GraphStorm an Easy-to-use and Scalable Graph Neural Network Framework: From Beginners to Heroes

Jian Zhang
AWS AI
Santa Clara, USA
jamezhan@amazon.com

Da Zheng
AWS AI
Santa Clara, USA
dzzhen@amazon.com

Xiang Song
AWS AI
Santa Clara, USA
xiangsx@amazon.com

Theodore Vasiloudis
AWS AI
Seattle, USA
thvasilo@amazon.com

Israt Nisa
AWS AI
New York, USA
nisisrat@amazon.com

Jim Lu
AWS AI
Seattle, USA
luzj@amazon.com

ABSTRACT

Applying Graph Neural Networks (GNNs) to real-world problems is challenging for machine learning (ML) practitioners due to two major obstacles. The first hurdle is the high barrier to learn programming GNNs from scratch. The second challenge lies in overcoming engineering difficulties when scaling GNN models for large graphs at an industry-level. To address these challenges, GraphStorm, an open-source framework, offers a solution by providing an easy-to-use user interface and an end-to-end GNN training/inference pipeline that seamlessly handles extremely large graphs in a distributed manner This tutorial aims to provide participants with a comprehensive understanding of GraphStorm, including its design principles, target users, and use cases, through presentations. The hands-on sections will enable attendees to walk through four practical GraphStorm use cases that can assist them in leveraging GNNs to address real-world business problems.

KEYWORDS

Graph Neural Networks, Distributed Training, GraphStorm

ACM Reference Format:
Jian Zhang, Da Zheng, Xiang Song, Theodore Vasiloudis, Israt Nisa, and Jim Lu. 2023. GraphStorm an Easy-to-use and Scalable Graph Neural Network Framework: From Beginners to Heroes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599179

1 TARGET AUDIENCE AND PREREQUISITES FOR THE TUTORIAL

Intent audience: This tutorial targets machine learning practitioners who are interested in or already working in graph machine learning tasks, and want to leverage easy-to-use and scalable tools to accelerate GNN adoption to address their own business problem, and researchers who are interested in experimenting their novel GNN models on large graphs.

Prerequisites: The attendees should have some knowledge with deep learning on graphs, and have used deep learning frameworks, e.g., Pytorch. Knowledge about graph neural network and DGL are better to have, but not required.

Takeouts after participation of the tutorial: We expect that the attendees will have an understanding of GraphStorm's basic information and application use cases. They will also know how to use GraphStorm in standalone mode to train GNN models for their own extensive graph data.

2 TUTORS

1. Jian Zhang, AWS AI, jamezhan@amazon.com
2. Da Zheng, AWS AI, dzzhen@amazon.com
3. Xiang Song, AWS AI, xiangsx@amazon.com

3 TUTORS' SHORT BIO

3.1 List of in-person presenters

1. Jian Zhang: Jian is a senior applied scientist at AWS AI, using ML techniques to help customers solve various problems, such as fraud detection, image generation. He has successfully developed and deployed GNN solutions for customers world-widely.
2. Da Zheng: Da is a senior applied scientist at AWS AI, leading the effort of building frameworks and algorithms to bring graph machine learning technologies in production. This includes DGL for GNN, DGL-KE for knowledge graph embeddings, DistDGL for scaling GNN training to billion-scale graphs, TGL for temporal GNNs, and more.
3. Xiang Song: Xiang is a senior applied scientist at AWS AI, leading the effort of building frameworks and services for industrial applications. This includes DGL and DistDGL for scaling GNN to large scale graphs, Neptune ML, an graph ML service designed for Amazon Neptune graph database.

3.2 List of contributors

1. Theodore Vasiloudis: Theodore is an applied scientist who works in distributed machine learning and data processing.
2. Israt Nisa: Israt is an applied scientist who specializes in developing scalable and high-performing modules for GNNs.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599179

Her expertise lies in optimizing low-level kernels specifically for GPUs.
3. Jim Lu: Jim is a senior product manager who leads the development of GraphStorm and its open source.

4 TUTORIAL OUTLINE

This tutorial will guide ML practitioners to master GraphStorm with presentations and hands-on practices. The presentation will briefly introduce GraphStorm's design ideas, target users, and use cases, which will help attendees to understand who can benefit from GraphStorm the most, and what GraphStorm is good at. [section format: slides]

The hands-on practices help attendees master four GraphStorm use cases in four sections.

