Sampling for Heterogeneous Graph Neural Networks

Feiyang Chen
005526097
fychen@cs.ucla.edu

Yongqian Li
004997466
yongqianli@g.ucla.edu

Ruoyu He
805729152
rhe9527@g.ucla.edu

YuanChing Lin
9493782067
yclin99@g.ucla.edu

Abstract

Graph sampling is a popular technique in training large-scale graph neural networks (GNNs), recent sampling-based methods have demonstrated impressive success for homogeneous graphs. However, objects around us and interactions between them in real life are often multi-modal and multi-typed, i.e.we live in a world of heterogeneous graphs. But only a few of the recent works have paid attention to sampling for heterogeneous graphs. In this work, we aim to study sampling methods for heterogeneous GNNs. We first analyze representative sampling algorithms on homogeneous graphs and evaluate their performance on heterogeneous graphs. At last, we will try to design a novel sampling method targeting heterogeneous graphs.

1 Proposal

Objects in the real world are often defined in terms of their relationship, such a set of objects and interactions between them is naturally represented as a graph. Graph neural networks (GNNs) [1] have recently received more and more attention since their power of modeling graph-structured data, which has shown impressive success in many practical applications, such as drug discovery [2], recommendation systems [3], traffic prediction [4], etc. However, training large-scale graph neural networks efficiently is still a challenge, especially because the computation of full-batch GCN's can be very expensive when the graph is large and dense[5]. To address this challenge, current sampling-based methods [5, 6, 7, 8] are developed during the past years and making great progress in homogeneous graphs. But our real world is mainly based on heterogeneous graphs, i.e. objects around us and interactions between them are often multi-typed. Unfortunately, only a few of the recent works have paid attention to sampling for heterogeneous graphs. In this work, we aim to study sampling methods for heterogeneous GNNs. We first analyze 3 representative sampling algorithms on homogeneous graphs, including node-wise sampling [5], layer-wise sampling [6], and subgraph-wise sampling [7]. Then we evaluate their performance on heterogeneous graphs. At last, we will try to design a novel sampling method targeting heterogeneous graphs. Our main motivation is making such sampling method achieve good efficiency without compromising on the effectiveness on heterogeneous graphs, which has not been explored before.

1.1 Node-wise Sampling

Node-wise sampling means that for each node at each layer, the network samples a small set of neighbors, and only aggregates the information in the sampled neighborhood. A classical example of node-wise sampling is **GraphSAGE** (SAmple and aggreGatE) [5], which focuses on increasing the efficiency of the neural network on prediction of the unseen nodes. The advantage of node-wise sampling is the reduced direct receptive field of each node at each layer. But the disadvantage is that the direct receptive field grows exponentially with respect to the number of GNN layers, not to mention the redundancy of nodes during the computation.

Preprint. Work in progress.

1.2 Layer-wise Sampling

Layer-wise sampling refers to sampling a small set of nodes for each layer, then in the aggregation step, for each node, only aggregate the information among the neighbor nodes that are in the sampled node set for the previous layer. A representative method is LAyer-Dependent ImportancE Sampling (**LADIES**)[6], which is based on layer-dependent sampling scheme to avoid exponential expansion of receptive field as well as guarantee the connectivity of the sampled adjacency matrix. Compared with existing GCN training methods including GraphSAGE[5] and FastGCN[9], LADIES has significantly lower memory cost, time complexity and estimation variance.

1.3 Subgraph-wise Sampling

Subgraph-wise sampling is the type of methods which samples subgraphs and then for each batch, use one sampled subgraph for training. A typical exemplary model for Homogeneous graphs is **Cluster-GCN** [7]. The key idea of this work is to leverage the fact that a mini-batch SGD algorithm's efficiency is proportional to the number of links between nodes in each batch ("within-batch links"): therefore, using efficient graph clustering algorithms to construct subgroups for each batch would be beneficial. Cluster-GCN applies graph clustering methods which results in much more within-cluster links than between-cluster links to partition the graph. Comparing with other SGD-based approaches, Cluster-GCN reaches better or similar training speed for shallow networks and has better memory usage. It is also able to train a very deep network with large embedding size.

2 Experiments

In our experiments, we will focus on node classification task for heterogeneous graphs. Relational Graph Convolutional Networks (R-GCNs)[10] is a method for modeling (heterogeneous) knowledge graphs, we will follow their algorithm as baseline model. The benchmark dataset is from Open Academic Graph (OAG)[11], including all CS papers (8.1G), all ML papers (1.9G), all NN papers (0.6G) spanning from 1900-2020. It is an academic graph in which papers, authors, venues, institutions, and fields are nodes, while the edge types also include different interaction relations among them. Besides, we will also consider other heterogeneous graphs datasets include DBLP [1]. DBLP is a computer science bibliography website and it publish is the same name dataset. The dataset contains four kinds of nodes and four relations: Author-label, Paper-author, paper-conference, and paper-type. The statistics are shown as Table 1 and Table 2. For these datasets, we will evaluate the above-mentioned sampling methods for heterogeneous graphs and use the results as baselines to compare with.

3 Schedule

3.1 Timeline

- Week 5: Apply 3 representative sampling methods (**GraphSAGE**, **LADIES**, **Cluster-GCN**) on homogeneous graphs, and **R-GCNs** algorithm on heterogeneous graphs as baseline;
- Week 6-7: Evaluate above sampling methods on heterogeneous graphs;
- Week 8-10: Explore a novel sampling method targeting heterogeneous graphs.

3.2 Workload distribution

- **GraphSAGE**: Ruoyu He;
- **LADIES**: Feiyang Chen;
- **Cluster-GCN**: Yongqian Li;
- **R-GCNs**: YuanChing Lin;
- Explore a novel sampling method and write the report: All members.

[1] https://dblp.org/

Datasets	#nodes	#edges	#papers	#authors	#fields	#venues	#institutes	#P-A	#P-F	#P-V	#A-I	#P-P
CS	11,732,027	107,263,811	5,597,605	5,985,759	119,537	27,433	16,931	15,571,614	47,462,559	5,597,606	7,190,480	31,441,552
OAG	178,663,927	2,236,196,802	89,606,257	88,364,081	615,228	53,073	25,288	300,853,688	657,049,405	89,606,258	167,449,933	1,021,237,518

Table 1: Open Academic Graph (OAG) Statistics

Datasets	#authors	#papers	#label	#conference	#type	#A-L	#P-A	#P-C	#P-T
DBLP	14,475	14,376	4	20	8,920	4,057	41,794	14,376	114,624

Table 2: DBLP Dataset Statistics

References

[1] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[2] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, 13(1):1–23, 2021.

[3] Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1623–1625, 2022.

[4] Frederik Diehl, Thomas Brunner, Michael Truong Le, and Alois Knoll. Graph neural networks for modelling traffic participant interaction. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 695–701. IEEE, 2019.

[5] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[6] Difan Zou, Ziniu Hu, Yewen Wang, Song Jiang, Yizhou Sun, and Quanquan Gu. Layer-dependent importance sampling for training deep and large graph convolutional networks. *Advances in neural information processing systems*, 32, 2019.

[7] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 257–266, 2019.

[8] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. GraphSAINT: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020.

[9] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*, 2018.

[10] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018.

[11] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, pages 2704–2710, 2020.