Boosting GNN's Generalization via Graph Data Augmentation
CS249 Project Proposal (Spring 2022)

Siddhant Patil Dipti Ranjan Sahu Nischal Reddy Chandra Shivam Patel

Abstract

Data augmentation, creating new plausible variations from the original training data, has been proved effective in improving the generalization of machine learning models on images and natural language. However, comparatively little work studies data augmentation for graphs (Zhao et al., 2020) and (Ding et al., 2022). Augmentation operations commonly used in vision and language cannot be naturally applied to graphs primarily because of two reasons. Firstly, graph is a more complicated data abstract, i.e. the graph structure is not euclidean while an image is. Secondly, the nodes in graphs are described by some (high-dimensional) characteristics while an image pixel is just scalars. This project aims to explore how to generate reasonable data on graphs that can boost the generalization of graph neural networks.

1. Problem Statement

The performance of most ML models, and deep learning models in particular, depends on the quality, quantity and relevancy of training data. However, insufficient data is one of the most common challenges in implementing machine learning. Data Augmentation is a technique that can be used to artificially expand the size of a training set by creating modified data from the existing one. It acts as a regularizer and helps reduce overfitting when training a machine learning model. It is closely related to oversampling in data analysis.

Data Augmentation have seen widespread adoption in fields such as computer vision (Shorten & Khoshgoftaar, 2019) and natural language processing (Feng et al., 2021). At the same time, graph neural networks (GNNs) have emerged as a rising approach for datadriven inference on graphs, achieving promising results on tasks such as node classification, link prediction and graph representation learning.

Despite the complementary nature of GNNs and data augmentation, few works present strategies for combining the two. One major obstacle is that, in contrast to other data, where structure is encoded by position, the structure of graphs is encoded by node connectivity, which is irregular. The hand-crafted, structured, data augmentation operations used frequently in CV and NLP therefore cannot be applied. Furthermore, this irregularity does not lend itself to easily defining new augmentation strategies.

Based on the components of interest, the graph augmentation can be divided into four categories.

- Node augmentation: adding new nodes or removing existing nodes
- Edge augmentation: adding new links or removing existing links
- Graph augmentation: adding new graphs or removing existing graphs (only for graph-level tasks such as graph classification)
- Feature augmentation: modifying the node features

For our project, we will try to innovate and improve on top of existing solutions. We will do this by trying to augment the model with other techniques and algorithms we have learned in lecture. We will include one (or a combination of several) above perspectives.

2. Datasets

We plan to use the following datasets.

- Citation graphs:
  - Cora: The Cora dataset consists of 2708 scientific publications classified into one of seven classes. The citation network consists of 5429 links.
  - Citeceer: The CiteSeer dataset consists of 3312 scientific publications classified into one of six classes. The citation network consists of 4732 links.
  - Pubmed: The Pubmed dataset consists of 19717 scientific publications pertaining to diabetes classified into one of three classes. The citation network consists of 44338 links.
- Social networks:
  - Flickr: Flickr dataset consists of 514k nodes and 3.2M edges that comprise results obtained from crawl of the Flickr photo-sharing social network from May 2006.
  - BlogCatalog: BlogCatalog is a graph dataset consisting of 88K nodes and 2.1M edges for a network of social relationships of bloggers listed in the BlogCatalog website.

- Information Networks:
  - IMDB: IMDb Movie Reviews dataset is a binary sentiment analysis dataset having 896.3K nodes and edges 3.8M. It consists of 50K reviews from the Internet Movie Database (IMDb) labeled as positive or negative.
  - Proteins: The database consists of proteins that are classified as enzymes or non-enzymes.

3. Solution Plan

We plan on starting with node augmentation and edge augmentation. Later on, we will move to feature augmentation. More specifically, we will keep the graph structure as learnable and optimize the structure through feedback signals from the downstream task. (Zheng et al., 2020) and (Luo et al., 2021) present some examples of this. On top of this, we also plan to use feature augmentation - we will perturb the features such that they improve the downstream tasks. One of the representative work is (Xu & Tong, 2021).

As described in (Zhao et al., 2022), augmentation strategies can be classified into two types: *task-dependent* and *task-independent* augmentations. In this project, we are going to focus on task-dependent augmentations, wherein we use these augmentations to improve the performance of a downstream task like node classification. The objective of a task-dependent augmentation can be written as:

$$\min_{\theta,\phi} \quad L_{all}(\{G_i\}, \{f_\theta(G_i)\}, \phi)$$

where θ and ϕ are the parameters of the augmentation and downstream task respectively, where as L_{all} is the merged augmentation and downstream objective.

There is a list of potential ideas for graph augmentation. We will check the feasibility of these ideas over time and learn new techniques as the class progresses.

4. Evaluation

Based on the given tasks at hand, it is evident that the majority of downstream tasks fall under the umbrella of classification. Hence, we plan to use the following evaluation metrics pertaining to classification primarily.

- F1-Score: Since we plan on evaluating the augmentation technique on task like node classification we would compute the accuracy of the model via augmentation followed by computing an F1 score.
- BCE Loss: We plan on using the Binary Cross Entropy loss function which helps evaluate the true and the predicted node labels

Some metric we plan on viewing include **Frechet Distance** as well as **Affinity and Diversity**, which together could help evaluate Data Augmentation. Affinity helps determine how much an augmentation shifts the training data distribution and diversity explains the complexity of the augmented sample for model and learning schedule

5. Potential Schedule

- **Week 5** - Complete Literature Review & Dataset Collection
- **Week 6** - Data cleaning, Data split, and Implementation of Baseline models
- **Week 7** - Run Baseline models, Setup Evaluation metrics
- **Week 8, 9** - Try out new innovative ideas
- **Week 10** - Summarise our experimental results and document our findings

Workload distribution: As the timeline is divided over weeks, all the team members will contribute equally in all the phases. Currently, all of us have picked up a paper for literature review and will implement the baseline models.

References

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

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Luo, D., Cheng, W., Yu, W., Zong, B., Ni, J., Chen, H., and Zhang, X. Learning to drop: Robust graph neural network via topological denoising. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Xu, Z. and Tong, H. Graph sanitation with application to node classification. *CoRR*, abs/2105.09384, 2021. URL https://arxiv.org/abs/2105.09384.

Zhao, T., Liu, Y., Neves, L., Woodford, O., Jiang, M., and Shah, N. Data augmentation for graph neural networks. *arXiv preprint arXiv:2006.06830*, 2020.

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

Zheng, C., Zong, B., Cheng, W., Song, D., Ni, J., Yu, W., Chen, H., and Wang, W. Robust graph representation learning via neural sparsification. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11458–11468. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zheng20d.html.