图神经网络的标签翻转对抗攻击

吴翼腾，刘伟，于洪涛
（信息工程大学，河南 郑州 450002）

摘 要：为扩展图神经网络对抗攻击类型以填补相关研究空白，提出了评估图神经网络对标签噪声稳健性的标签翻转对抗攻击方法。将对抗攻击的有效性机理提炼为矛盾数据假设、参数差异假设和同分布假设等 3 种基本假设，并基于 3 种假设建立标签翻转对抗攻击模型。采用基于梯度的攻击方法，理论证明了基于参数差异假设模型的攻击梯度与基于同分布假设模型的攻击梯度相同，建立 2 种攻击方法的等价关系。设计实验对比分析了基于不同假设建立模型的优势和不足；大量实验验证了标签翻转攻击模型的有效性。

关键词：图神经网络；对抗攻击；标签翻转；攻击假设；稳健性
中图分类号：TP18
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021167

Label flipping adversarial attack on graph neural network

WU Yiteng, LIU Wei, YU Hongtao
Information Engineering University, Zhengzhou 450002, China

Abstract: To expand the adversarial attack types of graph neural networks and fill the relevant research gaps, label flipping attack methods were proposed to evaluate the robustness of graph neural network aimed at label noise. The effectiveness mechanisms of adversarial attacks were summarized as three basic hypotheses, contradictory data hypothesis, parameter discrepancy hypothesis and identically distributed hypothesis. Based on the three hypotheses, label flipping attack models were established. Using the gradient oriented attack methods, it was theoretically proved that attack gradients based on the parameter discrepancy hypothesis were the same as gradients of identically distributed hypothesis, and the equivalence between two attack methods was established. Advantages and disadvantages of proposed models based on different hypotheses were compared and analyzed by experiments. Extensive experimental results verify the effectiveness of the proposed attack models.

Keywords: graph neural network, adversarial attack, label flipping, attack hypothesis, robustness

1 引言

对抗攻击指有目的地对输入数据施加微小扰动，以使学习模型输出错误的预测结果[1]。本文认为与对抗攻击直接关联的研究可上溯至 20 世纪 70 年代的统计诊断[2]。统计诊断最早研究了学习模型的微小扰动对统计推断带来的影响。近年来，随着深度学习的发展，其应用安全受到广泛关注。Szegedy 等[3]研究图像数据的卷积神经网络（CNN, convolutional neural network）安全问题时提出了“对抗样本”概念[4]。Zügner 等[5]提出处理图数据的深度学习模型图神经网络（GNN, graph neural network）的

收稿日期：2021−05−06；**修回日期**：2021−08−01
通信作者：于洪涛，yht_ndsc@126.com
基金项目：国家自然科学基金创新研究群体基金资助项目（No.61521003）；国家重点研发计划基金资助项目（No.2016QY03D0502）；郑州市协同创新重大专项基金资助项目（No.162/32410218）
Foundation Items: Foundation for Innovative Research Groups of The National Natural Science Foundation of China (No.61521003), The National Key Research and Development Program of China (No.2016QY03D0502), Zhengzhou City Collaborative Innovation Major Project (No.162/32410218)

对抗攻击。

图神经网络对抗攻击研究正处于快速发展阶段，相关研究成果活跃于国际学术会议[6-8]。分析最新研究成果发现，当前研究存在扰动类型不足、前提假设单一等问题。现有研究的扰动类型通常仅考虑增删节点和连边；攻击假设通常仅基于矛盾数据假设，具体说明如下。

1) 扰动类型不足。现有图神经网络对抗攻击的扰动类型[9-10]主要是特征扰动、增删连边和节点注入[5,11-13]，没有考虑将训练数据中的特定样本标签翻转为其他类别导致模型预测错误的标签翻转攻击。标签翻转攻击已被其他数据类型的对抗攻击场景广泛研究。例如，统计诊断[2,14]的经典模型均值漂移模型和局部影响分析模型[15-17]都详细研究了因变量扰动对模型统计推断的影响，是标签翻转攻击的最初形式。张宏坡等[18]提出了一种基于熵值法的标签翻转攻击方法，来评估朴素贝叶斯分类器对标签噪声的稳健性。Muñoz-González 等[19]实现了针对深度神经网络的标签翻转数据投毒攻击。文献[20]针对基于图的半监督学习模型（区别于图神经网络）建立了统一的标签翻转攻击架构。然而，针对图神经网络的对抗攻击还未见标签翻转攻击这种扰动类型。

2) 前提假设单一。现有图神经网络对抗攻击通常仅基于矛盾数据假设建立攻击模型，而未考虑攻击前后模型训练参数的显著差异以及测试数据和训练数据分布的一致性。本文基于以下 3 种攻击假设展开研究。①文献[5,8-9,13,21]将投毒攻击建模为寻求扰动方法，在训练集上构建一组存在矛盾的训练数据，使重训练的图神经网络在训练集上的损失最大，以降低测试数据预测准确率。本文将现有攻击方法概括为基于矛盾数据假设的攻击方法。②矛盾数据假设不能很好地对过拟合攻击场景建模。本文受统计诊断[2,14,22]研究工作的启发，引入参数差异假设，即“有效攻击前后图神经网络训练参数应该具有较大差异”的假设建立攻击模型。③对抗攻击方法忽视了机器学习最基本最常见的前提假设——同分布假设，即“随机划分的训练集和测试集应该具有相同分布”的假设。

本文基于 3 种攻击假设分别建立图神经网络的标签翻转攻击模型，以期在不同数据分布条件下分析图神经网络对标签噪声的敏感性和潜在安全漏洞。图神经网络应用广泛，它不仅应用于节点分类、图分类、链路预测等复杂网络任务，也可应用于文本分类、关系抽取等自然语言处理任务，还可应用于图像分类、目标检测等计算机视觉任务。在上述应用中，图神经网络的训练数据通常需要人工收集和标记，实际应用中多采用众包技术收集和标记数据。该收集和标记方式成本低廉且方便快捷；但无法充分保证标签质量，进而导致数据中存在不可避免的标签噪声，影响图神经网络的学习过程[18]，更无法避免精心设计的投毒数据。鉴于此，图神经网络标签翻转攻击的研究意义是从攻击者的角度研究标签翻转攻击，可以预知图神经网络在各项任务中的安全威胁，评估图神经网络对标签噪声的敏感性，为标签噪声的检出和过滤、设计稳健的图神经网络提供理论基础。本文主要工作如下。

1) 针对图神经网络对抗攻击扰动类型不足的问题，提出评估图神经网络对标签噪声稳健性的标签翻转对抗攻击方法。

2) 针对图神经网络对抗攻击前提假设单一的问题，首先基于经典的矛盾数据假设建立标签翻转攻击模型；然后引入参数差异假设和同分布假设，建立标签翻转攻击模型。

3) 理论分析证明了在一定条件下，基于同分布假设模型的攻击梯度与基于参数差异假设模型的攻击梯度相同，从而建立了 2 种攻击方法的等价关系，进一步增强了模型攻击机理的可解释性。

4) 实验对比分析了 3 种基本假设对应的损失度量在标签翻转攻击中的优势和不足；大量实验验证了标签翻转攻击模型的有效性。

2 基本概念

图表示为 $G(V,E)$，其中，V 为节点集合，E 为连边集合。设节点数 $|V|=N$，则无权无向图可用对称的邻接矩阵 $\boldsymbol{A}=\{0,1\}^{N\times N}$ 表示，$\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}$。图中每个节点有 n 维特征向量，节点特征用矩阵 $\boldsymbol{X}=\{0,1\}^{N\times n}$ 表示。文献[23-25]将图神经网络简化为 SGC（simple graph convolution），它具有低通滤波器的作用。本文以 SGC 为攻击和研究对象。SGC 模型的表达式为

$$\boldsymbol{Y}_{\mathrm{out}}=\mathrm{softmax}(\boldsymbol{\mathcal{A}}\boldsymbol{X}\boldsymbol{W}) \tag{1}$$

