

SBA-ST: 一种使用更小触发器逃避随机平滑防御的子图后门攻击方法

伍晓洁¹, 刘 强¹, 王煜恒¹, 付章杰²

¹国防科技大学 计算机学院 长沙 中国 410073

²南京信息工程大学 计算机学院/网络空间安全学院 南京 中国 210044

摘要 随着图神经网络在节点分类和图分类任务领域的快速发展,对其安全脆弱性的研究也越来越深入。在图分类领域的对抗性攻击中,越来越多的研究尝试在图上附加恶意触发器以实施有效的后门攻击,中毒图参与训练后生成的后门模型能够将添加有触发器的图错误预测为攻击者指定的目标类别。随机平滑防御是一种提升图学习模型鲁棒性的有效方法,能消除有限触发器大小条件下的后门攻击影响。当攻击者使用较大的触发器时,虽然后门攻击能够经验性地逃避随机平滑防御,但是易于被检测。因此,如何使用尽可能小的触发器来达到可观的攻击性能,仍然是一个具有挑战性的问题。本文提出了一种使用更小触发器逃避随机平滑防御的子图后门攻击方法(简称 SBA-ST),该方法以子图触发器的形式添加扰动,通过干扰对图分类任务影响度较高的节点,使得在随机平滑防御开启的情况下仍能够保持较强的攻击能力。具体来讲, SBA-ST 引入图注意力网络(GAT)模型和高斯混合模型(GMM)聚类分析以设计一个最佳后门注入位置选择机制。此外,该方法采用 Erdős-Rényi(ER)随机图生成模型以降低子图后门触发器生成的计算复杂度,其中,模型的参数为节点数量和边生成概率。在 5 个公开数据集上的对比实验结果表明, SBA-ST 获得了比 SBA 方法更高的后门分类准确率和平均攻击成功率,并能够使用比 SBA 明显更小的触发器来获得较小的后门攻击性能损失,从而验证了本文方法更佳的随机平滑防御逃避能力。

关键词 图对抗学习; 子图后门攻击; 图分类; 随机平滑认证

中图法分类号 TP391 DOI号 10.19363/J.cnki.cn10-1380/tn.2025.11.06

SBA-ST: A Subgraph Backdoor Attacking Method Using Smaller Triggers to Evade Randomized Smoothing Defense

WU Xiaojie¹, LIU Qiang¹, WANG Yuheng¹, FU Zhangjie²

¹ College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China

² School of Computer Science/ School of Cyber Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China

Abstract As graph neural networks (GNNs) rapidly advance in node classification and graph classification tasks, research into their security vulnerabilities has deepened. In the realm of adversarial attacks on graph classification, an increasing number of studies have attempted to implement effective backdoor attacks by attaching malicious triggers to graphs. The poisoned graphs, once included in training, lead the resulting backdoor models to misclassify trigger-containing graphs into attacker-specified target categories. Randomized smoothing defense is an effective method to enhance the robustness of graph learning models and it can eliminate the effects of backdoor attacks with limited trigger sizes. When adversaries use a large trigger size, although backdoor attacks can empirically evade the randomized smoothing defense mechanism, they are weak to be detected. Hence, how to achieve considerable attacking performance using a small enough trigger is still a challenging problem. In this paper, we propose a Subgraph Backdoor Attacking method using Smaller Triggers to evade randomized smoothing defense called SBA-ST. The proposed method adds disturbance in the form of a subgraph trigger and disturbs the nodes that have high impacts on the graph classification task, so that it can maintain strong attacking capability even when the randomized smoothing defense is enabled. Specifically, SBA-ST introduces graph attention network (GAT) model and Gaussian mixture model clustering analysis to design a selection mechanism of the optimal backdoor injection position. Furthermore, the proposed method utilizes the Erdős-Rényi (ER) random graph generation model, the parameters of which are the number of nodes and the probability of edge generation, to reduce the computational complexity of generating subgraph triggers. Comparative experiments over five public datasets

通信作者: 刘强, 博士, 副教授, Email: qiangliu06@nudt.edu.cn.

本课题得到国家重点研发计划项目“科技创新 2030”(No. 2022ZD0209105)和湖南省自然科学基金项目(No. 2021JJ30779)资助。

收稿日期: 2024-02-07; 修改日期: 2024-05-31; 定稿日期: 2025-10-15

show that, the SBA-ST method outperforms SBA in terms of backdoor classification accuracy and average success rate. Moreover, SBA-ST gains relatively small backdoor attacking performance loss with a significantly smaller trigger size compared to SBA, which demonstrates better evading ability of the proposed method in front of randomized smoothing defense.

Key words graph adversarial learning; subgraph backdoor attack; graph classification; randomized smoothing certification

1 引言

随着图数据结构被广泛应用于建模现实世界中的关系, 例如社交网络、蛋白质结构分析、通信网络等, 基于图的应用方兴未艾, 如社交异常账号检测、分子分类、社区检测、链接预测等。图神经网络(Graph Neural Networks, GNN)通过消息聚合机制能够有效地学习一个图表示, 该机制通过聚合来自一个节点的所有邻居节点的信息来更新其节点特征表示, 从而有效地保留了节点特征、邻居节点特征和局部图拓扑结构信息。因此, GNN 在节点分类、图分类、链接预测等任务上的性能表现优于传统机器学习技术^[1-3]。例如, 图卷积神经网络(Graph Convolutional Networks, GCN)利用邻接矩阵进行特征传播来学习节点表示。

然而, 图结构数据自身被证明易于遭受攻击^[3], 其中, 攻击者可轻易操纵图中攻击目标节点的局部特征结构, 如社交网络中的恶意账户或虚假链接, 从而改变图的属性, 进而影响 GNN 的性能。此外, GNN 也有经典深度神经网络固有的安全缺陷, 攻击者可以对图结构或节点特征施加不易察觉的扰动, 随后利用 GNN 模型的消息聚合机制来传播和聚合恶意信息, 导致出现了严重的安全隐患^[4]。

大多数现有针对 GNN 的对抗攻击研究主要从节点注入^[5,6]、增删边^[7-10]、边重接^[11]等方面进行恶意信息注入, 以此来扰动图结构或节点特征。虽然针对图像数据的后门攻击技术已取得了一定的进展^[12-14], 但是在图结构数据领域中的相关研究尚处于起步阶段^[15]。GNN 后门攻击旨在通过向训练数据集投毒并训练得到一个后门 GNN 模型, 当测试图携带有特定的结构这一后门触发条件时, 后门模型会将测试图误判为指定错误类别。类似于图像数据中使用子像素模块作为后门触发器, 图结构数据的后门触发器通常是子图结构^[16-19]。子图后门攻击(Subgraph Backdoor Attack, SBA^[16])被首次提出时, 是利用简单的随机图生成方法产生子图后门触发器并随机选择位置注入到训练图数据中, 再利用被投毒的数据集训练得到后门模型, 进而影响模型在图分类任务上的性能。然而, 在触发器大小有限的情况下, SBA 可被基于平滑的方法^[16]有效地防御, 该方法的工作原

理是对训练图进行多次随机下采样, 由于原图中的恶意节点保留在下采样子图中的概率远小于正常节点的概率, 使得随机子采样能够以较大的概率过滤掉被注入的后门对预测结果带来的不利影响, 再结合随机平滑 GNN 分类器训练过程中采用的投票原则, 使其对扰动具备一定的鲁棒性。另外, 触发器的生成还可以与模型优化过程相结合, 例如, GTA 方法^[20]通过形式化触发器生成目标函数与攻击目标函数, 并利用双层优化机制进行优化, 以自适应地生成具有最佳攻击性能的子图触发器。然而, 自适应触发器生成与模型训练联合优化需要计算大量参数, 高计算开销导致其在大规模图上进行攻击变得困难。TRAP 方法^[4]基于梯度计算为每个攻击对象生成自适应的触发器, 虽然能够获得比较理想的攻击性能, 但该方法在生成触发器时需要评估每条候选边对图分类任务的恶意程度, 计算复杂度较高; 此外, 实验结果表明该方法在随机平滑防御机制面前性能也下降明显。综上所述, 当攻击者使用较大的触发器发起后门攻击时, 虽然它能够经验性地逃避随机平滑防御, 但是易于被检测。因此, 如何使用尽可能小的触发器来达到可观的攻击性能, 仍然是一个具有挑战性的问题。

为了解决这一挑战, 本文扩展了 SBA 方法并提出了一种使用更小触发器逃避随机平滑防御的子图后门攻击方法(简称 SBA-ST)。为了突破 SBA 方法中由随机选择触发器注入位置而造成的攻击性能瓶颈, 本文设计了一个最佳后门注入位置选择机制, 该机制基于图注意力网络(Graph Attention Network, GAT)模型^[21], 利用节点的标签信息学习其在隐藏层的嵌入表示, 在部分没有节点标签的数据集中将节点度看作节点标签。由于 GAT 模型中使用了通用的自注意力网络, 它能够更好地将节点的局部和全局特征学习到节点表示向量中。此外, 为了找到对图分类任务影响度较高的节点, SBA-ST 方法在学习到的高维隐藏层嵌入表示上使用高斯混合模型(Gaussian Mixture Model, GMM)聚类方法进行聚类, 离聚类中心越近的节点, 其对分类结果的影响度将越高。在此基础上, 该机制综合考虑了节点自身重要性以及距聚类中心距离来选择后门注入位置, 以平衡攻击能力和隐蔽性。为了减小计算复杂度, SBA-ST 进一步

使用了 Erdős-Rényi(ER)随机图生成模型产生子图后门触发器,其中,模型的参数为节点数量和节点间边生成的概率。多组对比实验结果表明, SBA-ST 方法具有更强的后门攻击能力,并能够使用比 SBA 更小的触发器来逃避随机平滑防御。

综上,本文的主要贡献如下。

1) 提出了一种使用更小触发器逃避随机平滑防御的子图后门攻击方法 SBA-ST,该方法以子图触发器的形式添加扰动,通过干扰对图分类任务影响度较高的节点,使得在随机平滑防御开启的情况下仍能够保持较强的攻击能力;

2) SBA-ST 方法引入 GAT 模型捕捉节点的局部和全局特征,从而学习出高质量节点嵌入表示,通过 GMM 聚类分析,筛选出聚类中心附近重要性更大的节点,进而设计了一个最佳后门注入位置选择机制。此外,该方法采用 ER 随机图生成模型降低子图后门触发器生成的计算复杂度。

3) 基于 5 个公开数据集上的多组对比实验结果表明, SBA-ST 获得了比 SBA 方法具有更高的后门分类准确率和平均攻击成功率,并能够使用比 SBA 明显更小的子图触发器来获得较小后门攻击性能损失,从而验证了本文方法更佳的随机平滑防御逃避能力。

