Augmentation Methods for Graph Learning

Tong Zhao[1]*, Kaize Ding[2], Wei Jin[3], Gang Liu[4], Meng Jiang[4], Neil Shah[1]

Abstract

Data augmentation (DA) has recently seen increased interest in graph mining and graph machine learning (ML) owing to its ability to create additional training data and improving resulting trained model generalization. Despite these developments, the area is still quite underexplored, due to the challenges brought by complex, non-Euclidean structure of graph data, which limits the direct analogizing of traditional label-preserving augmentation operations in e.g. computer vision or natural language to this domain. In this tutorial, we present a comprehensive, systematic and structured survey of graph DA (GDA) approaches. We first overview necessary background in graph machine learning, graph neural networks (GNNs), and DA. Next, we present a breadth of literature, separated by heuristic augmentation approaches, learned augmentation approaches, and conclude. We anticipate this tutorial will be valuable for researchers in the graph ML domain, as well as practitioners utilizing these techniques for a wide variety of low-resource applications lacking significant labeled data.

1 Motivation and Rationale

Data driven inference has received a significant boost in generalization capability and performance improvement from DA techniques. These methods increase the amount of training data by creating plausible variations of existing data without additional human annotations, and have seen widespread adoption in fields such as computer vision (CV) and natural language processing (NLP). Such augmentations allow inference engines to learn to generalize better across those variations and attend to signal over noise. At the same time, graph ML methods such as GNNs have emerged as a rising approach for data driven inference on graphs. GNNs are neural networks that can be directly applied to graphs and provide a convenient and learnable method to address node-level, edge-level, and graph-level prediction tasks. Given a graph, GNNs follow a neighborhood aggregation scheme, where a node's representation is computed by recursively aggregating and transforming representation vectors of its neighboring nodes [26].

*[1]Snap Inc., [2]Arizona State University, [3]Michigan State University, [4]University of Notre Dame. Correspondence to Tong Zhao: <tzhao@snap.com>

Since GNNs use graph structure as the underlying computation graph for the network, a natural observation is that the performance of such methods is largely influenced by certain properties of the graph, such as correctness, sparsity, (skewed) degree distributions, and more [19]. GDA is thus a promising approach to alleviate these concerns by adjusting graph structure to the end of improving GNN model performance on target tasks. Indeed, several existing studies proposed heuristic augmentation methods [20, 7, 24, 9, 32] and learnable augmentation methods [14, 13, 33, 34, 28, 15, 18] on graph data, demonstrating improved performance by alleviating computational graph sparsity. Given the rapid rise of works in the GDA domain owing to their strong performance improvements, our tutorial aims to make a timely contribution by sensitizing the graph ML community towards this growing area of work, while (i) introducing background and motivation behind GDA, (ii) giving a bird's eye view of existing techniques, (iii) discussing key application areas, and (iv) discussing crucial challenges for future work in the space. We next introduce the content details, including the outline.

2 Tutorial Outline

1. Introduction and Background (10 mins.)
   (a) Machine Learning and Data Augmentation
   (b) Graph Machine Learning and Graph Neural Networks
2. Heuristic Graph Data Augmentation (25 mins.)
   (a) Data Removal (10 mins.)
      i. Node/Edge Dropping
      ii. Feature Masking
      iii. Subgraph Cropping
   (b) Data Addition (15 mins.)
      i. MixUp
      ii. Diffusion
      iii. Counterfactual Augmentation
3. Learned Graph Data Augmentation (30 mins.)
   (a) Graph Structure Learning (8 mins.)
   (b) Adversarial Training (8 mins.)
   (c) Graph Rationalization (7 mins.)
   (d) Automated Augmentation (7 mins.)
4. Break (10 mins.)
5. Graph Data Augmentation for Low-resource Learning (30 mins.)

(a) Self-supervised Learning on Graphs (15 mins.)
(b) Semi-supervised Learning on Graphs (15 mins.)

6. Future Directions and Conclusions (25 mins.)
   (a) Heterogeneity and Data Imbalance (6 mins.)
   (b) Domain Adaptation (6 mins.)
   (c) Efficiency and Scalability (6 mins.)
   (d) Theoretical Foundations (6 mins.)

Introduction and Background: We will introduce preliminaries required to understand the tutorial content, including a quick overview of DA in machine learning, graph ML, and GNNs. We will also present the structured outline and taxonomy for GDA methods.

Heuristic GDA Approaches: We will discuss several approaches in the context of heuristic GDA. These approaches involve stochastic or rule-based manipulation of node, edge, subgraph and feature data. We categorize these methods into

- **Data Removal:** Analogized from the commonly used augmentation techniques on computer vision, stochastic node/edge dropping [20, 7, 29], feature masking [29, 28], and subgraph cropping [23] methods provide efficient way of creating augmented graph data.
- **Data Addition:** To improve the generalization and robustness of GNNs, graph MixUp [24, 9], graph diffusion [14, 10] and counterfactual augmentation [32] methods are often used to create augmented graph data.