其中，$\boldsymbol{Y}_{\mathrm{out}}$ 为 SGC 模型输出，$\boldsymbol{\mathcal{A}}$ 为滤波矩阵，$\boldsymbol{X}$ 为输入特征向量，$\boldsymbol{W}$ 为参数矩阵。在文献[25]中，$\boldsymbol{\mathcal{A}}$ 的形式通常为

$$\boldsymbol{\mathcal{A}}=\tilde{\boldsymbol{L}}^{k},\tilde{\boldsymbol{L}}=\boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}},\tilde{\boldsymbol{A}}=\boldsymbol{I}+\boldsymbol{A},\tilde{\boldsymbol{D}}=\mathrm{diag}(\tilde{\boldsymbol{A}}\mathbf{1}) \quad (2)$$

其中，$\boldsymbol{I}$ 为单位矩阵，$\mathbf{1}$ 表示元素全为 1 的列向量。设 $\boldsymbol{Z}=\boldsymbol{\mathcal{A}}\boldsymbol{X}$，记 softmax(·) 为 $\sigma(\cdot)$，$\boldsymbol{Y}$ 表示 one-hot 编码的标签矩阵。使用交叉熵损失函数，并将其表达为矩阵形式为

$$\mathcal{L}=-\mathrm{tr}[\boldsymbol{Y}^{\mathrm{T}}\ln[\sigma(\boldsymbol{Z}\boldsymbol{W})]] \quad (3)$$

本文研究非指定目标、标签翻转的数据投毒攻击。非指定目标攻击不指定具体的一个或几个攻击目标，需要使测试集整体的预测准确率下降；标签翻转攻击允许将特定的训练样本标签翻转为其他类别；投毒攻击允许图神经网络对投毒的训练数据重新训练，使重训练的图神经网络在测试集的预测准确率下降。

3　标签翻转对抗攻击模型

3.1　基于矛盾数据假设的攻击模型

文献[5,8-9,13,21]建立了连边扰动的图神经网络投毒攻击模型。按上述文献定义，攻击方法可以统一概括为以下约束优化问题

$$\hat{\boldsymbol{A}}=\arg\max_{\hat{\boldsymbol{A}}}\mathcal{L}(\boldsymbol{W}^{*};\hat{\boldsymbol{A}}) \quad (4)$$

$$\text{s.t. } \boldsymbol{W}^{*}=\arg\min_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W};\hat{\boldsymbol{A}}),\left\|\hat{\boldsymbol{A}}-\boldsymbol{A}\right\|_{0}\leqslant\delta \quad (5)$$

其中，$\boldsymbol{A}$ 为原始邻接矩阵，$\hat{\boldsymbol{A}}$ 为扰动后的邻接矩阵，$\|\cdot\|_0$ 为矩阵中非 0 元素的个数，δ 为扰动开销，$\boldsymbol{W}^{*}$ 为扰动后得到的训练参数。投毒攻击允许对参数重新训练，是双层优化问题。

从上述投毒攻击的统一形式化表述中可以看出，现有攻击方法都是构造扰动后的样本数据，使图神经网络在扰动后数据集上损失函数达到最大，即图神经网络不能很好地拟合扰动后的训练集，扰动后的训练集中存在矛盾的训练数据。根据上述分析，现有方法可概括为基于“有效攻击的训练集中存在矛盾的训练数据”假设（简称矛盾数据假设）的攻击方法。

基于矛盾数据假设，图神经网络的标签翻转攻击模型可以表述为以下约束优化问题

$$\hat{\boldsymbol{Y}}=\arg\max_{\hat{\boldsymbol{Y}}}\mathcal{L}_{C}(\boldsymbol{W}^{*};\hat{\boldsymbol{Y}}) \quad (6)$$

$$\text{s.t. } \boldsymbol{W}^{*}=\arg\min_{\boldsymbol{W}}\mathcal{L}_{C}(\boldsymbol{W};\hat{\boldsymbol{Y}}),\left\|\hat{\boldsymbol{Y}}-\boldsymbol{Y}\right\|_{0}\leqslant\delta \quad (7)$$

其中，$\mathcal{L}_{C}=-\mathrm{tr}[\hat{\boldsymbol{Y}}^{\mathrm{T}}\ln[\sigma(\boldsymbol{Z}\boldsymbol{W})]]$ 表示基于矛盾数据假设的攻击损失函数和模型正向训练采用的损失函数。攻击模型分为 2 个阶段，即式(6)所示的投毒攻击和式(7)所示的对抗训练。基于矛盾数据假设，这 2 个阶段损失函数相同。

3.2　基于参数差异假设的攻击模型

受统计诊断学科启发，基于“有效攻击前后图神经网络模型参数应该具有较大差异”的假设（简称参数差异假设），本文建立图神经网络的标签翻转投毒攻击模型，可表示为如下约束优化问题

$$\hat{\boldsymbol{Y}}=\arg\max_{\hat{\boldsymbol{Y}}}\left\|\boldsymbol{W}^{*}-\boldsymbol{W}_{0}\right\|_{M}^{2} \quad (8)$$

$$\text{s.t. } \boldsymbol{W}^{*}=\arg\min_{\boldsymbol{W}}\mathcal{L}_{C}(\boldsymbol{W};\hat{\boldsymbol{Y}}),\left\|\hat{\boldsymbol{Y}}-\boldsymbol{Y}\right\|_{0}\leqslant\delta \quad (9)$$

其中，$\boldsymbol{W}_{0}=\arg\min_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W};\boldsymbol{Y})$ 为图神经网络的最佳训练参数；$\|\cdot\|_{M}^{2}$ 为矩阵的某种范数平方，用以衡量参数差异；$\boldsymbol{W}^{*}$ 为对邻接矩阵或特征矩阵扰动后图卷积网络的训练参数。

为衡量参数差异，从统计诊断经典文献[2,14-17,22,26]中引入 Cook 距离作为 $\|\cdot\|_{M}^{2}$ 度量攻击前后的参数差异。

定义 1　Cook 距离。参数矩阵 $\boldsymbol{W}^{*}$ 与 $\boldsymbol{W}_{0}$ 的 Cook 距离 CD 定义为

$$\mathrm{CD}=\mathrm{vec}^{\mathrm{T}}(\boldsymbol{W}^{*}-\boldsymbol{W}_{0})\nabla_{\boldsymbol{W}}^{2}\mathcal{L}(\boldsymbol{W}^{*})\mathrm{vec}(\boldsymbol{W}^{*}-\boldsymbol{W}_{0}) \quad (10)$$

其中，vec(·)表示矩阵按列优先拉直为列向量。

图神经网络通过优化求解算法求得使损失函数 $\mathcal{L}$ 达到最小值的参数矩阵 $\boldsymbol{W}_{0}$，即 $\boldsymbol{W}_{0}$ 满足 $\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_{0})=\mathbf{0}$。实用中 $\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_{0})\approx\mathbf{0}$。对 $\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})$ 在 $\boldsymbol{W}^{*}$ 处进行一阶泰勒展开并取 $\boldsymbol{W}=\boldsymbol{W}_{0}$ 得

$$\mathrm{vec}\left[\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_{0})\right]=\mathrm{vec}\left[\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{*})\right]+\nabla_{\boldsymbol{W}}^{2}\mathcal{L}(\boldsymbol{W}^{*})\mathrm{vec}(\boldsymbol{W}_{0}-\boldsymbol{W}^{*})+R(\mathrm{vec}(\boldsymbol{W}_{0}-\boldsymbol{W}^{*}))\approx\mathbf{0} \quad (11)$$