2 相关工作

图神经网络近些年来取得了广泛的关注并在许多领域发挥了重要的作用。然而, GNN 与其他深度神经网络类似,被证明同样具有脆弱性,易受到恶意攻击的影响。本节主要介绍广泛流行在图结构数据上的对抗攻击和最近兴起的图后门攻击的相关内容。

2.1 图对抗攻击

对抗攻击的思想在计算机视觉和自然语言处理领域^[22]早已被广泛运用,攻击者在原始数据中添加不可察觉的微小扰动形成对抗样本以降低训练模型的性能^[23-24],具有代表性的方法包括 PGD^[25]、JSMA^[26]、FGSM^[27]等。在图深度学习逐渐兴起的这些年里,研究者发现了图神经网络对对抗攻击所具有的脆弱性,掀起了在节点分类任务、图分类任务、链接预测、图表示学习等领域关于对抗攻击的研究热潮。攻击者通常会在白盒条件或黑盒条件^[28]下,扰动图结构数据(添加或删除边),或者节点特征(篡改节点特征),或者同时扰动图结构和节点特征^[7,29]。在白盒条件下,攻击者知道目标模型参数并可获得训练集数据。在黑盒条件下,攻击者不知道目标 GNN 模型的结构和参数,只能对一小部分训练数据集进行投毒。

许多工作^[7,9,29]研究了 GNN 针对对抗攻击的脆

弱性,攻击者通过扰动图结构或扰动节点特征,或者结合两种扰动方法,破坏 GNN 在节点分类、图分类、链接预测等任务上的性能。根据任务属于归纳学习或直推式学习的类别,可以在训练阶段或者推理阶段进行对抗性攻击,使用对抗样本来降低 GNN 的性能。例如, CLGA^[30]提出一种无监督的基于梯度上升的节点嵌入无监督对抗攻击方法,充分利用由于图对比学习模型生成的多个视图,在不知道它们的标签的情况下选择信息量最大的边,从而支持模型适应更多的任务(节点分类、链路预测等)。SRLIM^[31]认为在黑盒攻击条件下,代理模型应主要保持嵌入层和输入节点的拓扑一致性,于是提出基于等距映射的代理表示学习,以此保留代理嵌入中的拓扑结构。

基于梯度的对抗攻击也被应用于图领域, Gradient-based backdoors^[32]就将现有的梯度攻击方法扩展到图数据结构中,而 QEGC^[7]采用符号随机梯度下降(signSGD)方法来解决重新表述的攻击问题。根据攻击发生的阶段,还可以将对抗攻击分为投毒攻击和逃逸攻击,投毒攻击在训练 GNN 之前对训练图进行扰动,使得 GNN 在测试图上的预测性能下降,如 Nettack^[33]攻击。逃逸攻击在测试阶段加入扰动,此时 GNN 模型已被训练好且不能被攻击者修改,大多发生在黑盒攻击场景。

2.2 图后门攻击

后门攻击通过将后门触发器附加在原始训练数据上,并将其标签更改为攻击者指定的目标标签。使用中毒的训练数据集训练模型,在测试阶段,只要数据样本中具有后门触发器的结构,后门模型就会将该样本标签预测为攻击者指定的目标标签。触发器通常是一个足够小的子图,可以根据图数据精心制作^[4](密集连接的 Erdős-Rényi 图)或者通过触发器生成器产生^[16]。触发器也可以不具有特定的形状,通过不可察觉的扰动开销进行限制和优化。SBA^[16]主要依赖于向训练集中任意添加错误标记的输入,这增加了从输入数据中被检测出投毒的风险,同时也易被某些 GNN 解释工具检测到,因此, clean-label 后门攻击^[34]对具有目标标签的训练数据集子集进行采样,并在不更改其标签的情况下将触发器注入其中,这使得攻击者不依赖于有错误标签的输入,从而增加了后门攻击的隐蔽性。除了触发器自身的结构会影响后门攻击的效果,触发器的不同注入方法也是一个值得探索的问题。SBA 以替换的方式进行后门触发器注入,在目标投毒图数据中随机选取 t 个节点(触发器节点数量为 t),将该 t 个节点的原始属性与结构替换成触发器具有的属性与结构。Link-Backdoor^[17]攻击的

是链接预测结果, 它选取目标链接的两个节点为锚点, 作为触发器的注入位置。还有研究者利用 GNN 可解释性方法来快速选择后门触发器的最佳注入位置^[18], 利用 GNNExplainer^[19]来分析后门攻击的触发器注入位置对图分类任务中 GNN 的影响, 利用图的局部可解释性模型 GraphLIME^[35], 通过修改节点特征的不同子集来探索对节点分类任务的后门攻击最有效的特征子集。UGBA^[36]则通过 GCN 学习节点表示, 并通过聚类计算选择出投毒的节点, 在这些节点上自适应地生成后门触发器以攻击节点分类任务的准确性。

2.3 随机平滑

随机平滑作为提升机器学习模型对抗性攻击鲁棒性的先进技术, 通过对模型输入施加随机噪声, 利用含噪声输入的模型输出进行决策。此技术提供的概率保证赋予了其较传统防御机制更广的适应性和更高的安全性^[37], 而其简洁性又使其易于适用于不同模型与任务, 无须对原始架构做大幅修改。

随机平滑在图像、文本以及图结构数据等多种数据类型的对抗性防御中展示了其有效性。例如, 在自然语言处理领域, 通过在词嵌入空间引入噪声, 它可以抵御词级攻击^[38]; 在语音识别领域, 随机平滑已被用于提高模型对添加到输入语音信号中的微小扰动的鲁棒性^[39]。Salman 等人^[40]探索了通过改进噪声模型和聚合策略来提升随机平滑的效果和效率, 表明随机平滑技术在深度学习模型的防御体系中扮演着越来越重要的角色。Yan 等人^[41]提出了分层随机平滑, 使用不同的噪声分布初始化分层平滑, 对离散和连续域产生新的鲁棒性认证, 通过实验证明了分层平滑在图像和节点分类任务中能够保持模型较高的准确率并获得更具鲁棒性的认证。随着随机平滑不断发展, 其在对抗性鲁棒性方面的潜力得到了广泛认可, 为深度学习模型对抗复杂攻击提供了强有力的工具。

3 问题描述及形式化

3.1 威胁模型

给定一个包含 N 个图的带标签原始图数据集 $\mathcal{D} = \{(G_1, y_1), (G_2, y_2), \dots, (G_N, y_N)\}$, 其中 y_i 表示图 G_i 对应的标签, 且 $y_i \in \{1, 2, \dots, c\}$, $i = 1, \dots, N$ 。数据集 $\mathcal{D}$ 中的图均为属性图, 可表示为 $G = (V, E, X)$, 其中 $V = \{v_1, \dots, v_n\}$ 是规模为 n 的节点集, $E \subseteq V \times V$ 是边集, $X = \{x_1, \dots, x_n\}$ 是节点属性集合, x_i 是节点 v_i 的属性向量。图 G 的邻接矩阵为 $A \in R^{n \times n}$, 当节点

v_i 与节点 v_j 相连时 $A_{ij} = 1$, 否则 $A_{ij} = 0$ 。因此, 针对图分类问题的后门攻击可表示为

$$\begin{cases} f_b(\theta_b, G_{g_i}) = y_i \\ f_b(\theta_b, G) = f_o(\theta_o, G) \end{cases}, \quad (1)$$

其中, $f_b(\cdot)$ 为后门 GNN, $f_o(\cdot)$ 是原始 GNN, θ_o 与 θ_b 分别为原始模型和后门模型的参数集合, G_{g_i} 为注入了触发器的中毒图, G 为不受影响的良性数据。综上所述, 图后门攻击的目标是不仅使得后门分类器 $f_b(\cdot)$ 将中毒数据 G_{g_i} 分类为攻击者指定的类别 y_i , 而且不影响其他良性数据的分类正确率。

3.2 随机平滑防御

随机平滑是一种通过引入随机噪声至输入图, 以增强图神经网络(Graph Neural Network, GNN)对细微扰动的抵抗能力的技术。该方法促使网络从扰动数据中提取出稳定的特征, 从而提高其对抗性。然而, 研究表明, 当扰动达到一定的幅度时, 随机平滑的防御机制可能会失效^[16]。此外, 后门攻击的目的通常是在不被检测的情况下修改模型的行为。如果污染的数据点过多, 尤其是这些点在图结构中显得突出或不自然, 那么这种攻击更有可能被安全机制或在数据清洗过程中使用的异常检测算法所识别^[23]。

假设模型的输入数据为 a , 其维度为 s , 给定一个已训练好的 GNN 图分类器 f (在防御过程中 f 可以是后门分类器), 将输入 a 映射到有 c 个元素的标签集 $\{1, 2, \dots, c\}$ 中。随机下采样 $S(a, z)$ 被定义为保留 a 中的 z 个下采样特征, 并将剩余的特征设置为一个特殊值(例如 0)。随后, 随机平滑技术将随机下采样结果输入 f 进行预测, 属于第 j 个类别的概率定义为

$$p_j = \Pr(f(\theta_r, S(a, z)) = j), \forall j \in \{1, 2, \dots, c\}, \quad (2)$$

其中, θ_r 为模型参数。基于此, 平滑 GNN 分类器可形式化为式(3), 遵循投票原则取分类结果中标签数量最多的一个类别作为图分类的最终标签。

$$\bar{f}(x) = \arg \max_{j \in \{1, 2, \dots, c\}} \Pr(f(\theta_r, S(a, z)) = j) = \arg \max_{j \in \{1, 2, \dots, c\}} p_j \quad (3)$$

值得注意的是, 为了避免训练数据与测试数据分布不一致, 在训练随机平滑 GNN 分类器时也同样使用下采样后的数据。基于随机平滑机制的图后门攻击防御对测试图 G 进行 d 次随机下采样, 使得每次下采样后得到的子图保留了不同的图信息。随后, 防御机制使用已训练好的 GNN 图分类器 f 来预测这 d 个子图的标签, 根据式(3)以投票原则选择数量最多的标签 l 作为平滑 GNN 分类器 $\bar{f}$ 对测试图 G 的分类结果, 其中, $l = \arg \max_{j \in \{1, 2, \dots, c\}} p_j$ 。由于后门攻击通