Learned GDA Approaches: A downside of heuristic methods is their agnosticism to downstream tasks. In this section, we will discuss learned augmentation methods, which learn to augment with intended downstream goals in mind. These approaches involve modifying feature and/or structure data in a paremeterized fashion. We categorize relevant approaches into

- **Graph Structure Learning:** These methods aim to learn optimal graph structure for downstream tasks, treating the structure as parameters to be learned [33, 13, 8, 2, 17, 35]. Recent studies have demonstrated that structure learning is helpful in improving model generalization [33, 2] and robustness [13, 35].
- **Adversarial Training:** Graph adversarial training [6, 4, 12, 3, 1, 15] augments input graphs with adversarial patterns during model training, so that the trained models can tolerate the adversarial perturbation in data and generalize to out-of-distribution samples at test time.
- **Graph Rationalization:** Graph rationalization methods learn the subgraphs that have causal relationships with the labels [27, 16].
- **Automated Augmentation:** Several recent works explored automated solutions for finding the best augmentation strategies [28, 18].
- **Self-supervised learning:** DA techniques are often used to help create self-supervision especially in graph contrastive learning [29, 28, 36], where one aims to maximize the similarity between different augmented graphs.
- **Semi-supervised learning:** Similar to self-supervised methods, augmentations are used to create different views of the unlabeled data, which can be used for training with consistency loss [25, 7] in addition to labels.

Future Directions and Conclusions: Despite the rapidly growing body of work in this area, there are many unanswered questions. We discuss four future directions: heterogeneity and data imbalance, domain adaptation, efficiency and scalability, and theoretical foundations.

3 Target Audience and Prerequisites

The target audience of this tutorial includes researchers and practitioners who are interested in graph machine learning, and particularly those that may work in low-resource settings where labeled data is difficult to collect (e.g. trust and safety domain). We expect the audience to be familiar with basic ML fundamentals – (un)supervised learning, common loss types and graph data structures (low familiarity with graph mining techniques is OK). Additional experience in deep learning (e.g. training neural models) is also helpful.

- **Estimated participants:** We anticipate 50+ participants.

4 History and Related Tutorials

While the specific tutorial scope proposed here has not been exactly tackled before, we strongly believe in our tutorial's prospects for success, for a few reasons:

• **Our team has given numerous relevant tutorials across the graph ML and low-resource and data efficient learning domains at top conferences.** Together, we have given 18 tutorials in areas including graph ML, low-resource and data efficient learning, natural language processing and more. We intend for our tutorial to take a more targeted focus on augmentation methods for graph learning , while drawing on of some of our related tutorials on "Graph Minimally Supervised Learning" (WSDM 2022, KDD 2022) and "Data Efficient Learning on Graphs" (KDD 2021).

Our team has also given several fundamental graph ML tutorials, including "Graph Neural Networks: Models and Applications" (AAAI 2020) and "Graph Representation Learning: Foundations, Methods, Applications and Systems" (KDD 2021).

• **Our team is pushing research at the frontier of the GDA domain.** Our previous foundational work [33] "DA in GNNs" was published in AAAI'21 and has already accumulated over 117 citations and is a flagship work in the domain. Our team has also co-authored two surveys in the area: [31] and [5]. We maintain the GDA papers page with 140+ stars. Moreover, one of our organizers is the lead PI for over $545K in funding across NSF, Snap and Amazon related to GDA.

• **Our team has considerable research and technical presentation experience in the graph ML and broader ML communities.** Our team has diverse levels of academic tenure: 1 Assistant Professor, 2 industrial Research Scientists, and 3 PhD students of varying seniority. Together, our team has published over 130 top-tier conference papers in venues including ICLR, NeurIPS, AAAI, WWW, KDD, WSDM, SDM and more. We have several best papers at these, and routinely serve as organizing committee members, speakers, metareviewers and authors across them.