略去高阶项 $R(\mathrm{vec}(\boldsymbol{W}_{0}-\boldsymbol{W}^{*}))$ 可得

$$\mathrm{vec}\left(\boldsymbol{W}_{0}\right)\approx\mathrm{vec}\left(\boldsymbol{W}^{*}\right)+\left[-\nabla_{\boldsymbol{W}}^{2}\mathcal{L}(\boldsymbol{W}^{*})\right]^{-1}\mathrm{vec}\left[\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{*})\right] \quad (12)$$

基于以上分析，衡量参数差异的 Cook 距离可近似表示为

$$\mathrm{CD}=\mathrm{vec}^{\mathrm{T}}[\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{*})]\left[\nabla_{\boldsymbol{W}}^{2}\mathcal{L}(\boldsymbol{W}^{*})\right]^{-1}\mathrm{vec}[\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^{*})] \quad (13)$$

容易说明，矩阵 $\nabla_{\boldsymbol{W}}^{2}\mathcal{L}(\boldsymbol{W}^{*})$ 未必可逆。采用 Levenberg-Marquardt 修正或岭正则化方法[27-28]，将式(3)的损失函数 $\mathcal{L}$ 添加正则项 $\frac{\lambda}{2}\left\|\mathrm{vec}(\boldsymbol{W}^{*})\right\|_{2}^{2}$（$\lambda>0$）可得

$$\tilde{\mathcal{L}} = \mathcal{L} + \frac{\lambda}{2}\left\|\mathrm{vec}(\boldsymbol{W}^*)\right\|_2^2 \tag{14}$$

容易求得

$$\nabla_{\boldsymbol{W}}^2 \tilde{\mathcal{L}}(\boldsymbol{W}^*) = \nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{W}^*) + \lambda \boldsymbol{I} \tag{15}$$

其中，$\boldsymbol{I}$ 是与 $\nabla_{\boldsymbol{W}}^2 \mathcal{L}(\boldsymbol{W}^*)$ 同型的单位阵。此时 $\nabla_{\boldsymbol{W}}^2 \tilde{\mathcal{L}}(\boldsymbol{W}^*)$ 可逆。基于此，修正后的 Cook 距离 $\widetilde{\mathrm{CD}}$ 表示为

$$\widetilde{\mathrm{CD}} = \mathrm{vec}^{\mathrm{T}}[\nabla_{\boldsymbol{W}} \tilde{\mathcal{L}}(\boldsymbol{W}^*)]\left[\nabla_{\boldsymbol{W}}^2 \tilde{\mathcal{L}}(\boldsymbol{W}^*)\right]^{-1} \mathrm{vec}[\nabla_{\boldsymbol{W}} \tilde{\mathcal{L}}(\boldsymbol{W}^*)] \tag{16}$$

攻击模型可以表示为

$$\hat{\boldsymbol{Y}} = \arg\max_{\hat{\boldsymbol{Y}}} \widetilde{\mathrm{CD}}(\boldsymbol{W}^*) \tag{17}$$

$$\text{s.t. } \boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \mathcal{L}_C(\boldsymbol{W}; \hat{\boldsymbol{Y}}), \left\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\right\|_0 \leqslant \delta \tag{18}$$

基于参数差异假设的攻击模型同样分为投毒攻击和对抗训练 2 个阶段。在式(17)所示的投毒攻击阶段，使用修正后的 Cook 距离 $\widetilde{\mathrm{CD}}$ 作为损失函数；在式(18)所示的对抗训练阶段，训练数据为扰动后的训练数据，因此与基于矛盾数据假设的损失函数相同。

3.3 基于同分布假设的攻击模型

图神经网络对抗攻击问题忽略了一种机器学习最基本、最常见的前提假设，即同分布假设。无论图神经网络还是其他深度学习或机器学习方法，通常首先基于同分布假设展开研究，即认为训练集的数据分布与测试集一致，因此这些方法可以在训练集中学习数据模式，并有效地迁移至测试集。分析实际攻击场景，投毒攻击的目的是通过污染训练数据，使其训练的图神经网络在未污染的测试集上的错误率（损失函数）达到最大。若考虑训练测试集的同分布假设，相应的攻击模型应该改为通过污染训练数据，使其训练的图神经网络在未污染的训练集上的错误率（损失函数）达到最大。

根据以上分析，基于“随机划分的训练数据和测试数据对于图神经网络模型应该具有同分布性质”的假设（简称同分布假设），建立标签翻转投毒攻击模型，可表示为如下约束优化问题

$$\hat{\boldsymbol{Y}} = \arg\max_{\hat{\boldsymbol{Y}}} \mathcal{L}(\boldsymbol{W}^*; \boldsymbol{Y}) \tag{19}$$

$$\text{s.t. } \boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \mathcal{L}_C(\boldsymbol{W}; \hat{\boldsymbol{Y}}), \left\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\right\|_0 \leqslant \delta \tag{20}$$

根据 3.1 节和 3.2 节分析，无论何种攻击模型，在对抗训练阶段均采用扰动后的训练数据，因此式(20)与式(7)、式(18)相同。基于同分布假设的攻击模型在投毒攻击阶段，使用式(3)的交叉熵作为损失函数。从形式上，基于同分布假设的攻击模型与基于矛盾数据假设的攻击模型的差异仅体现在损失函数式(6)和式(19)中，式(6)使用扰动后的标签矩阵，式(19)使用未扰动的标签矩阵。但二者的建模思想是不同的。本文 3.4 节将证明，在某些较弱的条件下，基于同分布假设的攻击模型与基于参数差异假设的攻击模型等价。

3.4 攻击梯度及其等价定理

本节介绍上述攻击模型的符号记法，主要涉及 $\mathcal{L}$、$\mathcal{L}_C$、$\hat{\mathcal{L}}$、CD、$\widetilde{\mathrm{CD}}$ 之间的区别。$\mathcal{L}$ 由式(3)定义，$\mathcal{L}$ 代入的标签矩阵 $\boldsymbol{Y}$ 为未扰动的标签矩阵；$\mathcal{L}_C$ 由式(6)定义，$\mathcal{L}_C$ 代入的标签矩阵 $\hat{\boldsymbol{Y}}$ 为扰动后的标签矩阵；$\tilde{\mathcal{L}}$ 由式(14)定义，为带有正则项的 $\mathcal{L}$。Cook 距离 CD 由式(13)定义；$\widetilde{\mathrm{CD}}$ 由式(16)定义，表示由 $\tilde{\mathcal{L}}$ 定义的 Cook 距离 CD。

设图神经网络(1)采用梯度下降法训练

$$\boldsymbol{W}^t = \boldsymbol{W}^{t-1} - \alpha \nabla_{\boldsymbol{W}} \mathcal{L}_C(\boldsymbol{W}^*) \tag{21}$$

经过 $t = t_0$ 轮训练，可得模型的训练参数 $\boldsymbol{W}^* = \boldsymbol{W}^{t_0}$。具体地，可求得损失函数对参数的梯度为

$$\nabla_{\boldsymbol{W}} \mathcal{L}_C(\boldsymbol{W}) = \boldsymbol{Z}^{\mathrm{T}}[\sigma(\boldsymbol{Z}\boldsymbol{W}) - \hat{\boldsymbol{Y}}] \tag{22}$$

代入式(21)，有

$$\boldsymbol{W}^* = \boldsymbol{W}^0 - \alpha \sum_{t=0}^{t_0-1} \boldsymbol{Z}^{\mathrm{T}}[\sigma(\boldsymbol{Z}\boldsymbol{W}^t) - \hat{\boldsymbol{Y}}] \tag{23}$$