常只对少量的节点进行投毒以追求隐蔽性, 因此, 多次随机下采样过程能够以较大的概率将毒化节点的恶意影响过滤掉。

3.3 认证鲁棒性

随机平滑可以增强模型对输入扰动的稳健性, 另外, 通过量化分析随机噪声对模型预测结果的影响, 可以评估模型在实际操作中面对不确定性的表现。这种评估为模型的鲁棒性提供了数学上的保证。

假设攻击者向输入数据 $\mathbf{a}$ 中添加了一个有限的扰动 δ , 对于平滑 GNN 分类器, 认证鲁棒性定义了一个扰动边界, 当 δ 在该边界内时, 平滑 GNN 分类器能够给出 $\bar{f}(\mathbf{a} + \delta) = \bar{f}(\mathbf{a})$ 的高概率保证。子图后门攻击对图数据只执行少量的边删除或添加操作, 使得扰动是稀疏的, 因此, 可用 ℓ_0 范式约束扰动 δ 。平滑 GNN 分类器的认证鲁棒性可形式化为

$$\bar{f}(\mathbf{a} + \delta) = \bar{f}(\mathbf{a}), \forall \|\delta\|_0 \leq \varphi(\mathbf{a}), \quad (4)$$

其中, $\varphi(\mathbf{a})$ 是扰动边数, 表征了允许的扰动空间范围。受文献[17]的启发, $\varphi(\mathbf{a})$ 的最大取值可由式(5)计算:

$$\binom{s - \varphi(\mathbf{a})}{z} / \binom{s}{z} > \frac{K+1}{K} - \underline{p}_l, \quad (5)$$

其中, 不等式左边是在未受扰动的 $s - \varphi(\mathbf{a})$ 个项中下采样 z 个项的组合数与在总共 s 个项中下采样 z 个项的组合数之比, 可用来估计当扰动限制在 $\varphi(\mathbf{a})$ 范围内时, 模型预测结果仍然正确的可信度。 K 表示平滑 GNN 分类器处理的 K 分类任务, $\underline{p}_l$ 表示模型预测标签为 l 的概率下界。由投票原则可知, 在 K 分类任务中满足 $\frac{1}{K} < \underline{p}_l \leq 1$ 。设置概率阈值为 $\frac{K+1}{K} - \underline{p}_l \in [0, 1)$ 可在统计上提供一个更加保守的鲁棒性保证, 即意味着即使存在扰动, 模型的预测标签 l 仍然具有较高的置信度, 例如, $K=2$ 时, 概率阈值为 $1.5 - \underline{p}_l$ 。假设在 d 次独立下采样数据中, 被平滑 GNN 分类器预测为最终预测标签 l 的数据样本有 d_l 个, 此外, 将标签 l 视作一类, 其余预测标签类别视作另一个类, 则 d_l 服从参数为 d 与 $\underline{p}_l$ 的二项分布, 可以利用 Clopper-Pearson 方法估计 l 与 $\underline{p}_l$ [16]。

3.4 问题描述

本文研究工作假设基于随机平滑的防御机制已开启, 并且在攻击条件设置方面与 SBA 方法保持一致, 即攻击者能够获取和修改 GNN 模型相关信息, 并能够操纵小部分数据集进行攻击。值得强调的是, 本文工作的主要目标是对抗随机平滑防御机制,

因此, 在后门攻击能力和隐蔽性两者之间权衡, 本文更倾向于提升攻击能力。在给定的原始数据集 $\mathcal{D}$ 中按 3:2 比例随机划分得到干净训练数据集 $\mathcal{D}_{\text{ctr}}$ (clean train dataset) 和干净测试数据集 $\mathcal{D}_{\text{cte}}$ (clean test dataset)。以干净训练数据集为例, 可表示为 $\mathcal{D}_{\text{ctr}} = \{(G_1, y_1), (G_2, y_2), \dots, (G_M, y_M)\}$, 其中的干净图可表示为 $\mathcal{G}_{\text{tr}} = \{G_1, G_2, \dots, G_M\}$ 。

令随机平滑技术中的随机下采样的比例为 β , 在图 G 上进行一次下采样结果表示为 $S_\beta(G)$, 则进行 d 次下采样后得到的子图集合是 $\{S_\beta^1(G), \dots, S_\beta^d(G)\}$ 。此外, 攻击者按照 η 的比例向数据集 $\mathcal{D}_{\text{ctr}}$ 投毒后, 得到后门训练数据集 $\mathcal{D}_{\text{btr}}$ (backdoor train dataset) 和中毒的图数据集 $\mathcal{G}_{\text{tr}}^t = \{G_1^t, \dots, G_m^t\}$, 其中, $m = \text{Ceil}(M \times \eta)$, $\text{Ceil}(\cdot)$ 为向上取整操作, 并且 $\mathcal{G}_{\text{tr}}^t$ 内所有图的标签都更改为攻击者选定的目标标签 y_t 。进一步地, 将后门训练数据集 $\mathcal{D}_{\text{btr}}$ 中的干净图数据集表示为 $\mathcal{G}_{\text{tr}}^c$, 且满足 $\mathcal{G}_{\text{tr}}^t \cap \mathcal{G}_{\text{tr}}^c = \emptyset, \mathcal{G}_{\text{tr}}^t \cup \mathcal{G}_{\text{tr}}^c = \mathcal{G}_{\text{tr}}$ 。由于 $\mathcal{D}_{\text{btr}}$ 中的中毒图与干净图在模型训练前都需进行 d 次随机下采样, 因此, $\mathcal{D}_{\text{btr}}^j = \{(S_\beta^j(G_1), y_1), \dots, (S_\beta^j(G_M), y_M)\}$, $\mathcal{D}_{\text{btr}} = \{\mathcal{D}_{\text{btr}}^1, \dots, \mathcal{D}_{\text{btr}}^d\}$ 。后门测试数据集 $\mathcal{D}_{\text{bte}}$ (backdoor test dataset) 由 $\mathcal{D}_{\text{cte}}$ 中标签不是 y_t 的图数据组成。

假设原始 GNN 分类器为 f , 添加随机平滑技术后得到平滑 GNN 分类器 $\bar{f}$, 攻击者期望基于中毒训练数据集, 将 $\bar{f}$ 训练成为后门模型。在注入后门阶段, 后门触发器的结构由触发器生成模型的参数 θ_p 控制, θ_p 为 ER 随机图生成模型生成边的概率, 取值范围为 $[0, 1]$ 。另外, 后门触发器的注入位置由中毒节点选择模块生成, 选择的注入节点位置用集合表示为 $V_p = \text{SelectPosionedNode}(G, t)$, 其中, t 为输入的中毒节点数量参数, 中毒节点选择模块 $\text{SelectPosionedNode}(\cdot)$ 使用 GAT 模型学习图 G 中的节点嵌入表示, 并由高斯混合模型(GMM)进行节点聚类, 最后结合综合性节点评分规则选择 t 个最重要的节点作为触发器注入位置。

添加了随机平滑技术的 GNN 分类器针对干净图上的预测结果表示为 $\bar{f}_\theta(G)$, 其中, θ 为模型参数。注入子图后门触发器操作可形式化为 $I(G, V_p, \theta_p)$, 其中 $I(\cdot)$ 表示触发器注入操作, V_p 为触发器注入位置, θ_p 为 ER 模型生成边的概率, 则添加有随机平滑技术的 GNN 对中毒数据的预测结果可表示为

$\bar{f}_\theta(I(G, V_p, \theta_p))$, 其中 $j=1, 2, \dots, d$ 。进而, 平滑 GNN 分类器对下采样数据的输出可表示为 $\bar{f}_\theta(S^j(I(G, V_p, \theta_p)))$ 。综上所述, 攻击目标可形式化为

$$\begin{aligned} \min_{\theta_p} \sum_{G_i \in \mathcal{G}_u^l} \sum_{j=1, \dots, d} \mathcal{L}(\bar{f}_{\theta_i}(S_\beta^j(I(G_i, V_p, \theta_p))), y_i) \\ \text{s.t. } \theta_i^* = \arg \min_{\theta_i} \sum_{G_i \in \mathcal{G}_u^c} \sum_{j=1, \dots, d} \mathcal{L}(\bar{f}_{\theta_i}(S_\beta^j(G_i)), y_i) \quad (6) \\ + \sum_{G_i \in \mathcal{G}_u^l} \sum_{j=1, \dots, d} \mathcal{L}(\bar{f}_{\theta_i}(S_\beta^j(I(G_i, V_p, \theta_p))), y_i), \end{aligned}$$

其中, $\mathcal{L}$ 为交叉熵损失函数, θ_p 为触发器生成模型的参数, θ_i 为目标 GNN 模型的参数。式(6)旨在给定注入位置 V_p 时, 通过优化参数 θ_p 生成攻击性能最佳的触发器结构。

4 本文提出的 SBA-ST 方法

由于生成合适的子图后门触发器并选择最优注入位置具有挑战, 本文将后门攻击过程拆分成中毒

节点选择和子图触发器生成两个部分进行。因此, 本文提出的攻击方法需要解决的问题包括: 1) 如何选择对子图后门攻击最有用的 t 个(子图触发器大小)节点, 以便作为后门触发器的嵌入位置; 2) 如何选择合适的子图后门触发器, 以保证较高的攻击成功率和一定的不可察觉性。为此, 本文提出了如图 1 所示的子图后门攻击方法 SBA-ST, 该方法首先选取 GAT 方法在投毒目标图上训练一个节点编码模型, 以生成参与图分类任务的节点嵌入表示, GAT 中的通用自注意力框架能够使得节点更好地从邻居聚合特征向量, 从而捕获节点特征和图结构信息^[21]。随后, SBA-ST 方法进一步使用 GMM 进行聚类分析, 得到每个节点类别的聚类中心, 进而综合考虑节点自身重要性以及与聚类中心距离来选择后门注入位置, 如图 1 中红色节点所示。最后, 通过 ER 随机图生成模型产生子图后门触发器, 以实现降低计算复杂度的前提下保证较高的后门攻击效果。注入触发器时, 将被选中的投毒节点之间的边关系替换成触发器中的边关系。

