Adversarial Vulnerability of Graph Convolutional Networks

Samaneh Shamshiri Insoo Sohn

Division of Electronics and Electrical Engineering
Dongguk University

Samaneh.shamshiri@gmail.com isohn@dongguk.edu

Abstract

In recent years, graph neural networks (GNNs) have significantly advanced applications such as drug design, medical diagnoses, social network analysis, recommendation systems, and fraud detection. However, even state-of-the-art GNNs are vulnerable to adversarial attacks. These attacks exploit the model's performance by small crafted perturbations through manipulating the input graph (nodes, edges, or features) to mislead the model into making incorrect predictions or classifications. This paper reviews challenges posed by adversarial attack methods GNNs, and demonstrates the impact of two benchmark attacks; Nettack and Metattack on medical dataset.

1. Introduction

The vulnerability of traditional machine learning models to data perturbations is well-established [1]. Even minor modifications to the input can result in incorrect predictions. These perturbations, nearly indistinguishable from the original data to humans, are known as adversarial examples and can lead to misclassification. One of the most well-known examples is when a neural network misclassifies a stop sign as a speed limit sign due to subtle changes to the image, even though it still clearly appears as a stop sign to human observers [2]. Such examples highlight how machine learning models can fail dramatically in the face of adversarial perturbations, raising concerns about their use in safety-critical or scientific applications. As a result, researchers have increasingly focused on assessing the robustness of models across various domains, such as images, natural language, and speech. Recently, attention has shifted to the security concerns of Graph Neural Networks (GNNs). GNNs have become prominent in many real-world applications, representing complex relationships in systems like social networks, e-commerce platforms, biological networks, and traffic systems. For instance, GNNs are used in social networks for community detection and recommendation systems, in drug discovery for predicting molecular interactions, and in fraud detection for identifying suspicious activities. The robustness of GNNs in these areas is critical, as adversarial attacks could lead to severe consequences. For example, in financial networks, an adversarial attack could misclassify fraudulent transactions as legitimate, leading to financial loss. Similarly, in healthcare, an attack could result in incorrect predictions in drug design, potentially endangering patient safety [3]. Initial studies on GNNs' robustness [5] revealed their vulnerability to adversarial perturbations, particularly in node-level classification tasks. Adversarial attacks on GNNs can involve slight alterations to the graph structure, such as adding or removing edges, or modifying node features, which can significantly degrade the model's performance. This vulnerability is particularly concerning given the increasing deployment of GNNs in high-stakes environments.

Among the various types of GNNs, Graph Convolutional Networks (GCNs) [4] have gained significant attention due to their design tailored specifically for analyzing graphs. These state-of-the-art GCNs utilize a "message-passing" process, where nodes gather information from their neighbors at each convolutional layer. While GCNs have demonstrated strong performance in tasks like node classification and other graph analysis applications, they are not immune to adversarial attacks. Notably, Nettack [5] and Metattack [6] are two benchmark adversarial attacks that have been shown to significantly affect GCN performance. This paper focuses on evaluating the impact of Nettack and Metattack on GCNs. We examine how these attacks manipulate node features and graph structures to degrade GCN performance, and discuss the implications of these vulnerabilities in real-world applications.

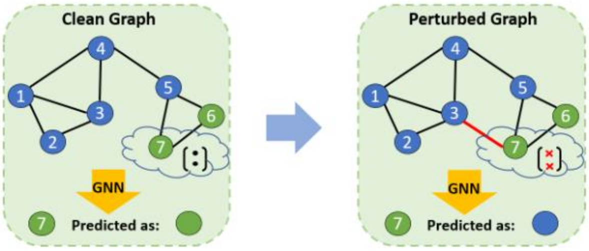

Figure 1: Adversarial attacks on GCN

2.Methodology of Attacks

Adversarial attacks on GCNs aim to cause incorrect predictions for node classification by introducing perturbations, such as adding or deleting edges and altering node features, as illustrated in Figure 1. Among the existing methods, Nettack is recognized as the state-of-the-art adversarial attack approach. The central idea of Nettack is to maximize the GCN's classification loss (i.e., the difference in classification outcomes between the original and modified GCN models) on the target node by applying perturbations within a defined perturbation space. Nettack is particularly effective because it generates nearly imperceptible perturbations by preserving the degree.

In addition to Nettack, Metattack is another well-known adversarial attack specifically designed for Graph Neural Networks (GNNs). Unlike Nettack, which focuses on targeted node-level attacks, Metattack is designed to perturb the graph globally. It achieves this by generating poisoning attacks using a meta-learning framework. This method allows the attacker to disrupt the entire graph structure, leading to widespread misclassifications across the network. Metattack's global approach makes it particularly dangerous in scenarios where maintaining the integrity of the entire graph is crucial, such as in social networks or recommendation systems.