对于离散的标签矩阵 $\boldsymbol{Y}$，优化问题式(6)、式(17)、式(19)属于 NP 难问题。现有文献主要依据攻击梯度实施扰动。攻击梯度是实现有效攻击的主要依据。

从攻击梯度的角度，定理 1 给出了参数差异假设与同分布假设的等价性。

定理 1 设

$$\widetilde{\mathrm{CD}} = \mathrm{vec}^{\mathrm{T}}[\nabla_{\boldsymbol{W}} \tilde{\mathcal{L}}(\boldsymbol{W}^*)]\left[\nabla_{\boldsymbol{W}}^2 \tilde{\mathcal{L}}(\boldsymbol{W}^*)\right]^{-1} \mathrm{vec}[\nabla_{\boldsymbol{W}} \tilde{\mathcal{L}}(\boldsymbol{W}^*)]$$

$$\tilde{\mathcal{L}} = \mathcal{L} + \frac{\lambda}{2}\left\|\mathrm{vec}(\boldsymbol{W}^*)\right\|_2^2$$

将权矩阵 $\nabla_{\boldsymbol{W}}^2 \tilde{\mathcal{L}}(\boldsymbol{W}^*)$ 视为常数矩阵，则有以下关系式成立

$$\nabla_{\hat{\boldsymbol{Y}}} \widetilde{\mathrm{CD}} = 2\nabla_{\hat{\boldsymbol{Y}}} \tilde{\mathcal{L}} \tag{24}$$

证明 设 $v(\boldsymbol{W}^*) = \mathrm{vec}[\nabla_{\boldsymbol{W}} \tilde{\mathcal{L}}(\boldsymbol{W}^*)]$，则

$$\mathrm{vec}(\nabla_{\hat{\boldsymbol{Y}}}\widetilde{\mathrm{CD}})=\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Y}}}\frac{\partial v(\boldsymbol{W}^*)}{\partial \boldsymbol{W}^*}\frac{\partial \widetilde{\mathrm{CD}}}{\partial v(\boldsymbol{W}^*)} \tag{25}$$

$$\frac{\partial \widetilde{\mathrm{CD}}}{\partial v(\boldsymbol{W}^*)}=2\left[\nabla^2_{\boldsymbol{W}}\tilde{\mathcal{L}}(\boldsymbol{W}^*)\right]^{-1}v(\boldsymbol{W}^*) \tag{26}$$

$$\frac{\partial v(\boldsymbol{W}^*)}{\partial \boldsymbol{W}^*}=\nabla^2_{\boldsymbol{W}}\tilde{\mathcal{L}}(\boldsymbol{W}^*) \tag{27}$$

$$\frac{\partial v(\boldsymbol{W}^*)}{\partial \boldsymbol{W}^*}\frac{\partial \widetilde{\mathrm{CD}}}{\partial v(\boldsymbol{W}^*)}=$$

$$2\nabla^2_{\boldsymbol{W}}\tilde{\mathcal{L}}(\boldsymbol{W}^*)\left[\nabla^2_{\boldsymbol{W}}\tilde{\mathcal{L}}(\boldsymbol{W}^*)\right]^{-1}v(\boldsymbol{W}^*)=2v(\boldsymbol{W}^*) \tag{28}$$

$$\mathrm{vec}(\nabla_{\hat{\boldsymbol{Y}}}\widetilde{\mathrm{CD}})=\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Y}}}2v(\boldsymbol{W}^*) \tag{29}$$

$$\nabla_{\hat{\boldsymbol{Y}}}\widetilde{\mathrm{CD}}=2\nabla_{\hat{\boldsymbol{Y}}}\tilde{\mathcal{L}} \tag{30}$$

证毕。

定理 1 表明，衡量参数差异的 Cook 距离的攻击梯度与基于同分布假设攻击模型的攻击梯度相同，因此从攻击梯度的意义上，上述 2 个模型等价。从而说明基于同分布假设的攻击方法的物理意义也是诱导图神经网络训练出一组异于原始参数的训练参数。

直接从基于同分布假设的损失函数出发，也可定性分析得出相同的结论：攻击后式(19)的损失函数 $\mathcal{L}$ 增大，但攻击前后损失函数中 $\boldsymbol{Y}$ 、$\boldsymbol{\mathcal{A}}$ 、$\boldsymbol{X}$ 保持不变，仅参数 $\boldsymbol{W}^*$ 发生改变，表明攻击前后图神经网络的训练参数具有较大差异。该结论解释了文献[13]中提及但未从理论上证明的实验现象。

等价性定理表明，求解攻击梯度时可采用带有正则项的 $\tilde{\mathcal{L}}$ 代替 Cook 距离 $\widetilde{\mathrm{CD}}$。由此，基于 3 种假设模型的攻击梯度可求解如下。

矛盾数据假设为

$$\mathrm{vec}(\nabla_{\hat{\boldsymbol{Y}}}\mathcal{L}_{\boldsymbol{C}})=\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Y}}}\frac{\partial \mathcal{L}_{\boldsymbol{C}}}{\partial \boldsymbol{W}^*} \tag{31}$$

参数差异假设为

$$\mathrm{vec}(\nabla_{\hat{\boldsymbol{Y}}}\tilde{\mathcal{L}})=\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Y}}}\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{W}^*}=$$

$$\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Y}}}\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^*}+\frac{\partial \frac{\lambda}{2}\left\|\mathrm{vec}(\boldsymbol{W}^*)\right\|_2^2}{\partial \boldsymbol{W}^*}\right)=\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Y}}}\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^*}+\mathrm{vec}(\lambda \boldsymbol{I})\right) \tag{32}$$

同分布假设为

$$\mathrm{vec}(\nabla_{\hat{\boldsymbol{Y}}}\mathcal{L})=\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Y}}}\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^*} \tag{33}$$

3.5　攻击算法

根据上述分析，以及式(31)～式(33)求得的攻击梯度，采用贪心算法实施扰动，可设计图神经网络的标签翻转对抗攻击算法如算法 1 所示。

算法 1　标签翻转对抗攻击算法

输入　邻接矩阵 $\boldsymbol{A}$，特征矩阵 $\boldsymbol{X}$，标签 $\boldsymbol{Y}$，攻击点数 n

输出　扰动列表 disturb_list

1) disturb_list = [];

2) for i = 1:n

3)　根据 disturb_list 更新标签矩阵 $\hat{\boldsymbol{Y}}$

4)　根据式(7)、式(18)和式(20)重训练图神经网络得到参数 $\boldsymbol{W}^*$

5)　根据式(31)、式(32)或式(33)计算攻击梯度 $\boldsymbol{G}$

6)　根据攻击梯度 $\boldsymbol{G}$ 更新扰动列表

7) end for

8) 返回 disturb_list

4　实验

4.1　对比实验

实验采用型号为 TITAN Xp 的 GPU 显卡，运行环境为 ubuntu 16.04 系统、cuda10.0、Python3.6 以及 Pytorch1.4。实验采用的数据集为 Polblogs[29]、Cora_ml、Cora[30]、Citeseer[31]，数据集的统计特性如表 1 所示。由于现有研究尚未考虑标签翻转攻击，将所提 3 种标签翻转攻击方法与随机翻转标签攻击方法 Random 以及 2 种经典的连边扰动投毒攻击方法 Mettack[13]和 Min-max[32]进行对比实验。Mettack 采用 approximating meta-gradients[13]。Min-max 的攻击方式设置为 negative cross-entropy[32]。本文提出的标签翻转攻击方法参数设置与 Mettack 保持一致。具体地，对于数据集 Polblogs、Cora_ml，图神经网络 SGC 正向训练的学习率取 $\alpha=0.1$；对于数据集 Cora、Citeseer，学习率取 $\alpha=0.01$。式(2)中 SGC 模型幂指数为 k = 1 和 k = 2。Mettack 和 Min-max 攻击方法允许对训练集中连边总数的 5%进行攻击，并控制扰动后的节点度不超过原始网络中节点度的最大值；标签翻转攻击方法 Random、$\mathcal{L}_{\boldsymbol{C}}$、$\tilde{\mathcal{L}}$、$\mathcal{L}$ 允许对训练数据中 5%的标签翻转至其他类别。基于参数差异假设的标签翻转攻击模型正则项 $\lambda=0.001$。实验划分数据集中 40%为训练集，60%为测试集，数据集随机划分 20 次，记录 20 次 SGC