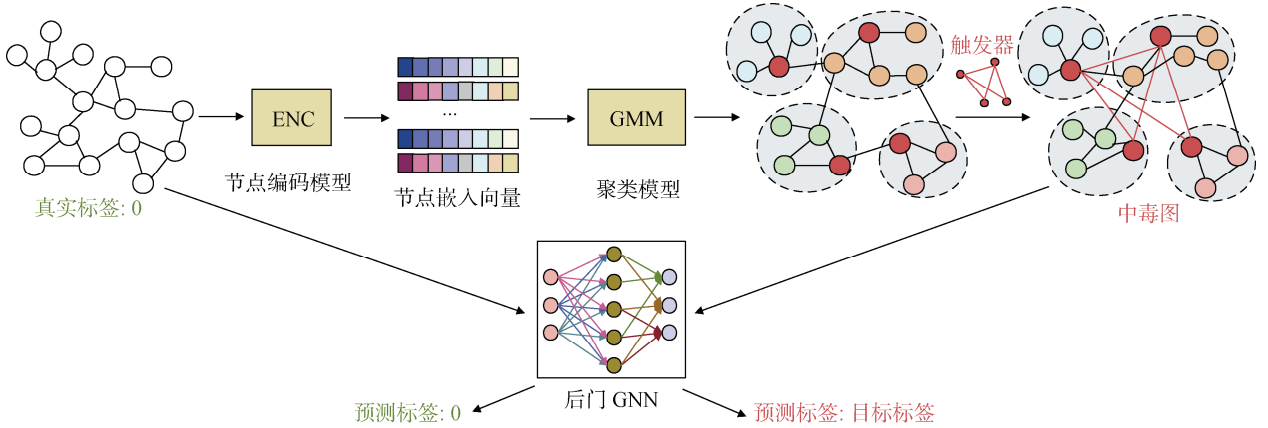


图 1 SBA-ST 方法的总体框架

Figure 1 Overall framework of the proposed SBA-ST method

4.1 中毒节点选择

直观来看, 攻击具有随机平滑防御机制的 GNN 分类器需要注入的后门能够在随机子采样后保留下来, 进而通过消息聚合机制将恶意的信息传递出去, 最终使得该图被分类为指定的错误类别。前文可知, 平滑 GNN 分类器 $\bar{f}$ 对图 G 的预测标签是在 d 个下采样子图的预测结果中通过投票原则确定的。如图 2 所示, 原始图数据的标签为 label:0, 假定标红节点为分类主要贡献节点, 按相同保留比例进行 d 次对随机下采样后得到的 d 个子图。值得注意的是, 在下采样子图过程中, 保留一定数量的重要贡献节点使得子图的标签同样能够被分类为 label:0, 而没有保留主要贡献节点的子图则会被误分类为 label:1。因

此, 选择对图分类任务具有重要性的节点作为污染目标, 在保持扰动范围限制的同时嵌入后门, 不仅可以保持攻击的隐蔽性, 还能有效地保留后门中的恶意信息。通过干扰这些关键节点, 可以使得平滑分类器对恶意信息更加敏感, 即使只有少数关键节点受到扰动, 其所携带的恶意信息也足以诱使平滑分类器产生误分类。

为了确定待投毒的关键节点, SBA-ST 考虑首先使用神经网络学习图的节点表示, 再利用 GAT 方法筛选出关键节点。相较于文献[36]中使用的 GCN 编码器, GAT 通过引入自注意力网络, 捕获节点特征和图结构信息, 从而学习到更有效的用于图分类任务的图嵌入表示。考虑到图表示学习到的节点特征具

有较高的维度, 对每个类使用 GMM 选择具有代表性的节点, 以此获得准确的聚类中心, 随后通过使用 Expectation maximization(EM)算法进行求解优化。训练节点编码器的过程可用公式表示为

$$\mathbf{H} = f_{\text{GAT}}(\mathbf{A}, \mathbf{X}), \quad \hat{\mathbf{Y}} = \text{softmax}(\mathbf{W} \cdot \mathbf{H}), \quad (7)$$

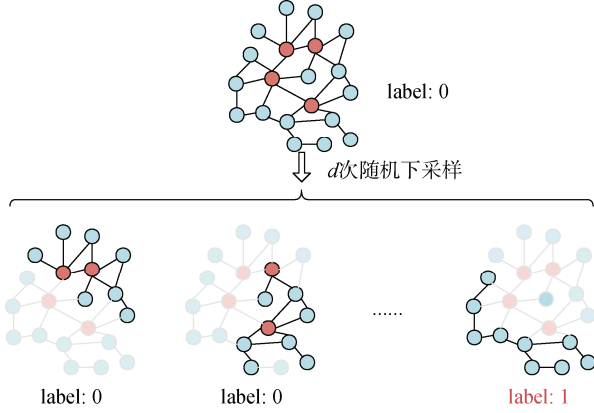


图 2 随机下采样示意图

Figure 2 An illustration of Randomized Subsampling

其中, $\mathbf{W}$ 表示用于图分类的可学习的权重矩阵。GAT 编码器的训练过程可表示为

$$l(\theta_e, \mathbf{W}) = \min_{\theta_e, \mathbf{W}} \sum_{i=1}^N \mathcal{L}_e(\hat{y}_i, y_i), \quad (8)$$

其中, θ_e 是 GAT 编码器的参数, $\mathcal{L}_e$ 表示编码器使用的交叉熵损失函数, y_i 表示节点 v_i 的真实标签, $\hat{y}_i$ 是节点 v_i 的预测标签。利用式(8)可训练得到 GAT 编码器, 获得节点嵌入表示, 随后在攻击者指定的目标标签 y_t 之外的每个类别上使用 GMM 计算出第 k 个类别的聚类中心 $\mathbf{h}_c^k$ 。直观认为, 节点离聚类中心越近, 它的重要程度越高。当节点自身的重要性较大时, 后门触发器的注入会使得恶意信息沿着多条边进行迅速传播, 从而增强攻击的不可察觉性, 但同时也会显著降低模型的预测性能。因此, 在筛选中毒节点时还考虑了节点的特征向量中心性, 以此权衡攻击性能和预测性能之间的关系。实际上, 节点 v_i 的特征向量中心性 $\text{EC}(v_i)$ 既取决于自身的重要性, 还取决于邻居的重要程度, 即 $\text{EC}(v_i) = c \cdot \sum_{j=1}^n \mathbf{A}_{ij} \cdot \text{EC}(v_j)$, $\mathbf{A}_{ij}$ 为邻接矩阵对应位置的值, c 为一个比例常数, 其参考取值可参考文献[42]。对属于第 k 簇的节点 v_i^k , 其得分计算为

$$s(v_i) = \|\mathbf{h}_i^k - \mathbf{h}_c^k\|_2 + \delta \cdot \text{EC}(v_i^k), \quad (9)$$

其中, δ 为控制特征向量中心性占比的参数。所有节

点的分数计算完成后, 选择 t 个分数最高的节点构成集合 V_P , $|V_P|$ 为后门触发器的大小。

综上所述, 算法 1 给出了中毒节点选择算法的伪代码。

算法 1 中毒节点选择算法

输入: 邻接矩阵 $\mathbf{A}$ 、属性矩阵 $\mathbf{X}$

中毒节点数量 t

输出: 中毒节点集合 V_P

- 1: 初始化 GAT 模型参数 θ_e
- 2: 根据式(9)最小化损失函数
- 3: $\mathbf{H} \leftarrow \text{GCN}(\theta_e, \mathbf{A}, \mathbf{X})$
- 4: 用 GMM 模型计算簇 k 的聚类中心 $\mathbf{h}_c^k$
- 5: 根据式(10)计算分数 s
- 6: $V_P \leftarrow \text{top}(s, t)$

现在站在认证鲁棒性视角进行理论分析, 当添加到图结构向量的扰动的 ℓ_0 范数满足公式(4)中 $\varphi(\cdot)$ 空间的约束时, 意味着平滑 GNN 分类器可对带有扰动的图 G 进行正确预测。图 G 中包含 $s = |G|(|G|-1)/2$ 个边项, 每个项的取值为 0 或 1。攻击者注入后门触发器可看作对图的二进制结构向量中的部分项进行翻转, 因此 ℓ_0 范式约束的是翻转边的数量。攻击者注入一个大小为 $t = |V_P|$ 的触发器, 最多翻转 $t(t-1)/2$ 条边。因此, 当添加到图中的触发器大小不大于阈值 T 时, 平滑 GNN 分类器会为扰动图预测正确的真实标签。公式(5)可转换为

$$\frac{\binom{|G|(|G|-1)/2 - T(T-1)/2}{z}}{\binom{|G|(|G|-1)/2}{z}} > \frac{K+1}{K} - \underline{p}_l, \quad (10)$$

其中, $\underline{p}_l$ 通过式(6)以 99% 的置信水平进行估计, 阈值 T 为认证触发器大小。由于 SBA-ST 方法选择重要性较大的节点进行扰动, 在限定了扰动产生位置的同时增加了随机下采样后恶意信息的保留概率。标签 l 是由投票原则选出的, 这种方法本质上偏好于那些在 d_l 次下采样样本中频繁出现的关键节点, 因为它们更可能影响投票结果。然而, 对关键节点的扰动会减少在 $s - \varphi(a)$ 采样集中包含这些节点的概率, 即降低了下采样数据 z 被正确预测的概率, 同时减少了 $\binom{s - \varphi(a)}{z}$ 的组合数, 这使得 SBA-ST 方法相较于 SBA 方法能够以更小的阈值 T 满足公式(10)。

4.2 子图后门触发器生成方法

图生成方法目前分为传统的图生成方法与深度

生成模型两大类, 其中, 传统的图生成方法包括了 ER 模型、随机块模型、优先链接模型等, 通过指定一个生成过程来定义如何构造图中的边; 深度生成模型则通过观察一组图 $\{G_1, \dots, G_l\}$ 学习到如何生成具有相似特性的图, 较为广泛应用的有变分自编码器、生成对抗网络和自回归模型。由于不同图对应的有效攻击子图并不相同, 使用深度生成模型并不是最佳选择。另外, 针对子图后门触发器生成问题, 文献[4]提出了通过优化攻击目标来自适应地生成最优的子图结构, 但这类方法通常基于梯度进行计算, 计算复杂度高。为了平衡好攻击性能与计算代价, SBA-ST 方法选择了 ER 模型作为子图后门触发器的生成方法, 该模型在传统图生成方法中是一种较为高效的子图触发器生成模型^[17]。ER 模型将任意一对节点之间出现边的概率定义为

$$P(A[u, v] = 1) = \theta_p, \forall u, v \in \mathcal{V}, u \neq v, \quad (11)$$

其中, $\theta_p \in [0, 1]$ 是控制图密度的参数, $\mathcal{V}$ 是生成图的节点集合。为此, SBA-ST 首先使用 ER 模型确定 $\mathcal{V}$ 的节点数量 $|V_p|$, 即生成的子图触发器的节点数量等于中毒节点数量, 通过交替优化与贪婪搜索的方式优化式(7)选择最优的概率。具体来说,

