Simple Baselines for OGBL-Vessel Dataset

1 Introduction

The ogbl-vessel dataset presents a significant challenge for various graph algorithms. In particular, it is observed that GCN and GraphSAGE cannot even outperform MLP, and all are outperformed by SEAL with a significant margin (approximately 50% vs. 80% Val/Test ROC-AUC). This report discusses techniques and improvements in handling the ogbl-vessel dataset, considering aspects like self-loops, node embeddings, preprocessing, predictors, and analysis of results.

2 Add Self-Loops

Some isolated nodes are present in the graph (val/test related nodes), possibly due to the data split. Their neighbor averaging output is zero, causing unnecessary imbalance in data distribution. The self-loops can improve GraphSAGE with an example script provided, yielding similar performance to MLP. However, isolated nodes are not the only reason for the poor performance of GraphSAGE on this dataset.

3 Node Embedding

Replacing or concatenating coordinate features with learnable or pretrained node embedding vectors can enhance GraphSAGE, achieving 60% Val/Test ROC-AUC. However, overfitting remains an issue.

4 Preprocessing (for Raw Coordinate Features)

Different preprocessing techniques are discussed, such as node-wise normalization, channel-wise normalization, max-min scaling, z-score scaling, and logarithmic scaling. These methods aim to control the magnitude of coordinates and avoid potential issues with large values.

5 Predictor

Various predictors are proposed to preserve the relative information of realistic coordinates. Examples include 'DIFF', 'CONCAT', 'MEAN', 'COS', 'SUM', 'MAX', where 'DIFF' or DiffLinear seems to be more meaningful. GCN/GraphSAGE and MLP can achieve significant improvements with these predictors.

6 Results

The following tables present the results for different models and preprocessing techniques:

6.1 MLP Results

Preprocessing	Model	Predictor	Train	Val	Test
node_norm.	MLP	Dot	50.35±0.00	50.40±0.00	50.28±0.00
channel_norm.	MLP	Dot	64.83±2.74	64.82±2.72	64.83±2.74
log	MLP	Dot	68.62±5.12	68.58±5.11	68.59±5.09
max-min	MLP	Dot	68.63±3.50	68.64±3.52	68.63±3.50
z-score	MLP	Dot	75.22±1.26	75.20±1.26	75.19±1.24
none	MLP	Dot	50.00±0.00	50.00±0.00	50.00±0.00
node_norm.	MLP	Diff	82.77±5.66	82.76±5.67	82.77±5.66
channel_norm.	MLP	Diff	85.38±0.01	85.39±0.02	85.40±0.02
log	MLP	Diff	93.24±0.14	93.23±0.14	93.22±0.14
max-min	MLP	Diff	94.03±0.03	94.04±0.03	94.02±0.03
z-score	MLP	Diff	**94.16±0.02**	**94.15±0.02**	**94.14±0.01**
none	MLP	Diff	94.01±0.03	94.01±0.03	94.00±0.03

6.2 GraphSAGE Results

Preprocessing	Model	Predictor	Train	Val	Test
node_norm.	SAGE	Dot	50.35±0.00	50.40±0.00	50.28±0.00
channel_norm.	SAGE	Dot	50.50±0.00	50.53±0.01	50.36±0.01
log	SAGE	Dot	50.47±0.03	50.50±0.02	50.38±0.03
max-min	SAGE	Dot	50.71±0.03	50.71±0.03	50.55±0.03
z-score	SAGE	Dot	50.83±0.02	50.77±0.02	50.62±0.01
none	SAGE	Dot	50.00±0.00	50.00±0.00	50.00±0.00
node_norm.	SAGE	Diff	79.97±1.95	72.00±1.11	71.99±1.10
channel_norm.	SAGE	Diff	91.90±0.31	76.11±0.62	76.11±0.62
log	SAGE	Diff	96.74±0.38	82.75±0.79	82.74±0.80
max-min	SAGE	Diff	97.81±0.23	84.71±0.54	84.72±0.53
z-score	SAGE	Diff	98.06±0.02	85.64±0.38	85.65±0.38
none	SAGE	Diff	98.11±0.06	87.71±0.07	87.71±0.07

7 Conclusion

Empirically, the coordinate features are robust enough to obtain over 90% Train/Val/Test ROC-AUC. In many cases, graph convolution may cause overfitting and significant degradation. Future work may consider other metrics, such as Hits@xx, to further analyze model performance.

8 Reference

- Open Graph Benchmark
- ogbl-vessel