初始准确率平均值和攻击后准确率平均值，实验结果如表 2 所示。表 2 中准确率的最小值用黑体标出。

表 1 数据集统计特性

数据集	节点数	连边数	特征维数	分类数
Polblogs	1 222	16 714	1 490	2
Cora_ml	2 810	7 981	2 879	7
Cora	2 485	5 069	1 433	7
Citeseer	2 110	3 668	3 703	6

表 2 准确率

k	方法	Polblogs	Cora_ml	Cora	Citeseer
1	未扰动	94.54%	86.45%	85.04%	74.20%
	Min-max	92.62%	84.87%	82.38%	73.40%
	Mettack	93.32%	86.03%	84.81%	74.79%
	Random	94.02%	86.11%	84.20%	73.53%
	$\mathcal{L}_C$	91.87%	83.54%	80.12%	69.10%
	$\tilde{\mathcal{L}}$	**90.84%**	**79.99%**	**79.22%**	69.01%
	$\mathcal{L}$	91.07%	**79.99%**	79.25%	**68.93%**
2	未扰动	95.60%	88.05%	87.08%	74.52%
	Min-max	94.61%	87.61%	85.74%	74.12%
	Mettack	93.70%	87.26%	86.74%	74.86%
	Random	95.41%	88.04%	86.78%	74.04%
	$\mathcal{L}_C$	94.75%	87.35%	85.02%	70.77%
	$\tilde{\mathcal{L}}$	**93.56%**	82.85%	82.23%	**69.69%**
	$\mathcal{L}$	93.65%	**82.94%**	**82.18%**	69.78%

综合分析以上数据，可以得出如下结论。

1) 考虑标签翻转攻击类型，可以实现有效的投毒攻击。在相同扰动比例比较基准下，基于 3 种假设的标签翻转攻击效果优于基于 Mettack 和 Min-max 方法的连边扰动攻击效果。实验结果证明了标签翻转这一新的攻击类型的有效性。

2) 对于标签翻转攻击，随机翻转标签几乎无法实施有效攻击。对于 5%的扰动，图神经网络 SGC 的预测准确率没有明显下降，表明通过式(5)重训练，图神经网络可以抵抗随机攻击。而针对本文所提的投毒加扰方式，SGC 的预测准确率下降明显。这种恶意注入的标签噪声可能在数据标记阶段产生，图神经网络在实用中存在潜在威胁。该标签噪声同时具有不易察觉性，将在 4.2 节分析。

3) 对于本文提出的 3 种标签翻转攻击方法，本文条件下的实验结果表明，基于参数差异假设和同分布假设的标签翻转攻击模型的攻击效果优于基于矛盾数据假设的攻击效果。基于参数差异假设的攻击效果几乎与基于同分布假设的攻击效果相同，与等价性定理得出的理论结果一致。

表 2 列出了 5%的扰动各攻击方法的攻击结果。为进一步比较不同方法的攻击效果，说明扰动量对攻击效果的影响，其他实验条件不变，采用 1%~10%的扰动并记录准确率下降的平均值。选取标签翻转攻击方法 Random、$\mathcal{L}_C$ 与 $\mathcal{L}$（$\tilde{\mathcal{L}}$ 的实验结果与 $\mathcal{L}$ 类似），实验结果如图 1 所示。

总体而言，基于矛盾数据假设和同分布假设的攻击方法相比于随机翻转标签攻击有更明显的攻击效果。随机翻转标签攻击无法抵抗图神经网络重训练。无论 $k = 1$ 和 $k = 2$，基于同分布假设的攻击效果均优于基于矛盾数据假设的攻击效果，具体原因将在 4.3 节详细分析。实验结果说明了本文所提出的基于不同假设的标签翻转攻击模型的有效性和攻击假设的合理性。

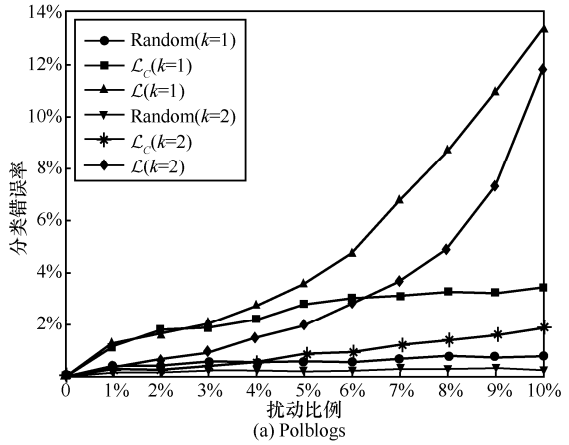
4.2 扰动的难以分辨性分析

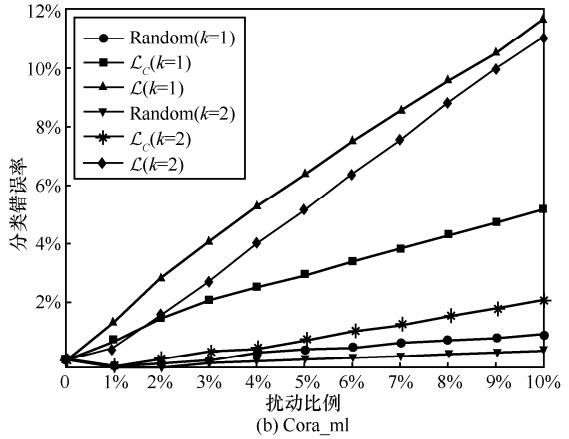
本节从图结构统计特征[33]和标签类别分布两方面分析标签翻转攻击的难以分辨性。

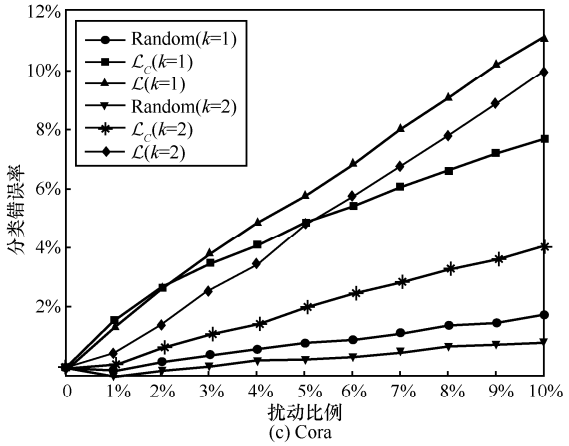
1) 图结构统计特征。标签翻转攻击不对图结构实施扰动，图结构的各项统计特征例如度分布、节点特征相似度、模体等局部结构特征均保持不变。

2) 标签类别分布。图 2 绘制了 Cora_ml 数据集 $k = 1$ 时原始标签分布和 4 种标签翻转攻击方法扰动后的标签分布（图 2 随机选取 20 次实验中的某次实验结果作为示例）。其他数据集的实验结果与之类似，实验结论相同。