1) 第一步, 初始化 θ_p 、 θ_t 取值, 在后门训练数据集上得到优化后的后门模型参数 θ_t^* , 可表示为

$$\min_{\theta_t} \mathcal{L}(\theta_t, \theta_p) = \sum_{G_i \in \mathcal{G}_{tr}^c} \sum_{j=1, \dots, d} l(\bar{f}_{\theta_t}(S_{\beta}^j(G_i)), y_i) + \sum_{G_i \in \mathcal{G}_{tr}^t} \sum_{j=1, \dots, d} l(\bar{f}_{\theta_t}(S_{\beta}^j(I(G_i, V_p, \theta_p))), y_i), \quad (12)$$

2) 第二步, 保持模型参数, 在后门测试集上测试并记录攻击成功率;

3) 第三步, 按照预定义的搜索步长 ω 更新 θ_p , 重复第二步, 其中, ω 为常数, 作用是将连续的搜索空间形式化为可遍历的离散空间, 并通过贪婪搜索的思想选择攻击成功率最高情况下的 θ_p 和 θ_t^* 作为近似最优结果。

综上所述, 算法 2 给出了后门模型优化算法的伪代码。

算法 2 后门模型优化算法

输入: 中毒节点集合 V_p 、随机下采样比例 β

邻接矩阵 A 、属性矩阵 X 、搜索步长 ω

输出: 后门模型参数 θ_t^*

1: 初始化参数 θ_p 、 θ_t 、 $i = 0$

2: **WHILE** $\theta_p \leq 1$ **DO**

3: 根据式(13)优化得到 $\theta_t^{(i)}$

4: **IF** $\bar{f}_{\theta_t^{(i)}}(\mathcal{D}_{btr})$ better than $\bar{f}_{\theta_t^*}(\mathcal{D}_{btr})$ **DO**

5: $\theta_t^* = \theta_t^{(i)}$

6: $\theta_p \leftarrow \theta_p + \omega$

7: $i \leftarrow i + 1$

8: **return** θ_t^*

5 性能评价

5.1 实验准备

(1) 实验数据集

实验选用了常见的几种图分类应用数据集, 化合物数据集 MUTAG、NCI-109、Mutagencity, 生物蛋白质数据集 DD, 社交网络数据集 IMDBBINARY、COLLAB, 详细实验数据信息见表 1。

1) MUTAG 是硝基芳香族化合物集合, 旨在判断化合物是芳香族还是杂芳族, 包括 188 个化合物样品;

2) NCI-109 数据集包含 4127 个化合物, 目标是判断化合物是否有阻碍癌细胞增长的性质;

3) DD 是判断蛋白质为酶或非酶的数据集;

4) IMDBBINARY 是一个电影协作数据集, 每个演员代表一个节点, 边代表两个演员是否出现在同一个影片里, 图标签分为爱情片和动作片两个类别;

5) COLLAB 是一个科学合作数据集, 每个节点代表一个研究者, 图标签分为高能物理、凝聚态物理和天体物理三个类别。

表 1 数据集信息

Table 1 Information of datasets

数据集	图数量	平均节点数	图类别
MUTAG	188	17.7	2
NCI-109	4127	29.7	2
DD	1178	284.3	2
IMDBBINARY	1000	19.8	2
COLLAB	5000	74.5	3

(2) 性能评价指标定义

令干净的 GNN 图分类器为 f_c , 后门分类器为 f_b , 干净测试集 $\mathcal{D}_{cte} = \{(G_1, y_1), \dots, (G_S, y_S)\}$, 后门测试集 $\mathcal{D}_{bte} = \{(G_1, y_1), \dots, (G_M, y_M)\}$, 且 $M < S$, 则本文使用的性能评价指标详细定义如下。

1) 干净分类准确率(CleanACC): 指 GNN 图分类器 f_c 在干净测试集 $\mathcal{D}_{cte}$ 上预测标签的准确率, 即

$\text{CleanACC} = \frac{\sum_{i=1}^S \mathbb{I}(f_c(G_i) = y_i)}{S}$, 其中, $\mathbb{I}$ 为指示函数。

2) 后门分类准确率(BackdoorACC): 指后门分类器 f_b 在干净测试集 $\mathcal{D}_{\text{cte}}$ 上预测标签的准确率, 即

$$\text{BackdoorACC} = \frac{\sum_{i=1}^S \mathbb{I}(f_b(G_i) = y_i)}{S}。$$

3) 平均攻击成功率(Average Success Rate, ASR): 指后门分类器在后门测试集 $\mathcal{D}_{\text{bte}}$ 上将测试图预测为攻击者选定的目标标签的比例, 即 $\text{ASR} = \frac{\sum_{i=1}^M \mathbb{I}(f_b(G_i) = y_i)}{M}$, 其中, y_i 是指定的目标标签。

4) 准确率下降程度(DropACC): 指分类器受到后门攻击后, 在测试集上预测准确率下降的程度, 即 $\text{DropACC} = \text{CleanACC} - \text{BackdoorACC}$ 。

5) 随机平滑前后后门分类准确率下降程度(DropBackdoorACC): 指分类器在遭受后门攻击的情况下, 关闭随机平滑防御条件下与开启防御条件下的后门分类准确率差值, 即

$$\text{DropBackdoorACC} = \text{BackdoorACC}_{\text{NoDef}} - \text{BackdoorACC}_{\text{Def}}$$

其中, $\text{BackdoorACC}_{\text{NoDef}}$ 是关闭随机平滑防御条件下的 BackdoorACC 值, $\text{BackdoorACC}_{\text{Def}}$ 是开启随机平滑防御条件下的 BackdoorACC 值。

6) 随机平滑前后平均攻击成功率下降程度(DropASR): 指分类器在遭受后门攻击的情况下, 关闭随机平滑防御条件下与开启防御条件下的平均攻击成功率差值, 即 $\text{DropASR} = \text{ASR}_{\text{NoDef}} - \text{ASR}_{\text{Def}}$, 其中, $\text{ASR}_{\text{NoDef}}$ 是关闭随机平滑防御条件下的 ASR 值, ASR_{Def} 是开启随机平滑防御条件下的 ASR 值。

训练 GNN 分类器使用干净训练数据集 $\mathcal{D}_{\text{ctr}}$, 它从原始数据集中随机抽取 2/3 的图组成, 剩下 1/3 构成干净测试集 $\mathcal{D}_{\text{cte}}$ 。后门训练数据集从干净训练集中随机抽取 η 比例的图, 并注入自适应生成的子图后门触发器, 将标签改为攻击者指定的目标标签 y_i , 在代码中默认 y_i 为 label 0。后门测试集则是从干净测试集中选择真实标签不是 y_i 的图, 然后注入自适应触发器, 以此测试后门攻击的成功率。

本实验使用的对比方法是来自文献[17]的随机子图后门攻击方法。

5.2 对比实验设置

首先, 本文按照经验将随机平滑中的随机下采样比例 β 和次数 d 分别选择为 0.1 和 100^[17], 训练轮次设置为 100, 此外, θ_p 初始化为 0.1, 搜索步长 ω 设置为 0.05, 通过在公开实验数据集上进行对照实

验, 以探究不同攻击方法对该条件下的随机平滑防御的对抗能力。通过在干净数据集上进行图分类任务的训练得到基准分类准确率, 以便于比较后门攻击对干净数据集的分类准确率的影响。中毒节点数量 t 根据图数据集规模不同选择不同的值, 具体来讲, 在 NCI-109、MUTAG 和 IMDBBINARY 三个数据集实验中初始设置为 4, 在 DD 和 COLLAB 数据集实验中分别初始设置为 7 和 10。最后, 将训练数据集与测试数据集划分比例设置为 3:2, 投毒比例 η 为 $[0, 0.2]$, 并随机初始化 θ_e 和 θ_i 模型参数。

所有实验均在配备 RTX A4000-16G GPU、2.30GHz Xeon 5218 CPU 和 128GB 内存的服务器上进行, 所有代码基于 PyTorch 深度学习框架编写。对比实验详细环境配置情况如表 2 所示。

表 2 对比实验详细环境配置一览表

Table 2 Detailed environment configuration list for comparison experiments

环境配置项	详细配置情况
操作系统	Ubuntu 18.04 LTS
GPU 配置	NVIDIA RTX A4000 16GB
Pycharm 版本	Pycharm 2023.2.4(Professional Edition)
Pytorch 版本	torch 1.12.1+cu116
python 环境	python 3.8

5.3 不同方法的综合性能对比

为了便于后续方法性能对比, 本文基于干净数据集评价未被攻击的图学习模型(即干净 GNN 模型)的分类性能。表 3 展示了干净 GNN 模型在 5 个数据集上的 CleanACC 指标结果, 该结果将作为后续实验结果对比的基准性能结果。

表 3 基于不同数据集的干净 GNN 模型分类准确率结果

Table 3 Results of clean accuracy of the clean GNN model over different datasets

数据集	NCI-109	MUTAG	IMDBBINARY	DD	COLLAB
CleanACC	0.697	0.762	0.733	0.732	0.761

进一步地, 为了评价本文提出的 SBA-ST 方法相比 SBA 方法^[17]的攻击性能增益, 本文选取文献[11]中的攻击对象 GIN^[43]图分类模型实施子图后门攻击, 并在后门测试集上评估后门攻击的成功率、在干净测试集上评估其分类准确率, 基于此计算出不同后门攻击方法对非目标数据集分类能力的影响。表 4 给出了 SBA 和 SBA-ST 两种方法在 5 个数据集上针

对 GIN 模型实施后门攻击的性能对比结果, 其中, 括号中第一个和第二个数值分别是 SBA 和 SBA-ST 方法的实际性能表现, 并且粗体数据标识了对比之下较优的结果。从表 4 可知, 在 NCI-109、IMDBBINARY、DD 和 COLLAB 数据集对比实验中, 本文方法的 DropACC 指标分别为 0.067、0.023、0.071 和 0.081, 均比 SBA 的低, 验证了本文方法注入的后门更具不可察觉性。另一方面, NCI-109、IMDBBINARY、DD 和 COLLAB 数据集的对比实验结果显示, SBA-ST 方法的 BackdoorACC 指标相比 SBA 方法分别提高了 2.7%、7.0%、1.8%、4.1%; 此外, 在所有数据集对比实验中, 本文方法的 ASR 指标表现都优于对比方法, 尤其是 MUTAG 数据集, 攻击成功率提高了 39.0%,