Section 1: GraphStorm Quick-start. This section provides an overview of the standalone GraphStorm environment setup, including installation of Docker, building GraphStorm Docker images from source code. This section will also guide attendees to go through an example that train a GraphStorm built-in RGCN model on an OGB-Arxiv graph. [section format: hands-on with scripts]

Section 2: Use Your Own Graphs. This section outlines the process of preparing users' own graph data and employing GraphStorm's tools to process customer graphs into the format suitable for using GraphStorm. In this section, we will use a synthetic transaction dataset as an example of user's own data, and then use GraphStorm launch commands to train built-in models on a fraud detection task. Additionally, this section also provides guidance on tunning GraphStorm's configurations to optimize performance. [section format: hands-on with scripts and code snippets]

Section 3: User Your Own GNN Models. This section explains the procedure of adapting user-defined GNN models via GraphStorm's custom model APIs, which enable the modified models to be trained within the GraphStorm framework. The section will encompass an overview of GraphStorm's custom APIs and utilize the Heterogeneous Graph Transformer (HGT) as an exemplar of user-defined GNN models to help attendees learn the adaptation skills. [section format: hands-on with scripts and code snippets]

Section 4: Distributed GraphStorm. This section illustrates the standard configuration and deployment of a GraphStorm cluster that comprises multiple GraphStorm instances. With such clusters, GraphStorm is capable of training GNN models on vast graphs, measured in billions of nodes and tens of billions of edges. Furthermore, this section employs a GraphStorm cluster to train on the OGB-Papers100M graph. Owing to possible constraints of attendees' devices, this section is limited to demonstration only. [section format: slides and demo videos]

5 SIMILAR/HIGHLY RELATED TUTORIAL

KDD 2020: Scalable Graph Neural Networks with Deep Graph Library: https://dl.acm.org/doi/abs/10.1145/3394486.3406712

KDD 2021: All You Need to Know to Build a Product Knowledge Graph:https://naixlee.github.io/Product_Knowledge_Graph_Tutorial_KDD2021/

KDD 2022: Graph Neural Networks in Life Sciences: Opportunities and Solutions:https://dl.acm.org/doi/abs/10.1145/3534678.3542628

Although the aforementioned tutorials primarily address graph neural networks (GNNs) in a theoretical context, introducing GNN libraries and applications, the proposed hands-on tutorial specifically targets a ready-to-use framework. Notably, this novel framework significantly reduces the barrier to GNN adoption and expands GNN application areas, particularly in settings where large-scale graph data is prevalent.

REFERENCES

[1] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*. 2704–2710.

[2] George Karypis and Vipin Kumar. 1997. METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. (1997).

[3] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[4] Vasimuddin Md, Sanchit Misra, Guixiang Ma, Ramanarayan Mohanty, Evangelos Georganas, Alexander Heinecke, Dhiraj Kalamkar, Nesreen K Ahmed, and Sasikanth Avancha. 2021. Distgnn: Scalable distributed training for large-scale graph neural networks. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–14.

[5] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 593–607.

[6]]velivckovicgraph Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. [n. d.]. Graph Attention Networks. In *International Conference on Learning Representations*.

[7] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* (2019).

[8] Hongxia Yang. 2019. Aligraph: A comprehensive graph neural network platform. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 3165–3166.

[9] Chenguang Zheng, Hongzhi Chen, Yuxuan Cheng, Zhezheng Song, Yifan Wu, Changji Li, James Cheng, Hao Yang, and Shuai Zhang. 2022. ByteGNN: efficient graph neural network training at large scale. *Proceedings of the VLDB Endowment* 15, 6 (2022), 1228–1242.

[10] Da Zheng, Minjie Wang, Quan Gan, Xiang Song, Zheng Zhang, and George Karypis. 2021. Scalable graph neural networks with deep graph library. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 1141–1142.

6 ENGAGEMENT STRATEGIES

This tutorial will provide detailed demonstrations and step-by-step hands-on materials, allowing attendees to easily learn GraphStorm. Since most attendees will likely have access to proper computing resources, they will have the opportunity to experiment with GraphStorm on their own during the tutorial. Furthermore, the speakers will solicit feedback from attendees on the tutorial topics and adjust the length or depth of content based on the majority of the audience's preferences.

7 SOCIETAL IMPACTS

This tutorial is expected to have a significant impact on the graph machine learning community. It will facilitate the rapid adoption of Graph Neural Network models into real-world applications that require handling industry-level graphs. In addition, the tutorial is likely to drive innovation within the academic community. Researchers will be able to leverage GraphStorm to conduct GNN research and test their ideas on very large graphs, which are previously challenging.