中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

硕 士 学 位 论 文

一种基于近边界数据的模型所有权推断方法研究

Research on Model Ownership Inference Based on Near-boundary Data

论文作者 杨宗稳 指导教师 蒲凌君副教授

申请学位 工学硕士 培养单位 南开大学

学科专业 计算机科学与技术 研究方向 模型的知识产权保护

答辩委员会主席 评 阅 人

南开大学研究生院

二〇二三年四月

南开大学学位论文使用授权书

本人完全了解《南开大学关于研究生学位论文收藏和利用管理办法》关于南开大学 (简称“学校”) 研究生学位论文收藏和利用的管理规定，同意向南开大学提交本人的学位论文电子版及相应的纸质本。

本人了解南开大学拥有在《中华人民共和国著作权法》规定范围内的学位论文使用权，同意在以下几方面向学校授权。即：

1. 学校将学位论文编入《南开大学博硕士学位论文全文数据库》，并作为资料在学校图书馆等场所提供阅览，在校园网上提供论文目录检索、文摘及前 16 页的浏览等信息服务；

2. 学校可以采用影印、缩印或其他复制手段保存学位论文；学校根据规定向教育部指定的收藏和存档单位提交学位论文；

3. 非公开学位论文在解密后的使用权同公开论文。

本人承诺：本人的学位论文是在南开大学学习期间创作完成的作品，并已通过论文答辩；提交的学位论文电子版与纸质本论文的内容一致，如因不同造成不良后果由本人自负。

本人签署本授权书一份（此授权书为论文中一页），交图书馆留存。

学位论文作者暨授权人 (亲笔) 签字：_____

20　　年　　月　　日

南开大学研究生学位论文作者信息

论 文 题 目	一种基于近边界数据的模型所有权推断方法研究				
姓　　名	杨宗稳	学号	2120200439	答辩日期	
论 文 类 别	博士 ☐　学历硕士 ☑　专业学位硕士 ☐　同等学力硕士 ☐　划☑选择				
学院（单位）	计算机学院		学科/专业 (专业学位) 名称	计算机科学与技术	
联 系 电 话	13102257615		电子邮箱	yzw@mail.nankai.edu.cn	
通讯地址 (邮编)：300000					
非公开论文编号			备注		

注：本授权书适用我校授予的所有博士、硕士的学位论文。如已批准为非公开学位论文，须向图书馆提供批准通过的《南开大学研究生申请非公开学位论文审批表》复印件和“非公开学位论文标注说明”页原件。

南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：_____ 年 月 日

非公开学位论文标注说明

(本页表中填写内容须打印)

根据南开大学有关规定，非公开学位论文须经指导教师同意、作者本人申请和相关部门批准方能标注。未经批准的均为公开学位论文，公开学位论文本说明为空白。

论文题目			
申请密级	☐ 限制 (⩽2 年) ☐ 秘密 (⩽10 年) ☐ 机密 (⩽20 年)		
保密期限	20 年 月 日至 20 年 月 日		
审批表编号		**批准日期**	20 年 月 日

南开大学学位评定委员会办公室盖章 (有效)

注：限制 ★2 年 (可少于 2 年); 秘密 ★10 年 (可少于 10 年); 机密 ★20 年 (可少于 20 年)

摘要

这里输入中文摘要。

关键词： 毕业论文；模板

Abstract

This is the abstract.

Key Words: Thesis; template

目录

摘要 …… I
Abstract …… II
第一章 绪论 …… 1
第一节 研究背景与意义 1
第二节 相关研究现状 1
第三节 本文主要工作 1
第四节 本文组织架构 1
第二章 技术背景 …… 2
第一节 深度神经网络 2
第二节 对抗性攻击 2
第三节 对抗生成网络 2
第四节 深度神经网络模型窃取攻击 2
第五节 深度神经网络模型的知识产权保护 2
第六节 本章小结 2
第三章 基于对抗生成网络特征提取的近边界数据研究 …… 3
第一节 近边界对抗性样本 3
第二节 CW 生成近边界对抗性样本 3
第三节 近边界数据私有化 5
第四节 本章小结 6
第四章 基于近边界数据的模型所有权推断方法研究 …… 7
第一节 理论驱动 7
4.1.1 所有权验证局限性 7
4.1.2 利用数据推断模型所有权 7
第二节 近边界数据推断模型所有权 9
4.2.1 设计目标 10
4.2.2 方法概述 10

4.2.3 假设检验 12
第三节 本章小结 12
第五章 基于近边界数据的模型所有权推断方法分析 ····· 14
第一节 实验设置 14
第二节 生成初始近边界数据的算法选择 14
第三节 数据近边界特性的评估与扩展 16
第四节 推断模型所有权 16
第五节 微调目标分类边界的影响 16
第六节 可伸缩性扩展 18
第七节 本章小结 18
第六章 总结与展望 ····· 19
第一节 工作总结 19
第二节 工作展望 19
参考文献 ····· 20
图索引 ····· 21
表索引 ····· 22
致谢 ····· 23
个人简历 ····· 24