上述对比实验结果进一步表明, 本文方法具有比基准方法更为强大的攻击能力。

进一步地, 为了对比不同攻击方法对抗随机平滑防技术的能力, 本文在原始分类模型上启用了防御机制, 并在训练前对训练图数据和测试图数据进行了随机下采样, 得到了如表 5 所示的对比实验结果。从表 5 可知, 针对干净测试数据集的分类精度方面, 在所有数据集对比实验中, SBA-ST 方法相比 SBA, 其 DropACC 结果更小并且 BackdoorACC 结果更大; 此外, 本文方法的 ASR 指标性能也都全面优于对比方法, 性能提升幅度从 18.2%到 38.1%不等。综上所述, 本文方法在面临随机平滑防御时拥有相对更稳定、更优越的攻击能力表现。

表 4 随机平滑防御未开启条件下 SBA 和 SBA-ST 方法的性能对比

Table 4 Performance comparison between SBA and SBA-ST methods when disabling randomized smoothing defense

数据集	NCI-109	MUTAG	IMDBBINARY	DD	COLLAB
BackdoorACC	(0.613, 0.630)	(0.746 , 0.682)	(0.660, 0.710)	(0.657, 0.669)	(0.652, 0.680)
ASR	(0.666, 0.706)	(0.523, 0.857)	(0.626, 0.660)	(0.612, 0.755)	(0.725, 0.866)
DropACC	(0.084, 0.067)	(0.016 , 0.080)	(0.073, 0.023)	(0.075, 0.071)	(0.109, 0.081)

表 5 随机平滑防御开启条件下 SBA 和 SBA-ST 方法的性能对比

Table 5 Performance comparison between SBA and SBA-ST methods when enabling randomized smoothing defense

数据集	NCI-109	MUTAG	IMDBBINARY	DD	COLLAB
BackdoorACC	(0.605, 0.616)	(0.634, 0.680)	(0.620, 0.650)	(0.635, 0.661)	(0.622, 0.650)
ASR	(0.489, 0.671)	(0.428, 0.809)	(0.520, 0.640)	(0.489, 0.735)	(0.562, 0.779)
DropACC	(0.092, 0.081)	(0.128, 0.082)	(0.113, 0.083)	(0.097, 0.063)	(0.139, 0.111)

为了更直观地展示不同方法在随机平滑防御机制开启前后的性能变化情况, 本文从另一个视角给出 SBA 和 SBA-ST 在后门分类准确率下降程度和平均攻击成功率下降程度两个方面的对比结果, 如图 3 和图 4 所示。从图 3 可知, 在 BackdoorACC 指标上, SBA-ST 方法在 MUTAG 和 DD 两个数据集上的性能下降幅度要小于 SBA 方法, 特别是 MUTAG 数据集的性能下降幅度显著小于对比方法, 这一对比结果表明, 本文方法在中小规模的图数据上具有更为鲁棒的后门攻击性能。另外, 虽然本文方法在 NCI-109 和 IMDBBINARY 两个数据集上的性能下降幅度略大于对比方法, 但是结合前文针对表 5 实验数据的分析, 本文方法的综合性能表现依然占优。从图 4 可知, 在 DropASR 指标上, 本文方法的性能是全面优于比较方法的, 特别是在大多数数据集上, SBA-ST 的性能下降幅度显著低于 SBA, 从而验证了本文方法在当前的随机平滑防御条件下具有更强的攻

击性能。

本文以 MUTAG 数据集为例, 通过实验探究投毒比例 η 对本文方法攻击性能的影响, 其中, 随机平滑防御机制处于开启状态, η 取值范围设置为 $\{0, 0.05, 0.1, 0.15, 0.2\}$, 结果如图 5 所示。从图 5 可知, 随着投毒比例从 0 增加到 0.2, 本文方法的 BackdoorACC 指标结果由 0.825 下降到 0.698, 然而, 其 ASR 指标由 0.333 上升到 0.857。上述实验结果说明, BackdoorACC 和 ASR 是相互矛盾的两个指标。考虑到在后门攻击场景, 攻击者更加希望在足够隐蔽的情况下取得足够的攻击成功率, 因此, 对于 MUTAG 数据集来说, 最优的投毒比例将被选择为 0.15, 此时, SBA-ST 能够在 BackdoorACC 性能下跌不超过 0.1 的前提下取得超过 0.7 的攻击成功率, 取得了在攻击隐蔽性和攻击能力之间良好的折中。值得说明的是, 上述 η 参数选择过程也适用于其他数据集实验。

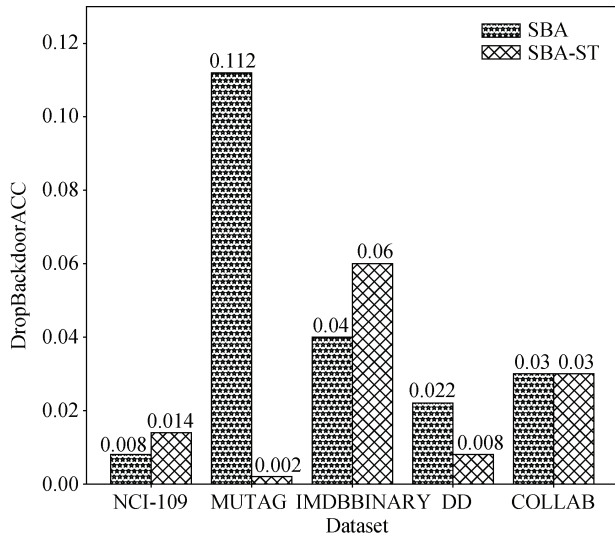


图3 SBA和SBA-ST的DropBackdoorACC性能对比

Figure 3 Performance comparison of DropBackdoorACC between SBA and SBA-ST

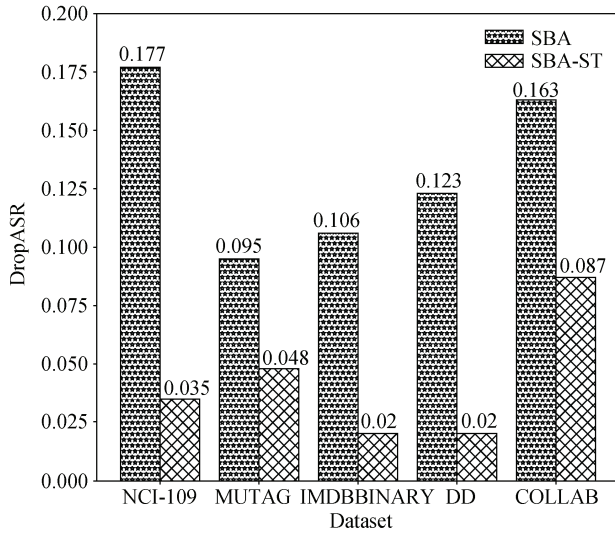


图4 SBA和SBA-ST的DropASR性能对比

Figure 4 Performance comparison of DropASR between SBA and SBA-ST

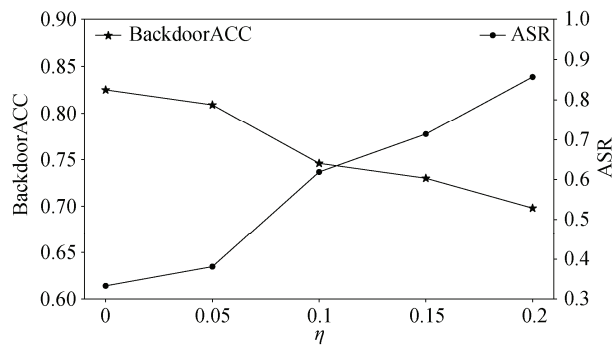


图5 在MUTAG数据集上投毒比例对本文方法攻击性能的影响

Figure 5 Impact of poisoning ratio on the proposed method's performance over the MUTAG dataset

5.4 不同随机平滑防御强度下的性能对比

本节的对比实验通过修改随机下采样比例 β 和触发器大小,以全面评价不同平滑防御强度下不同方法的性能优劣,其他参数与5.2节中设置相同。

首先,本文评价不同的随机下采样比例 β 对后门攻击效能的影响,将 β 分别设置为5%、10%、20%、30%,并进行了相应的对比实验,结果见图6,其中,虚线和实现分别表示SBA方法和SBA-ST方法的性能表现,并且每一个方法显示了BackdoorACC和ASR两个指标结果随 β 的变化情况。由图6可知,SBA方法的BackdoorACC指标基本不受 β 值变化的影响,而SBA-ST方法的BackdoorACC性能下降程度则随着 β 的增大而减小,原因是SBA方法通过随机选取中毒节点的方式实施均匀扰动,而SBA-ST方法倾向于针对那些对图分类结果具有更大影响的节点注入扰动信息。由于防御后的GNN分类器倾向于在预测最终标签时保留对分类结果影响较大的节点,

因此,随着下采样比例 β 的增加,SBA-ST方法投毒的节点被保留的概率增大,从而使得恶意信息对分类结果产生了较大影响。在NCI09和IMDBBINARY数据集实验中,见图6(a)和图6(c),当 β 取0.05和0.1时,SBA方法的BackdoorACC指标的下降幅度小于SBA-ST方法,原因是SBA-ST针对重要节点的投毒策略导致更多恶意信息被保留,进而降低了随机平滑后的BackdoorACC指标,导致下降区间扩大。另外,图6结果表明,两个方法的ASR性能下降程度随着 β 的增大而减小,这是因为较高的下采样比例将更有可能保留被毒害的节点,从而帮助攻击者成功攻击防御后的GNN分类器。此外,在大多数情况下,SBA-ST方法的ASR性能下降幅度小于SBA方法,但是当下采样比例接近0.3时,两个方法的ASR下降幅度相当且为一个极小值。通过上述结果分析,本文认为:1)当下采样比例较大时,基于随机平滑的防御策略无法有效过滤被扰动的节点,导致防御效能大幅下降。2)本文提出的SBA-ST方法倾向于扰动对分类结果影响大的重要节点,即使下采样比例 β 处于较低水平时,SBA-ST也能够保持比SBA方法更加稳定的攻击能力,从而验证了平滑分类器对于关键节点注入的恶意信息更加敏感。

随后,本文进一步评价不同的触发器大小(TriggerSize)对后门攻击效能的影响,将TriggerSize设置为原图节点数的1%、5%、10%、20%、30%,并进行了相应的对比实验。值得说明的是,本次对比实验主要评价不同的触发器大小对两个方法的ASR下降幅度的影响情况,结果见图7。