观察标签类别分布可知，扰动后各类别标签分布与原始标签分布差异不大。若实际场景中需严格保持标签类别分布不变，只需对攻击算法中的扰动筛选策略稍作调整。例如，不仅依据攻击梯度大小次序筛选扰动元素，而且限定后一轮攻击翻转至前一轮的原始标签类别。如此迭代，可保持标签类别分布不变。或者，基于前文实验证明的随机翻转标签无法实施有效攻击的结论，为保持标签类别分布不变，投毒攻击后再采用随机扰动平衡标签分布（简称随机平衡策略）。基于随机平衡策略，得到 Cora_ml 数据集 $k = 1$ 时基于矛盾数据假设 $\mathcal{L}_C$ 和同分布假设 $\mathcal{L}$ 的实验结果如图 3 所示（$\tilde{\mathcal{L}}$ 的实验结果与 $\mathcal{L}$ 类似）。图 3 中同时对比了 4.1 节不使用该策略的实验结果。可以看出，加入随机平衡策略的攻击结果仍然具有可用性，与原攻击效果差异不明显。







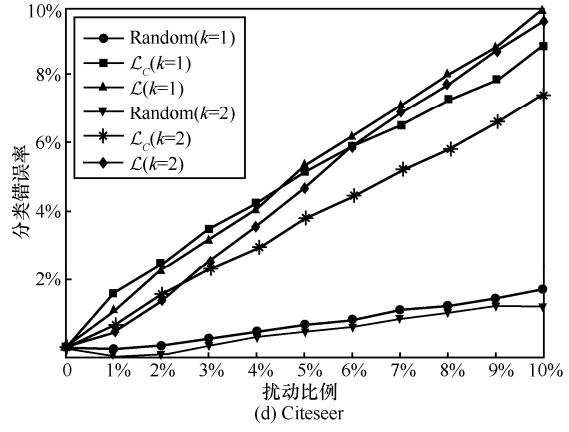


图 1　不同扰动量的攻击效果对比

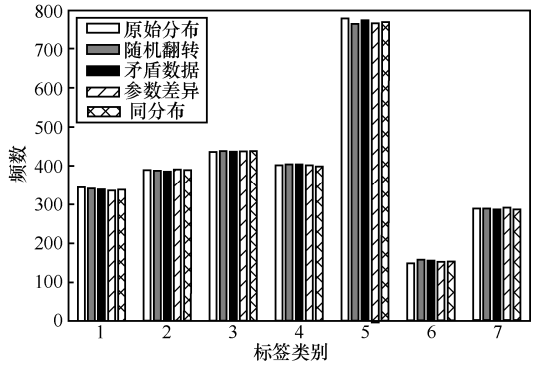


图 2　标签类别分布对比

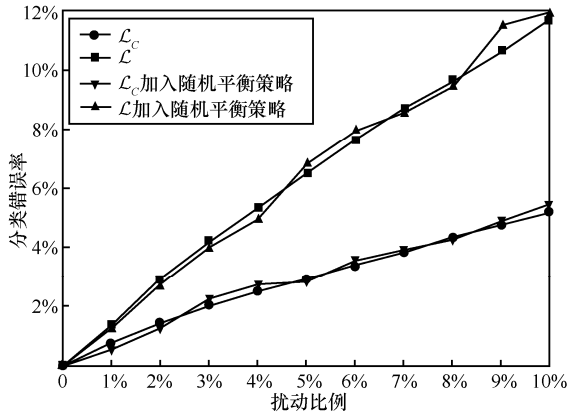


图 3　加入随机平衡策略实验结果对比

4.3　模型的比较分析

本节详细分析矛盾数据假设、参数差异假设与同分布假设及相应损失函数的合理性，并与随机翻转标签进行对比分析。实验记录各方法在扰动量为1%~10%下的训练准确率和测试准确率；对每种攻击方法（包括随机攻击 Random）均同时计算 3 种损失函数值，即 $\mathcal{L}_C,\tilde{\mathcal{L}},\mathcal{L}$（随机攻击方法没有专门的损失函数，只计算随机扰动后其他 3 种损失函数值）。实验设置与 4.1 节相同，所有数值均为 20 次实验的平均值。选取数据集 Cora_ml 的实验结果如表 3 和表 4 所示。其他数据集和不同参数下的实验结果与所列结果相似。综合分析表 3 和表 4 的数据，可得出以下结论。

1) 基于 3 种假设建模的损失函数具有有效性。对于未受扰动的原始数据，4 种攻击方法对应的基于矛盾数据假设的初始损失函数值 $\mathcal{L}_{C_1}$、$\mathcal{L}_{C_2}$、$\mathcal{L}_{C_3}$、$\mathcal{L}_{C_4}$ 和基于同分布假设的初始损失函数值 $\mathcal{L}_1$、$\mathcal{L}_2$、$\mathcal{L}_3$、$\mathcal{L}_4$ 相同，当 $k=1$ 时，约为 77.85；当 $k=2$ 时，约为 158.60。由于未对数据进行投毒扰动，标签矩阵 $\boldsymbol{Y}$ 与 $\hat{\boldsymbol{Y}}$ 相同，且初始参数 $\boldsymbol{W}^*$ 相同，因此损失函数值相同。基于参数差异假设的初始损失函数 $\tilde{\mathcal{L}}_1$、$\tilde{\mathcal{L}}_2$、$\tilde{\mathcal{L}}_3$、$\tilde{\mathcal{L}}_4$，当 $k=1$ 时，约为 102.78；当 $k=2$

时，约为 181.85。因为基于参数差异假设的损失函数与其他二者相比添加了正则项，导致损失函数的初值不同。基于 3 种假设的标签翻转攻击方法均表现出随着扰动量的增加，训练测试准确率递减、损失函数递增的趋势，证明了基于 3 种假设建立损失函数的有效性。而对于随机翻转标签攻击方法，随着扰动量增加，损失函数虽有上升趋势，但上升并不明显，尽管扰动量达到 10%，依据参数差异或同分布损失 $\tilde{\mathcal{L}}$、$\mathcal{L}$ 来衡量，损失函数值也达不到基于 3 种假设攻击方法 3%的扰动损失，而扰动后的准确率大致处于基于 3 种假设攻击方法 1%扰动量的攻击效果，随机翻转标签的攻击效果不理想。

2) 基于参数差异假设和同分布假设建模的损失函数 $\tilde{\mathcal{L}}$、$\mathcal{L}$ 相比基于矛盾数据假设建模的损失函数 $\mathcal{L}_C$ 更具有效性。从两方面分析如下。① 在相同的扰动量前提下，对比分析基于 3 种假设攻击方法的训练准确率、测试准确率和对应的损失函数值。例如，对于 5%的扰动，基于矛盾数据假设的训练准确率和测试准确率分别为 97.83%和 83.54%，基于同分布假设的训练准确率和测试准确率分别为 93.96%和 79.99%。可见在该扰动下，基于同分布假设的攻击方法更有效。然而，采用基于矛盾数据假设攻击方法计算得出的矛盾数据损失 $\mathcal{L}_{C_1}=399.91$，基于同分布假设攻击方法计算得出的矛盾数据损失 $\mathcal{L}_{C_3}=197.10$，损失函数值与攻击效果不一致；而对应 2 种方法的同分布损失 $\mathcal{L}_1=279.80$，$\mathcal{L}_3=298.47$，损失函数值与攻击效果一致。对于其他扰动量可得出同样结论。基于参数差异假设的有关结果与同分布假设相同。② 分析随机翻转标签的矛盾数据损失 $\mathcal{L}_C$ 和同分布损失 $\mathcal{L}$。对于相同的扰动量，矛盾数据损失值 $\mathcal{L}_C$ 大于同分布损失值 $\mathcal{L}$，$k=2$ 时更显著。原因是由式(6)可知矛盾数据损失 $\mathcal{L}_C$ 含有 2 个变量，即扰动后的标签矩阵 $\hat{\boldsymbol{Y}}$ 和重训练参数 $\boldsymbol{W}^*$。而根据式(19)，同分布损失 $\mathcal{L}$ 只含有一个变量，即重训练参数 $\boldsymbol{W}^*$。即使训练参数 $\boldsymbol{W}^*$ 不变，矛盾数