5 Supplementary References

- Our past tutorials on "Graph Minimally Supervised Learning" (WSDM 2022, KDD 2022) and "Data Efficient Learning on Graphs" (KDD 2021).
- Our 2 past surveys on GDA: Zhao et al'22 [31], Ding et al'22 [5]
- Our curated reading list of GDA papers.[1]

6 Tutors

- **Tong Zhao** (tzhao@snap.com): Research Scientist at Snap Inc. Tong is a Research Scientist in the Computational Social Science group at Snap Research. His research focuses on graph machine learning, representation learning, and data augmentation methods on graphs. His work has resulted in 20+ conference and journal publications, in top venues such as ICML, KDD, AAAI, WWW, TNNLS, etc. Several of his works were pioneer studies of GDA, which is the topic of this tutorial (e.g., [33, 34, 32]). He is also the leading author of a survey paper on GDA [31].
- **Kaize Ding** (kding9@asu.edu): is currently a Ph.D. candidate at Arizona State University. His research interests are broadly in data mining and machine learning, with a particular focus on graph neural networks, and data-efficient learning. He has published over 30 papers on top conferences and journals such as WWW, WSDM, EMNLP, AAAI, and TNNLS. He is also the leading author of a GDA paper [5]. More details can be found at http://www.public.asu.edu/~kding9/.
- **Wei Jin** (jinwei2@msu.edu): Wei Jin is a Ph.D. student of Computer Science and Engineering at Michigan State University. He works in the area of graph learning and data-centric AI. He has published his research in top conference proceedings (e.g., ICLR, KDD, ICML, WSDM and AAAI). He has also served as (senior) program committee member at a number of these. He also presented multiple tutorials about graph neural network at top conferences such as AAAI and KDD.
- **Gang Liu** (gliu7@nd.edu): Gang Liu is a second-year Ph.D. student of Computer Science and Engineering at the University of Notre Dame. His research interests include graph machine learning and data augmentation. He has publications in top venues such as ICML and KDD on the topics of GDA as the leading or active author.
- **Meng Jiang** (mjiang2@nd.edu): Meng Jiang is an assistant professor at the Department of Computer Science and Engineering in the University of Notre Dame. His research interests include data mining, machine learning, and natural language processing. The awards he received include best paper finalist in KDD 2014, the best paper award in KDD-DLG workshop 2020, and ACM SIGSOFT Distinguished Paper Award in ICSE 2021. He received Notre Dame Faculty Research Award in 2019, NSF CRII Award in 2019, NSF III Grant (Comprehensive Methods to Learn to Augment Graph Data) in 2022, and NSF CAREER award in 2022.
- **Neil Shah** (nshah@snap.com): Neil Shah is a Lead Research Scientist and Manager at Snap Research, working on machine learning algorithms and applications on large-scale graph data. His work has resulted in 50+ conference and journal publications, in top venues such as ICLR, NeurIPS, KDD, WSDM, WWW, AAAI and more, including several best-paper awards. He has also served as an organizer, chair and senior program committee member at a number of these. He is the corresponding author of a key early pioneering work in GDA [33], several further GDA studies [34, 30], and many early works in unsupervised graph anomaly detection and applications [21, 11, 22].

[1] https://github.com/zhao-tong/graph-data-augmentation-papers

References

[1] J. Chen, X. Lin, H. Xiong, Y. Wu, H. Zheng, and Q. Xuan, *Smoothing adversarial training for gnn*, IEEE Transactions on Computational Social Systems, (2020).

[2] Y. Chen, L. Wu, and M. Zaki, *Iterative deep graph learning for graph neural networks: Better and robust node embeddings*, in NeurIPS, vol. 33, 2020, pp. 19314–19326.

[3] Q. Dai, X. Shen, L. Zhang, Q. Li, and D. Wang, *Adversarial training methods for network embedding*, in WWW, 2019, pp. 329–339.

[4] Z. Deng, Y. Dong, and J. Zhu, *Batch virtual adversarial training for graph convolutional networks*, arXiv preprint arXiv:1902.09192, (2019).

[5] K. Ding, Z. Xu, H. Tong, and H. Liu, *Data augmentation for deep graph learning: A survey*, arXiv preprint arXiv:2202.08235, (2022).

[6] F. Feng, X. He, J. Tang, and T.-S. Chua, *Graph adversarial training: Dynamically regularizing based on graph structure*, TKDE, 33 (2019), pp. 2493–2504.

[7] W. Feng, J. Zhang, Y. Dong, Y. Han, H. Luan, Q. Xu, Q. Yang, E. Kharlamov, and J. Tang, *Graph random neural networks for semi-supervised learning on graphs*, NeurIPS, 33 (2020).

[8] L. Franceschi, M. Niepert, M. Pontil, and X. He, *Learning discrete structures for graph neural networks*, in ICML, 2019, pp. 1972–1982.

[9] X. Han, Z. Jiang, N. Liu, and X. Hu, *G-mixup: Graph data augmentation for graph classification*, in ICML, 2022.

[10] K. Hassani and A. H. Khasahmadi, *Contrastive multi-view representation learning on graphs*, in ICML, 2020.

[11] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, *Fraudar: Bounding graph fraud in the face of camouflage*, in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 895–904.