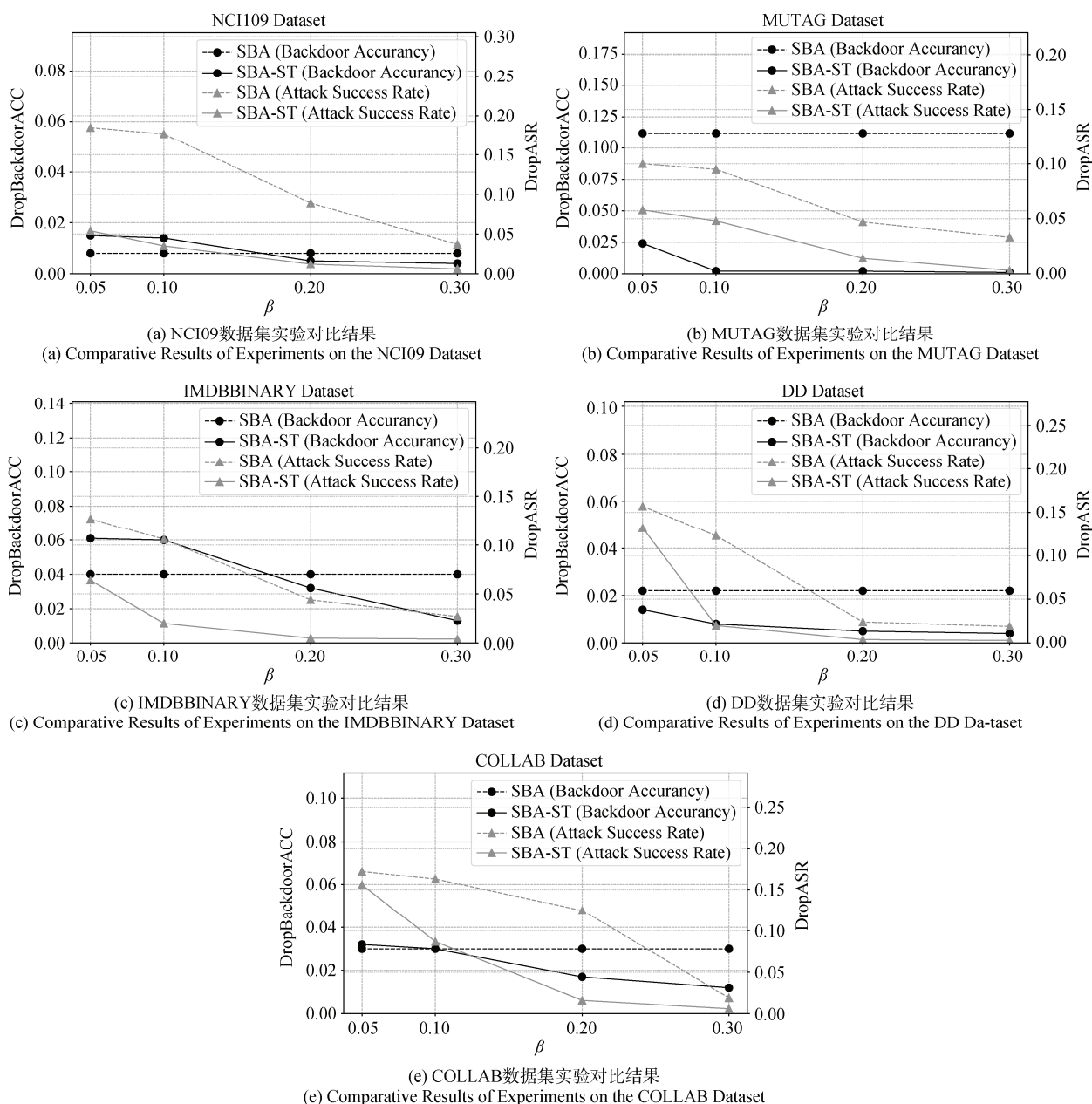


图 6 不同数据集实验中不同下采样比率对 DropBackdoorACC 和 DropASR 的影响

Figure 6 Impact of different subsampling ratios on DropBackdoorACC and DropASR with respect to different datasets

从对比实验结果可以看到, 在 5 个数据集上, 随着 TriggerSize 增大, 两个方法的 DropASR 下降幅度普遍呈减少趋势, 说明一个更大的后门触发器能够携带更多恶意信息, 从而有利于增强攻击者的攻击能力。从认证鲁棒性的角度分析, 当 TriggerSize 较小时, 经过防御的 GNN 分类器能够有效地处理引入的扰动, 导致 DropASR 显著下降, 显示随机平滑技术在过滤微小扰动方面的强大能力。然而, 随着触发器尺寸的增加, DropASR 的下降趋势放缓, 表明触发器尺寸的增加能够提高逃避该防御机制的能力。以图 7(a)所示的 NCI-109 数

据集实验结果为例, 当触发器大小为 1% 时, SBA 和 SBA-ST 方法均呈现出显著的 ASR 性能下降, 表明防御机制有效阻挡了攻击。然而, 当触发器大小增大至 5% 甚至更大的时候, SBA-ST 方法的 ASR 下降幅度缩小为 0.053, 甚至 0.011 (此时触发器大小为 30%)。相较之下, SBA 方法的 ASR 下降幅度在触发器大小设置为 30% 时仅缩小至 0.075。这一结果进一步验证了本文方法在面临随机平滑防御时具有相对稳定的攻击能力, 也验证了该方法在 NCI-109 数据集实验中使用比 SBA 方法更小的触发器能达到更好的攻击效果。

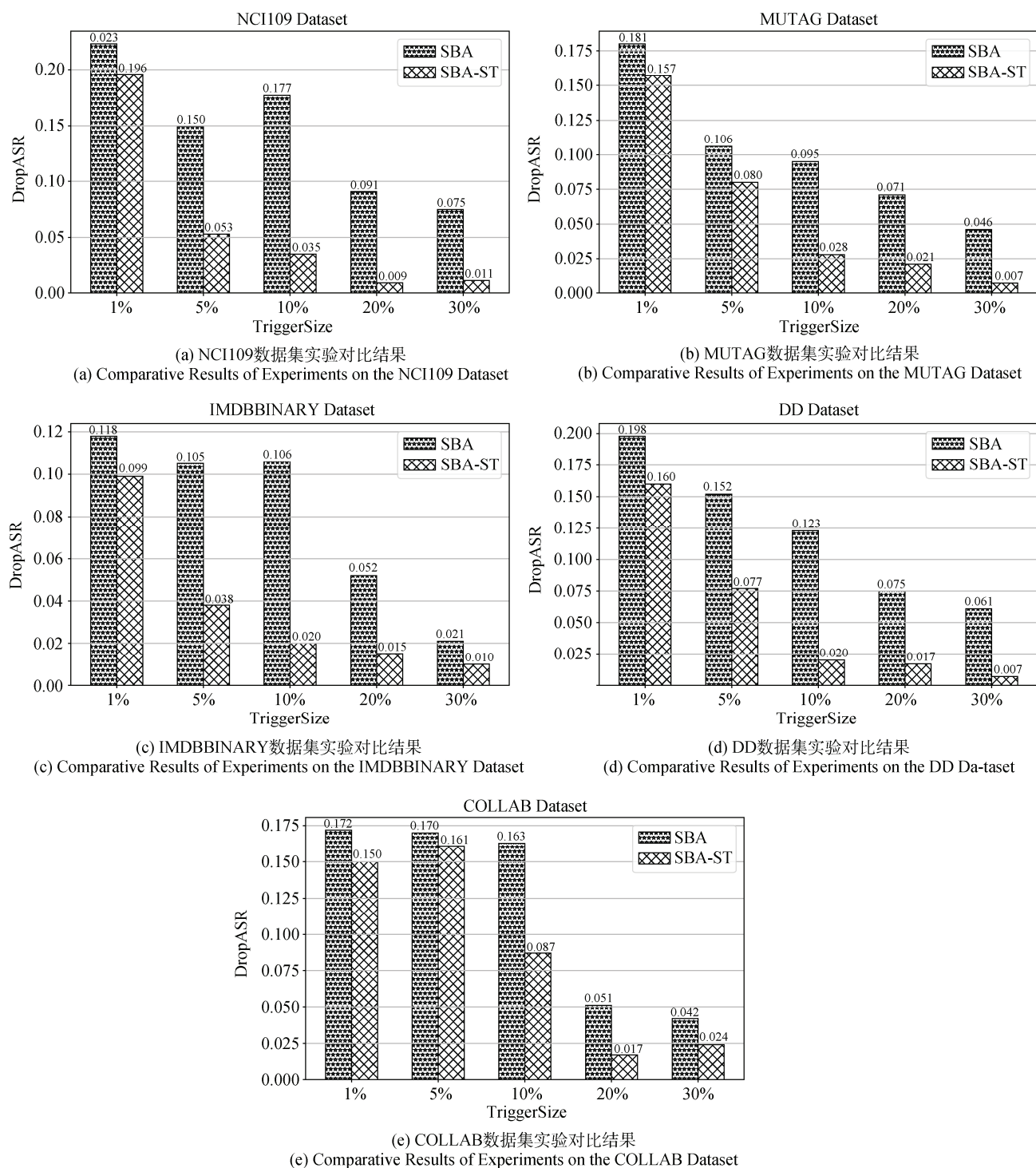


图 7 不同数据集实验中不同触发器大小对 DropASR 的影响

Figure 7 Impact of different trigger sizes on DropASR with respect to different datasets

从图 7 可以看到, 基于 MUTAG、IMDBBINARY、DD 和 COLLAB 四个数据集的对比实验结果呈现了与 NCI-109 数据集类似的趋势, 因此, 上述分析结论在这四个数据集实验中同样有效。考虑到已有工作证明了在实际攻击场景中, 明显的扰动容易被采用的多种检测方法识别^[23], 因此, 当 TriggerSize 选择为相对较小的取值时(比如 10%), SBA-ST 在 NCI-109、

MUTAG、IMDBBINARY 和 DD 数据集上的 DropASR 均低于 0.05, 而 SBA 方法的 DropASR 则超过或接近 0.1。以图 7(c)所示的 IMDBBINARY 数据集为例, 为了达到与 SBA-ST 在 TriggerSize 为 10%条件下相当的攻击性能, SBA 需要增大 TriggerSize 至 30%, 从而验证了 SBA-ST 可以使用比 SBA 更小的触发器来逃避随机平滑防御。

6 总结

针对图后门攻击这一研究热点, 本文扩展了 SBA 方法并提出了一种使用更小触发器逃避随机平滑防御的子图后门攻击方法 SBA-ST, 该方法以子图触发器的形式添加扰动, 通过干扰对图分类任务影响度较高的节点, 使其在随机平滑防御机制开启的情况下仍能够保持较强的攻击能力, 通过引入 GAT 模型和 GMM 聚类分析来捕捉节点的局部和全局特征, 从而学习出高质量节点嵌入表示, 并筛选出聚类中心附近重要性更大的节点, 进而设计了一个最佳后门注入位置选择机制。基于 5 个公开数据集上的对比实验结果表明, 本文提出的 SBA-ST 方法相比对比方法具有更强的后门攻击能力, 并能够使用比 SBA 更小的触发器来逃避随机平滑防御。