3. Experimental Analysis

First we trained the GCN on raw data without attack. In this work we used Mutag dataset which is a collection of nitroaromatic compounds that have been gathered to predict their mutagenicity on Salmonella typhimurium. There are 188 graphs in Mutag dataset with average number of 17.93 nodes and 19.79 edge for binary classification tasks. We used ReLU as the non-linear activation function for the GCN and SGD as the optimizer. The learning rate was set to 0.01, and the momentum to 0.9. The accuracy of the GCN on benign data before attack is 84.21%.

Following the experimental setup, we generate perturbed dataset by attacking GCN using Nettack, and Mettack. Figure 2 depicts the experimental results on perturbed GCN with Nettack and Metattack respectively.

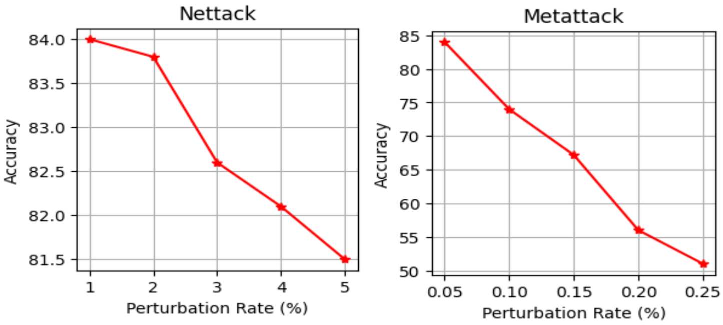

Figure 2. Experimental results according to perturbation rate on perturbed datasets using Nettack, Metattack, and respectively. The x-axis is the perturbation rate, and the y-axis is the accuracy.

As can be seen in Figure 2 the slope of performance decline as the attack strength increases. In terms of Nettack, the GCN's accuracy starts at around 84% and decreases steadily with increasing perturbation rates. Even at a 5% perturbation rate, the accuracy only drops by a few percentage points, indicating that Nettack requires a higher perturbation rate to cause significant degradation in performance. The decrease in accuracy is relatively modest, showing that Nettack's effect is more gradual. This could imply that Nettack may be less aggressive but can still steadily degrade performance with increasing perturbations. On the other hand, with Metattack, the GCN starts at a higher accuracy of 85%, but the accuracy drops dramatically with very small increases in the perturbation rate. By the time the perturbation rate reaches 0.25%, the accuracy has plummeted to around 50%. This suggests that Metattack is much more effective at disrupting the GCN with minimal perturbations. The significant drop in accuracy (from 85% to 50%) within a small perturbation range indicates that Metattack is highly effective at quickly degrading the GCN's performance. This suggests that Metattack is a more potent adversarial attack compared to Nettack, particularly in environments where even small perturbations can be devastating.

4. Conclusion

While GCNs have demonstrated impressive performance on various graph-related tasks, their vulnerability to adversarial attacks remains a significant concern. In this study, we assessed the impact of two state-of-the-art attack methods, Nettack and Metattack, on the performance of GCNs. Our experiments revealed that although the GCN achieved an accuracy of 84.21% on the Mutag dataset, this accuracy was substantially reduced when subjected to different levels of perturbation. These results underscore the need for robust defense mechanisms against adversarial attacks in GCNs. To address this, recent research has proposed several advanced defense algorithms. Looking ahead, our future work will focus on enhancing the robustness of GCNs by leveraging complex network structures, such as scale-free networks, to rewire dormant edges and mitigate the impact of adversarial attacks.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00252328).

References

[1] I. J. Goodfellow, J. Shlens, and Ch. Szegedy, “Explaining and harnessing adversarial examples”. In International Conference on Learning Representations (ICLR), 2015.

[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song. “Robust physical-world attacks on deep learning visual classification”. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018.

[3] S. Shamshiri, K. J. Han, and I. Sohn, DB COVIDNet: A Defense Method against Backdoor attacks, Mathematics , 2023

[4] M. Welling, T.N. Kipf, “Semi-supervised classification with graph convolutional networks” In Proceedings of the J. International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2016.

[5] D. Zugner, A. Akbarnejad, and S. G ¨ unnemann, “Adversarial attacks ¨ on neural networks for graph data,” in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2847–2856.

[6] D. Zügner, S. Günnemann, “Adversarial attacks on graph neural networks via meta learning”, in: Proceedings of the 7th International Conference on Learning Representations, 2019.