表 3　　$k=1$ 时数据集 Cora_ml 基于不同假设的比较分析

假设	损失函数	扰动量										
		无扰动	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
矛盾数据	训练准确率	99.65%	98.51%	98.45%	98.32%	98.13%	97.83%	97.53%	97.18%	96.84%	96.42%	95.95%
	测试准确率	86.45%	85.71%	85.04%	84.43%	83.96%	83.54%	83.10%	82.61%	82.12%	81.70%	81.25%
	$\mathcal{L}_{C_1}$	77.84	243.80	295.95	337.10	371.11	399.91	428.22	454.47	479.39	503.16	527.14
	$\tilde{\mathcal{L}}_1$	102.77	195.60	224.89	250.35	273.50	295.88	318.07	339.39	360.19	380.71	404.29
	$\mathcal{L}_1$	77.84	180.13	209.41	234.72	257.65	279.80	301.76	322.91	343.53	363.89	387.28
参数差异	训练准确率	99.66%	97.83%	96.78%	95.79%	94.93%	93.94%	92.89%	91.98%	91.10%	90.23%	89.30%
	测试准确率	86.45%	85.05%	83.64%	82.28%	81.05%	79.99%	78.85%	77.80%	76.78%	75.78%	74.65%
	$\mathcal{L}_{C_2}$	77.85	175.58	181.74	186.29	194.25	199.43	202.84	205.52	209.44	214.92	219.92
	$\tilde{\mathcal{L}}_2$	102.78	189.84	217.24	248.54	277.21	310.66	345.39	381.92	417.88	452.05	492.05
	$\mathcal{L}_2$	77.85	173.25	200.42	231.60	260.22	293.66	328.38	364.88	400.80	434.92	474.88
同分布	训练准确率	99.66%	97.85%	96.83%	95.79%	94.97%	93.96%	92.95%	91.95%	91.02%	90.25%	89.29%
	测试准确率	86.42%	85.12%	83.60%	82.28%	81.12%	79.99%	78.85%	77.80%	76.77%	75.80%	74.68%
	$\mathcal{L}_{C_3}$	77.87	173.47	181.53	186.50	192.86	197.10	200.70	202.68	206.56	212.37	216.85
	$\tilde{\mathcal{L}}_3$	102.80	189.06	216.25	248.45	278.29	315.38	351.02	389.25	425.06	460.38	501.39
	$\mathcal{L}_3$	77.87	172.48	199.48	231.60	261.39	298.47	334.14	372.40	408.17	443.43	484.41
随机攻击	训练准确率	99.66%	98.34%	98.21%	97.96%	97.81%	97.59%	97.42%	97.21%	97.01%	96.88%	96.68%
	测试准确率	86.46%	86.70%	86.59%	86.48%	86.23%	86.11%	86.00%	85.86%	85.76%	85.68%	85.56%
	$\mathcal{L}_{C_4}$	77.89	148.36	158.79	166.53	175.08	181.54	188.93	195.55	202.00	207.06	214.97
	$\tilde{\mathcal{L}}_4$	102.82	163.38	171.55	179.66	187.36	194.66	201.20	208.75	215.47	221.05	228.50
	$\mathcal{L}_4$	77.89	147.08	155.27	163.37	171.10	178.37	184.90	192.39	199.09	204.64	212.08

表 4　　**$k=2$ 时数据集 Cora_ml 基于不同假设的比较分析**

假设	损失函数	扰动量										
		无扰动	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
矛盾数据	训练准确率	96.49%	94.98%	94.88%	94.88%	94.83%	94.69%	94.58%	94.35%	94.11%	93.88%	93.52%
	测试准确率	88.05%	88.21%	87.99%	87.79%	87.68%	87.35%	87.08%	86.87%	86.56%	86.27%	85.96%
	$\mathcal{L}_{C_1}$	158.58	344.39	420.66	479.22	528.69	572.72	611.05	648.31	683.45	716.71	750.49
	$\tilde{\mathcal{L}}_1$	181.84	266.66	292.22	316.58	338.95	360.68	381.91	403.02	424.26	445.29	468.01
	$\mathcal{L}_1$	158.58	252.93	279.12	303.77	326.28	348.09	369.31	390.45	411.68	432.71	455.41
参数差异	训练准确率	96.51%	94.58%	93.38%	92.38%	91.25%	90.05%	88.83%	87.70%	86.54%	85.48%	84.47%
	测试准确率	88.06%	87.64%	86.42%	85.11%	83.91%	82.85%	81.63%	80.48%	79.23%	77.91%	76.60%
	$\mathcal{L}_{C_2}$	158.57	269.89	278.94	285.41	292.11	297.75	299.60	302.78	305.69	308.04	312.26
	$\tilde{\mathcal{L}}_2$	181.84	265.07	288.56	315.32	341.76	369.31	395.93	426.15	457.87	489.82	530.75
	$\mathcal{L}_2$	158.57	249.86	273.22	299.88	326.27	353.83	380.44	410.72	442.51	474.52	515.59
同分布	训练准确率	96.50%	94.62%	93.53%	92.42%	91.37%	90.13%	88.86%	87.72%	86.51%	85.47%	84.58%
	测试准确率	88.05%	87.71%	86.50%	85.35%	84.01%	82.94%	81.67%	80.46%	79.23%	78.05%	76.92%
	$\mathcal{L}_{C_3}$	158.61	268.73	278.24	282.96	290.70	295.65	296.99	302.46	303.10	305.86	310.32
	$\tilde{\mathcal{L}}_3$	181.87	265.31	288.51	317.72	345.52	373.91	401.51	432.71	466.63	501.98	546.42
	$\mathcal{L}_3$	158.61	250.14	273.19	302.31	330.08	358.48	386.07	417.36	451.37	486.81	531.40
随机攻击	训练准确率	96.51%	94.89%	94.80%	94.77%	94.65%	94.60%	94.55%	94.46%	94.44%	94.36%	94.28%
	测试准确率	88.05%	88.34%	88.29%	88.18%	88.12%	88.04%	88.02%	87.97%	87.86%	87.80%	87.74%
	$\mathcal{L}_{C_4}$	158.59	236.73	254.61	268.65	282.00	293.60	306.80	319.12	330.61	339.74	353.35
	$\tilde{\mathcal{L}}_4$	181.85	243.15	249.12	254.59	259.88	264.98	270.04	275.63	280.81	285.24	291.25
	$\mathcal{L}_4$	158.59	228.03	234.21	239.81	245.25	250.44	255.61	261.28	266.54	271.03	277.14

据假设引入扰动后的标签矩阵 $\hat{\boldsymbol{Y}}$ 同样会使损失函数增大。若参数差异假设或同分布假设与实际攻击场景更吻合，引入 $\hat{\boldsymbol{Y}}$ 实际上弱化了损失函数对参数差异的度量。通过以上分析可以得出结论，基于同分布假设和基于参数差异假设的损失函数与攻击效果具有更好的一致性，损失函数值越大，攻击效果越好；而基于矛盾数据假设的损失函数与攻击效果的一致性一般。从而解释了 4.1 节的实验现象，进而说明了本文场景种基于参数差异假设和基于同分布假设的攻击方法优于基于矛盾数据假设攻击方法的原因。