下一步, 本文的未来工作将聚焦在进一步提高图后门攻击的隐蔽性, 例如, 协同式图后门攻击机制。

参考文献

- [1] Yang S Q, Verma S, Cai B R, et al. Variational Co-Embedding Learning for Attributed Network Clustering[J]. *Knowledge-Based Systems*, 2023, 270: 110530.
- [2] Xu B B, Cen K T, Huang J J, et al. A Survey on Graph Convolutional Neural Network[J]. *Chinese Journal of Computers*, 2020, 43(5): 755-780.
(徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. *计算机学报*, 2020, 43(5): 755-780.)
- [3] Chen L, Li J T, Peng Q B, et al. Understanding Structural Vulnerability in Graph Convolutional Networks[C]. *The Thirtieth International Joint Conference on Artificial Intelligence*, 2021: 2249-2255.
- [4] Yang S Q, Doan B G, Montague P, et al. Transferable Graph Backdoor Attack[C]. *The 25th International Symposium on Research in Attacks, Intrusions and Defenses*, 2022: 321-332.
- [5] Tao S C, Cao Q, Shen H W, et al. Adversarial Camouflage for Node Injection Attack on Graphs[J]. *Information Sciences*, 2023, 649: 119611.
- [6] Ju M X, Fan Y J, Ye Y F, et al. Black-Box Node Injection Attack for Graph Neural Networks[EB/OL]. 2022: arXiv: 2202.09389. <https://arxiv.org/abs/2202.09389>.
- [7] Mu J M, Wang B H, Li Q, et al. A Hard Label Black-Box Adversarial Attack Against Graph Neural Networks[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 108-125.
- [8] Shao J L, Wang Y Q, Shi B S, et al. Adversarial for Social Privacy: A Poisoning Strategy to Degrade User Identity Linkage[EB/OL]. 2022: arXiv: 2209.00269. <https://arxiv.org/abs/2209.00269>.
- [9] Zang X, Xie Y, Chen J, et al. Graph Universal Adversarial Attacks: A few Bad Actors Ruin Graph Learning Models[C]. *The Thirtieth International Joint Conference on Artificial Intelligence*, 2021: 3328-3334.
- [10] Liu Z H, Luo Y, Wu L R, et al. Are Gradients on Graph Structure Reliable in Gray-Box Attacks? [C]. *The 31st ACM International Conference on Information & Knowledge Management*, 2022: 1360-1368.
- [11] Ma Y, Wang S H, Derr T, et al. Graph Adversarial Attack via Rewiring[C]. *The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021: 1161-1169.
- [12] Li Y M, Jiang Y, Li Z F, et al. Backdoor Learning: A Survey[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(1): 5-22.
- [13] Tan T J L, Shokri R. Bypassing Backdoor Detection Algorithms in Deep Learning[C]. *2020 IEEE European Symposium on Security and Privacy*, 2020: 175-183.
- [14] Chen D W, Fu A M, Zhou C Y, et al. Federated Learning Backdoor Attack Scheme Based on Generative Adversarial Network[J]. *Journal of Computer Research and Development*, 2021, 58(11): 2364-2373.
(陈大卫, 付安民, 周纯毅, 等. 基于生成式对抗网络的联邦学习后门攻击方案[J]. *计算机研究与发展*, 2021, 58(11): 2364-2373.)
- [15] Sun L C, Dou Y T, Yang C, et al. Adversarial Attack and Defense on Graph Data: A Survey[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(8): 7693-7711.
- [16] Zhang Z X, Jia J Y, Wang B H, et al. Backdoor Attacks to Graph Neural Networks[C]. *The 26th ACM Symposium on Access Control Models and Technologies*, 2021: 15-26.
- [17] Zheng H B, Xiong H Y, Ma H N, et al. Link-Backdoor: Backdoor Attack on Link Prediction via Node Injection[J]. *IEEE Transactions on Computational Social Systems*, 2024, 11(2): 1816-1831.
- [18] Xu J, Xue M H, Picek S. Explainability-Based Backdoor Attacks Against Graph Neural Networks[C]. *The 3rd ACM Workshop on Wireless Security and Machine Learning*, 2021: 31-36.
- [19] Ying R, Bourgeois D, You J X, et al. GNNExplainer: Generating Explanations for Graph Neural Networks[EB/OL]. 2019: arXiv: 1903.03894. <https://arxiv.org/abs/1903.03894>.
- [20] Xi Z, Pang R, Ji S, et al. Graph Backdoor[C]. *Proceedings of the 30th USENIX Security Symposium*, 2021: 1523-1540.
- [21] Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks[EB/OL]. 2017: arXiv: 1710.10903. <https://arxiv.org/abs/1710.10903>.
- [22] Tong X, Wang B J, Wang R Z, et al. Survey on Adversarial Sample of Deep Learning towards Natural Language Processing[J]. *Computer Science*, 2021, 48(1): 258-267.
(全鑫, 王斌君, 王润正, 等. 面向自然语言处理的深度学习对抗样本综述[J]. *计算机科学*, 2021, 48(1): 258-267.)
- [23] Ren K, Zheng T H, Qin Z, et al. Adversarial Attacks and Defenses in Deep Learning[J]. *Engineering*, 2020, 6(3): 346-360.
- [24] Tan Q Y, Zeng Y M, Han Y, et al. Survey on Backdoor Attacks Targeted on Neural Network[J]. *Chinese Journal of Network and Information Security*, 2021, 7(3): 46-58.
(谭清尹, 曾颖明, 韩叶, 等. 神经网络后门攻击研究[J]. *网络与信息安全学报*, 2021, 7(3): 46-58.)
- [25] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[EB/OL]. 2017: arXiv:

- 1706.06083. <https://arxiv.org/abs/1706.06083>.
- [26] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]. *2016 IEEE European Symposium on Security and Privacy*, 2016: 372-387.
- [27] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. <https://arxiv.org/abs/1412.6572>.
- [28] Liu X M, Xie L H, Wang Y P, et al. Adversarial Attacks and Defenses in Deep Learning[J]. *Chinese Journal of Network and Information Security*, 2020, 6(5): 36-53.
(刘西蒙, 谢乐辉, 王耀鹏, 等. 深度学习中的对抗攻击与防御[J]. *网络与信息安全学报*, 2020, 6(5): 36-53.)
- [29] Zhang H, Yuan X L, Zhou C, et al. Projective Ranking-Based GNN Evasion Attacks[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(8): 8402-8416.
- [30] Zhang S X, Chen H X, Sun X G, et al. Unsupervised Graph Poisoning Attack via Contrastive Loss Back-Propagation[C]. *The ACM Web Conference 2022*, 2022: 1322-1330.
- [31] Liu Z H, Luo Y, Zang Z L, et al. Surrogate Representation Learning with Isometric Mapping for Gray-Box Graph Adversarial Attacks[C]. *The Fifteenth ACM International Conference on Web Search and Data Mining*, 2022: 591-598.
- [32] Chen L, Peng Q B, Li J T, et al. Neighboring Backdoor Attacks on Graph Convolutional Network[EB/OL]. 2022: arXiv: 2201.06202. <https://arxiv.org/abs/2201.06202>.
- [33] Ma J Q, Deng J W, Mei Q Z. Adversarial Attack on Graph Neural Networks as an Influence Maximization Problem[C]. *The Fifteenth ACM International Conference on Web Search and Data Mining*, 2022: 675-685.
- [34] Xu J, Picke S. Poster: Clean-label Backdoor Attack on Graph Neural Networks[C]. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022: 3491-3493.
- [35] Huang Q, Yamada M, Tian Y, et al. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(7): 6968-6972.
- [36] Dai E Y, Lin M H, Zhang X, et al. Unnoticeable Backdoor Attacks on Graph Neural Networks[C]. *The ACM Web Conference 2023*, 2023: 2263-2273.
- [37] Rekavandi A M, Farokhi F, Ohrimenko O, et al. Certified Adversarial Robustness via Randomized A-Smoothing for Regression Models[C]. *The 38th International Conference on Neural Information Processing Systems*, 2025: 134127-134150.
- [38] Jia R, Raghunathan A, Gökseel K, et al. Certified Robustness to Adversarial Word Substitutions[C]. *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019: 4127-4140.
- [39] Qin Y, Carlini N, Goodfellow I, et al. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition[EB/OL]. 2019: arXiv: 1903.10346. <https://arxiv.org/abs/1903.10346>.
- [40] Hadi S, Grey Y, Jerry L, et al. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers[C]. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019: 11292-11303.
- [41] Yan S, Jan S, Aleksandar B, et al. Hierarchical Randomized Smoothing[C]. *Advances in Neural Information Processing Systems*, 2023.
- [42] Ren X L, Lyu L Y. Review of Ranking Nodes in Complex Networks[J]. *Chinese Science Bulletin*, 2014, 59(13): 1175-1197.
(任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. *科学通报*, 2014, 59(13): 1175-1197.)
- [43] Xu K, Hu W H, Leskovec J, et al. How Powerful Are Graph Neural Networks? [EB/OL]. 2018: arXiv: 1810.00826. <https://arxiv.org/abs/1810.00826>.



伍晓洁 于 2022 年在华中科技大学计算机科学与技术专业获得学士学位。现在国防科技大学计算机科学与技术专业攻读硕士学位。研究领域为图对抗学习、图深度学习。研究兴趣包括图对抗攻击、图异常检测。Email: wuxiaojie1015@nudt.edu.cn



刘强 于 2014 年在国防科技大学计算机科学与技术专业获得博士学位。现任国防科技大学计算机学院副教授, 硕士生导师。中国计算机学会(CCF)会员(49698M)。研究领域为网络安全、人工智能安全。研究兴趣包括网络安全、深度学习对抗样本、图学习安全。Email: qiangliu06@nudt.edu.cn



王煜恒 于 2022 年在国防科技大学计算机科学与技术专业获得学士学位。现在国防科技大学计算机科学与技术专业攻读硕士学位。研究领域为图对抗学习、图深度学习。研究兴趣包括图神经网络的鲁棒性以及图对抗样本检测。Email: wyh_0@nudt.edu.cn



付章杰 于 2012 年在湖南大学计算机科学与技术专业获得博士学位。现任南京信息工程大学计算机学院、网络空间安全学院执行院长, 教授, 博士生导师。中国计算机学会(CCF)高级会员(38656S)。研究领域为数字取证、区块链安全、人工智能安全等。Email: wwwfzj@126.com