Technical Report for OGB Link Property Prediction

Liang Yao[a], Qiang Liu[a], Hongyun Cai[a], Shengong Ji[a], Feng He[a], Xu Cheng[a]

[a] *Tencent Inc., Shenzhen, China*

Abstract

This technical report introduces our solution for two OGB Link Property Prediction challenges, ogbl-collab and ogbl-citation2.

1. Introduction of Open Graph Benchmark Challenge

The Open Graph Benchmark (OGB) is a collection of realistic, large-scale, and diverse benchmark datasets for machine learning on graphs [1]. The ogbl-collab and ogbl-citation2 are two datasets for link prediction. The ogbl-collab dataset is an undirected graph, representing a subset of the collaboration network between authors indexed by Microsoft Academic Graph (MAG) [2]. Each node represents an author and edges indicate the collaboration between authors. All nodes come with 128-dimensional features, obtained by averaging the word embeddings of papers that are published by the authors. The task is to predict the author collaboration relationships in a particular year given the past collaborations. The evalution metric is Hits@50. The ogbl-citation2 dataset is a directed graph, representing the citation network between a subset of papers extracted from MAG. Each node is a paper with 128-dimensional Word2Vec features that summarizes its title and abstract, and each directed edge indicates that one paper cites another. The task is to predict missing citations given existing citations. The evaluation metric is Mean Reciprocal Rank (MRR).

2. Method

Feature propagation [3, 4] has shown its success in node classification tasks, but it remains unexplored in link prediction tasks. In this experiment, given the original feature vector X and adjacent matrix A, we replace the feature vector X with propagated features $[X, \tilde{A}X, \tilde{A}^2X]$, where $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$

Table 1: Test performance for different methods. We run all models 10 times and and report mean ± standard deviation.

Method	ogbl-collab (Hits@50)	ogbl-citation2 (MRR)
PLNLP	0.7046 ± 0.0040	–
PLNLP + $[X, \tilde{A}X, \tilde{A}^2X]$	$\mathbf{0.7087 \pm 0.0033}$	–
MLP	0.1991 ± 0.017	0.2900 ± 0.0018
MLP + $[X, \tilde{A}X, \tilde{A}^2X]$	0.2839 ± 0.0127	0.3224 ± 0.0017

is the normalized symmetric adjacency matrix. The results are organized into Table 1. Our method achieves better results compared with the state of the art model PLNLP [5] or MLP as the backbone model. To the best of our knowledge, our method ranks No.1 on leaderboard for ogbl-collab until now.

References

[1] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, J. Leskovec, Open graph benchmark: Datasets for machine learning on graphs, arXiv preprint arXiv:2005.00687 (2020).

[2] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft academic graph: When experts are not enough, Quantitative Science Studies 1 (1) (2020) 396–413.

[3] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, K. Weinberger, Simplifying graph convolutional networks, in: International conference on machine learning, PMLR, 2019, pp. 6861–6871.

[4] E. Rossi, F. Frasca, B. Chamberlain, D. Eynard, M. Bronstein, F. Monti, Sign: Scalable inception graph neural networks, arXiv preprint arXiv:2004.11198 (2020).

[5] Z. Wang, Y. Zhou, L. Hong, Y. Zou, H. Su, Pairwise learning for neural link prediction, arXiv preprint arXiv:2112.02936 (2021).