[12] W. Hu, C. Chen, Y. Chang, Z. Zheng, and Y. Du, *Robust graph convolutional networks with directional graph adversarial training*, Applied Intelligence, (2021).

[13] W. Jin, Y. Ma, X. Liu, X. Tang, S. Wang, and J. Tang, *Graph structure learning for robust graph neural networks*, in KDD, 2020, pp. 66–74.

[14] J. Klicpera, S. Weissenberger, and S. Günnemann, *Diffusion improves graph learning*, NeurIPS, 32 (2019).

[15] K. Kong, G. Li, M. Ding, Z. Wu, C. Zhu, B. Ghanem, G. Taylor, and T. Goldstein, *Robust optimization as data augmentation for large-scale graphs*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 60–69.

[16] G. Liu, T. Zhao, J. Xu, T. Luo, and M. Jiang, *Graph rationalization with environment-based augmentations*, in KDD, 2022.

[17] D. Luo, W. Cheng, W. Yu, B. Zong, J. Ni, H. Chen, and X. Zhang, *Learning to drop: Robust graph neural network via topological denoising*, in WSDM, 2021, pp. 779–787.

[18] Y. Luo, M. McThrow, W. Y. Au, T. Komikado, K. Uchino, K. Maruhash, and S. Ji, *Automated data augmentations for graph classification*, arXiv preprint arXiv:2202.13248, (2022).

[19] J. Palowitch, A. Tsitsulin, B. Mayer, and B. Perozzi, *Graphworld: Fake graphs bring real insights for gnns*, KDD, (2022).

[20] Y. Rong, W. Huang, T. Xu, and J. Huang, *Dropedge: Towards deep graph convolutional networks on node classification*, arXiv preprint arXiv:1907.10903, (2019).

[21] N. Shah, A. Beutel, B. Gallagher, and C. Faloutsos, *Spotting suspicious link behavior with fbox: An adversarial perspective*, in 2014 IEEE International conference on data mining, IEEE, 2014, pp. 959–964.

[22] N. Shah, A. Beutel, B. Hooi, L. Akoglu, S. Gunnemann, D. Makhija, M. Kumar, and C. Faloutsos, *Edgecentric: Anomaly detection in edge-attributed networks*, in 2016 IEEE 16Th international conference on data mining workshops (ICDMW), IEEE, 2016, pp. 327–334.

[23] Y. Wang, W. Wang, Y. Liang, Y. Cai, and B. Hooi, *Graphcrop: Subgraph cropping for graph classification*, arXiv preprint arXiv:2009.10564, (2020).

[24] ———, *Mixup for node and graph classification*, in WWW, 2021.

[25] Y. Wang, W. Wang, Y. Liang, Y. Cai, J. Liu, and B. Hooi, *Nodeaug: Semi-supervised node classification with data augmentation*, in KDD, 2020, pp. 207–217.

[26] M. Welling and T. N. Kipf, *Semi-supervised classification with graph convolutional networks*, in ICLR, 2017.

[27] Y. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, *Discovering invariant rationales for graph neural networks*, in ICLR, 2021.

[28] Y. You, T. Chen, Y. Shen, and Z. Wang, *Graph contrastive learning automated*, arXiv preprint arXiv:2106.07594, (2021).

[29] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, *Graph contrastive learning with augmentations*, NeurIPS, 33 (2020).

[30] T. Zhao, T. Jiang, N. Shah, and M. Jiang, *A synergistic approach for graph anomaly detection with pattern mining and feature learning*, IEEE Transactions on Neural Networks and Learning Systems, (2021).

[31] T. Zhao, G. Liu, S. Günnemann, and M. Jiang, *Graph data augmentation for graph machine learning: A survey*, arXiv preprint arXiv:2202.08871, (2022).

[32] T. Zhao, G. Liu, D. Wang, W. Yu, and M. Jiang, *Learning from counterfactual links for link prediction*, in ICML, 2022, pp. 26911–26926.

[33] T. Zhao, Y. Liu, L. Neves, O. Woodford, M. Jiang, and N. Shah, *Data augmentation for graph neural networks*, in AAAI, vol. 35, 2021, pp. 11015–

11023.
[34] T. Zhao, B. Ni, W. Yu, Z. Guo, N. Shah, and M. Jiang, *Action sequence augmentation for early graph-based anomaly detection*, in CIKM, 2021.
[35] C. Zheng, B. Zong, W. Cheng, D. Song, J. Ni, W. Yu, H. Chen, and W. Wang, *Robust graph representation learning via neural sparsification*, in ICML, 2020, pp. 11458–11468.
[36] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, *Graph contrastive learning with adaptive augmentation*, in WWW, 2021, pp. 2069–2080.