Report on Graph Neural Network Model

Introduction

This report details the implementation and evaluation of a Graph Neural Network (GNN) model based on the GraphSAGE architecture. The model was trained on the Lipophilicity dataset from MoleculeNet, using the PyTorch Geometric library. The primary objective was to predict lipophilicity values of molecular graphs.

Dataset

The Lipophilicity dataset, provided by MoleculeNet, contains molecular graphs with node-level features and edge connections. Key preprocessing steps included:

- Shuffling and splitting the dataset into training (80%), validation (20%).
- Normalizing node features and converting them to `float32` format.

Model Architecture

The model employs the GraphSAGE architecture, which aggregates features from neighboring nodes in a graph. The key components are:

GraphSAGE Layers

- Six GraphSAGE convolutional layers, each followed by ReLU activation.
- Skip connections between layers to enhance gradient flow and model expressivity.

Pooling

- A global mean pooling operation aggregates node-level representations to graph-level representations.

Fully Connected Layers

- A hidden dense layer with ReLU activation.
- An output dense layer for final predictions.

Training and Evaluation

The dataset was shuffled and split into training and validation sets. The model's performance was evaluated using the **R^2 score**, which measures the proportion of variance explained by the predictions.

Results

The training and validation results are as follows:

- **Training R^2 Score:** 0.9715
- **Validation R^2 Score:** 0.6875

Visualization of training loss and validation performance trends over epochs is provided in the notebook.

Conclusion

The GraphSAGE model demonstrated its ability to learn from graph-structured data and predict lipophilicity values effectively. Future work may involve fine-tuning the hyperparameters or exploring alternate GNN architectures to improve performance further.