5　结束语

标签翻转对抗攻击在统计诊断、垃圾邮件检测、图像中的对抗样本以及基于图的半监督学习等领域得到了广泛研究。本文针对图神经网络对抗攻击扰动类型不足的问题，提出并实现了图神经网络的标签翻转对抗攻击。首先，提炼出对抗攻击有效性机理的矛盾数据假设、参数差异假设和同分布假设。然后，基于 3 种假设建立攻击模型并实验验证。有效攻击的核心是基于攻击假设建立攻击模型进而求解攻击梯度。攻击梯度是实施有效攻击的主要依据。为保持标签扰动的不易察觉性，可增加限制条件保持标签分布不变，这容易通过修改扰动筛选策略来实现。本文得出以下结论：1) 对于处理图数据的深度学习模型图神经网络，标签翻转对抗攻击具有有效性；2) 采用基于梯度的攻击方法，参数差异假设与同分布假设建立的攻击模型等价；3) 本文场景中基于参数差异假设和同分布假设的标签翻转攻击方法优于基于矛盾数据假设的攻击方法。

由于实际场景中某一样本难以界定唯一的归属类别，或样本本身存在错误标注，这可能大幅降低图神经网络模型的预测能力，因此标签翻转对抗攻击研究为图神经网络的模型诊断和稳健的图神经网络设计提供了必要前提。后续研究工作可基于标签翻转攻击原理，对图数据中的异常点、强影响点、离群点等进行检测和模型诊断，从而改善数据质量；并设计图神经网络结构，建立能够防御对抗

攻击干扰的稳健图神经网络。

参考文献：

[1] YUAN X Y, HE P, ZHU Q L, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[2] 韦博成, 鲁国斌, 史建清. 统计诊断引论[M]. 南京: 东南大学出版社, 1991.
WEI B C, LU G B, SHI J Q. Introduction to statistical diagnosis[M]. Nanjing: Southeast University Press, 1991.

[3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.

[4] 司念文, 张文林, 屈丹, 等. 基于对抗补丁的可泛化的 Grad-CAM 攻击方法[J]. 通信学报, 2021, 42(3): 23-35.
SI N W, ZHANG W L, QU D, et al. Generalized Grad-CAM attacking method based on adversarial patch[J]. Journal on Communications, 2021, 42(3): 23-35.

[5] ZÜGNER D, AKBARNEJAD A, GÜNNEMANN S. Adversarial attacks on neural networks for graph data[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2018: 2847-2856.

[6] MA J, DING S, MEI Q. Towards more practical adversarial attacks on graph neural networks[C]//Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2020: 4756-4766.

[7] LI J, ZHANG H L, HAN Z C, et al. Adversarial attack on community detection by hiding individuals[C]//Proceedings of The Web Conference 2020. New York: ACM Press, 2020: 917-927.

[8] BOJCHEVSKI A, GÜNNEMANN S. Adversarial attacks on node embeddings via graph poisoning[J]. arXiv Preprint, arXiv: 1809.01093, 2018.

[9] CHEN L, LI J, PENG J, et al. A survey of adversarial learning on graphs[J]. arXiv Preprint, arXiv: 2003.05730, 2020.

[10] XU H, MA Y, LIU H C, et al. Adversarial attacks and defenses in images, graphs and text: a review[J]. International Journal of Automation and Computing, 2020, 17(2): 151-178. 673-683.

[11] SUN Y W, WANG S H, TANG X F, et al. Adversarial attacks on graph neural networks via node injections: a hierarchical reinforcement learning approach[C]//Proceedings of The Web Conference 2020. New York: ACM Press, 2020: 673-683.

[12] WU Y T, LIU W, HU X B, et al. Parameter discrepancy hypothesis: adversarial attack for graph data[J]. Information Sciences, 2021, 577: 234-244.

[13] ZÜGNER D, GÜNNEMANN S. Adversarial attacks on graph neural networks via meta learning [J]. arXiv Preprint, arXiv: 1902.08412, 2019.

[14] 韦博成, 林金官, 解锋昌. 统计诊断[M]. 北京: 高等教育出版社, 2009.
WEI B C, LIN J G, XIE F C. Statistical diagnostics[M]. Beijing: Higher Education Press, 2009.

[15] COOK R D. Detection of influential observation in linear regression[J]. Technometrics, 1977, 19(1): 15-18.

[16] COOK R D. Influential observations in linear regression[J]. Journal of the American Statistical Association, 1979, 74(365): 169-174.

[17] COOK R D, WEISBERG S. Residuals and influence in regression[M]. New York: Chapman and Hall, 1982.

[18] 张宏坡, 程宁, 张博, 等. 一种基于熵值法的标签翻转攻击方法: CN112700081A[P]. 2021-04-23.
ZHANG H P, CHENG N, ZHANG B, et al. A label flipping attack method based on entropy: CN112700081A[P]. 2021-04-23.

[19] MUÑOZ-GONZÁLEZ L, BIGGIO B, DEMONTIS A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2017: 27-38.

[20] LIU X, SI S, ZHU X, et al. A unified framework for data poisoning attack to graph-based semi-supervised learning[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2019: 9780-9790.

[21] JIN W, LI Y, XU H, et al. Adversarial attacks and defenses on graphs: a review and empirical study[J]. arXiv Preprint, arXiv:2003.00653, 2020.

[22] 费宇, 陈飞, 喻达磊. 线性和广义线性混合模型及其统计诊断[M]. 科学出版社, 2013.
FEI Y, CHEN F, YU D L, et al. Linear and generalized linear mixed models and their statistical diagnosis[M]. Beijing: Science Press, 2013.

[23] LI Q M, WU X M, LIU H, et al. Label efficient semi-supervised learning via graph filtering[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 9574-9583.

[24] NT H, MAEHARA T. Revisiting graph neural networks: all we have is low-pass filters[J]. arXiv Preprint, arXiv:1905.09550, 2019.

[25] WU F, SOUZA A, ZHANG T, et al. Simplifying graph convolutional networks[C]//International conference on machine learning. Long Beach: PMLR, 2019: 6861-6871.

[26] WEI B C, SHIH J Q. On statistical models for regression diagnostics[J]. Annals of the Institute of Statistical Mathematics, 1994, 46(2): 267-278.

[27] HOERL A E, KENNARD R W. Ridge regression: biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[28] MARQUARDT D W. An algorithm for least-squares estimation of nonlinear parameters[J]. Journal of the Society for Industrial and Applied Mathematics, 1963, 11(2): 431-441.

[29] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93.

[30] MCCALLUM A K, NIGAM K, RENNIE J, et al. Automating the construction of Internet portals with machine learning[J]. Information Retrieval, 2000, 3(2): 127-163.

[31] ADAMIC L A, GLANCE N. The political blogosphere and the 2004 US election: divided they blog[C]//Proceedings of the 3rd International Workshop on Link Discovery. New York: ACM Press, 2005: 36-43.

[32] XU K D, CHEN H G, LIU S J, et al. Topology attack and defense for graph neural networks: an optimization perspective[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 3961-3967.

[33] 陈晋音, 黄国瀚, 张敦杰, 等. 一种面向图神经网络的图重构防御方法[J]. 计算机研究与发展, 2021, 58(5): 1075-1091.
CHEN J Y, HUANG G H, ZHANG D J, et al. GRD-GNN: graph reconstruction defense for graph neural network[J]. Journal of Computer Research and Development, 2021, 58(5): 1075-1091.

[作者简介]

吴翼腾（1992− ），男，吉林省吉林市人，信息工程大学博士生，主要研究方向为人工智能安全、对抗机器学习。

刘伟（1992− ），男，河北保定人，信息工程大学硕士生，主要研究方向为人工智能安全、自然语言处理。

于洪涛（1970− ），男，辽宁丹东人，博士，信息工程大学研究员、博士生导师，主要研究方向为大数据和人工智能。