothly considering the aforementioned dataset path to yield the expected results.
4. **Report (.pdf):** A comprehensive report documenting all findings from the experiments conducted. Report Format: ***group_number_assignment3.pdf***
5. The report must present detailed results for each dataset across various k values in a tabular format. Ensure that the best results for each dataset are highlighted in bold.
6. ***As this is an open-ended research problem, grading will be based on the classification performance on the test set. Students are encouraged to conduct extensive hyperparameter tuning on the validation set to enhance performance.***

Problem Statement:

Down-Sample and Pool : In this assignment you need to implement a new graph pooling algorithm for graph classification tasks. Steps of the algorithm are as follows:

Down-Sample & Pool $(A^{(l)}, X^{(l)}, k, \mathbf{m}^{(l+1)})$

- Input : $A^{(l)}$ **:** Adjacency Matrix and $X^{(l)}$ **:** Feature Matrix at layer l
- ***Down-Sampling Layer :*** Down-sample most important **k**%-nodes which are crucial for the downstream task. Use **gPool layer** to adaptively select some nodes to form a smaller graph based on their scalar projection values on a trainable projection vector.
- ***Hierarchical Pooling :*** Use a down-sampled set of nodes to learn a new coarsened graph corresponding to a $\mathbf{m}^{(l+1)}$-cluster of nodes in the graph. Use **diffpool layer** to learn the new coarsened graph $A^{(l+1)}$and their modified node feature set $X^{(l+1)}$

Model Architecture: Model architecture for graph classification task are as follows:

$GNN_1 \rightarrow GNN_2 \rightarrow$ ***Down-Sample & Pool***$_1 \rightarrow GNN_3 \rightarrow GNN_4 \rightarrow$ ***Down-Sample & Pool***$_2 \rightarrow$ ***Classification Head***

Tasks:

Implement the aforementioned model architecture for graph classification task. Different design choices (hyper-parameters) are as follows :

1. ***GNN Model -*** Students need to conduct experiments on GCN [Kipf et.al.] model and report comparative results in the report. **(3 Marks)**
2. ***K in Down-Sampling Layer :*** Students need to conduct experiments with different values of k = {90%, 80%, 60%} at each down-sampling layer to downsample important k%-nodes. Report comparative results. **(1+1+1=3 Marks)**
3. ***M in Hierarchical Pooling Layer :*** Students need to conduct experiments with m = 6 and 3 for ***Down-Sample & Pool***$_1$ and ***Down-Sample & Pool***$_2$ respectively **(1 Mark)**

Datasets:

You need to implement above Tasks on three benchmark protein structure dataset:

a. ***D&D -*** Binary Classification Task
b. ***ENZYMES -*** 6-Class Classification Task

Divide each dataset into three parts: 80% for Training, 10% for Validation, and 10% for Testing. Train your model using the Training set, and use the Validation set to tune and validate it. Once you've selected the best model based on its performance on the Validation set, evaluate it on the Test set and report the classification accuracy on this Test set.

You need to report all your findings after conducting experiments on all these datasets (7 marks each - Total 14 Marks)