第一章 绪论

第一节 研究背景与意义

机器学习的发展

模型知识产权问题描述

模型知识产权相关研究

第二节 相关研究现状

研究问题

研究现状

第三节 本文主要工作

揭示[1] 现有问题，确认数据驱动推断所有权的有效性

利用对抗性样本抵御模型窃取

基于 DCGAN 生成私有数据

广泛实验验证有效性

第四节 本文组织架构

第一章

第二章

第三章

第四章

第五章

第六章

第二章　技术背景

引言

第一节　深度神经网络

神经网络相关概念

第二节　对抗性攻击

对抗性攻击相关概念

第三节　对抗生成网络

对抗生成网络相关概念

第四节　深度神经网络模型窃取攻击

深度神经网络模型窃取攻击相关概念

第五节　深度神经网络模型的知识产权保护

深度神经网络模型的知识产权保护相关概念

第六节　本章小结

小结

第三章 基于对抗生成网络特征提取的近边界数据研究

本章将从近边界对抗性样本出发，引出近边界数据，并详细阐述生成私有近边界数据的方法。

第一节 近边界对抗性样本

在第五章第三节中，本文通过大量的实验证明了近边界数据在大多数模型窃取攻击中，其近边界特征在盗窃模型中被保留。因此，近边界数据可以作为推断深度神经网络模型所有权的依据使用。下面给出近边界数据的定义：

定义 1 近边界数据。给定一个数据样本 x，一个阈值 θ, 如果数据样本 x 满足 $|g_i(x) - g_j(x)| \leq \theta$，其中 $i \neq j$ 并且 $min(g_i(x), g_j(x)) \geq \max_{k \neq i,j} g_k(x)$，$g_k(x)$ 代表数据样本 x 决策为类别 k 的概率，则数据样本 x 被称为近边界数据。

第二节 CW 生成近边界对抗性样本

尽管近边界在模型的知识产权保护中表现出显著的效果，但是自然的近边界数据在样本空间中的占比很低，甚至可以忽略，因此如何得到一定规模的近边界数据样本仍然很困难。

根据最近的一些研究[2]，对抗性样本通常被用于确定分类器的分类边界。具体而言，对抗性样本有两个分类：原始分类和目标分类。其中，原始分类是该样本不经过特殊处理的原始分类结果，目标分类是对原始样本添加微小噪声后的分类结果。如图3.1所示，对抗性样本对分类边界的跨越体现在，在视觉上对抗性样本和原始样本几乎没有差别，但是分类结果却是目标分类。

本文认为该特征可以帮助从对抗性样本中获得较多的近边界数据。因此，本文测试了几种常用的生成对抗性样本的方法，以帮助我们构建近边界数据。

***Fast Gradient Sign Method*(*FGSM*)** :FGSM[3] 是最经典的构建对抗性样本的方法之一，它是一种基于梯度生成对抗性样本的方法，属于无目标攻击方式。只需要对原始样本添加微小的扰动 η，如3.1，3.2所示，即可生成样本 x 的对抗性样本 $\tilde{x}$。

原始样本

对抗性样本

图 3.1 原始样本与对抗性样本对比

$$\eta = \epsilon \cdot sign(\bigtriangledown_x J(\theta, x, y^*)) \tag{3.1}$$

$$\tilde{x} = clip(x + \eta) \tag{3.2}$$

其中 $sign$ 是符号函数，x 表示原始样本，y^* 表示 x 的真实类别，θ 表示模型权重参数，J 表示分类器损失函数，$\bigtriangledown_x$ 表示对原始样本 x 求偏导，$clip$ 函数是将样本投射回可行数据域，ϵ 用来控制变化幅度。

FGSM 生成对抗性样本的速度非常快，但其结果非常依赖 ϵ 的选择，因此探索不同的 ϵ 是使用该方法的重点。除此之外，我们还测试了许多 FGSM 的进阶版本如 IGSM 和 RFSGM, 它们引入了迭代加入噪声和弱扰动的方法。IGSM 迭代式地使样本跨越分类边界直至成功，RFGSM 则是增加了扰动的多样性，可以更精细地生成对抗性样本。在实际结果中我们发现 FGSM 生成对抗性示例尽管速度非常快，但位于分类边界附近的数据比例却极低。IGSM 和 RFGSM 效果要比 FGSM 好，但仍认为不符合我们的期望。在大量的测试中，我们发现 CW 能够生成大量在分类边界附近的样本，具体的测试结果在第五章第二节中。

Carlini and Wagner's methods(CW) :CW[4] 方法同样是添加噪声到对抗性样本中，但其具有三种变体：CW-L_0，CW-L_2 和 CW-L_∞，不同的变体使用不同的方法来衡量噪声的大小，其中 CW-L_2 在实验中效果最为突出，因此本文使用该方法作为生成对抗性样本的选择。具体而言，CW-L_2 对于给定的初始样本迭代搜索一个小噪声使示例变为对抗性样本，这种思路使得生成的对抗性样本都集中在分类边界附近，但相应地，CW-L_2 牺牲了效率。

在这一阶段，我们只是在源模型的样本空间中挑选一部分数据作为初始样本添加小噪声，针对性地生成了目标分类对抗性样本。在此阶段源模型的训练和原始数据均不受任何影响，防御者只需要针对性的生成对抗性示例即可。然

而，近边界数据作为推断所有权的重要证据，直接生成对抗性样本也极易受到盗窃者的复制。因此，我们需要将生成的近边界数据私有化，具体操作将在第三章第三节中给出。

第三节 近边界数据私有化

由于通过生成对抗性样本的方法构建近边界数据这一步骤十分容易复现，并且现在大多数模型训练使用的数据都来源于公开数据。因此我们需要从公开的训练数据中构建自己私有化的近边界数据，以防止模型所有者的近边界数据被轻易模仿。在本文中，我们希望可以通过训练一种模型学习第三章第二节中近边界对抗性样本的特征，并以此生成新的近边界数据。这种新的数据从视觉上不一定和原始数据类似，但其原始的特征以及添加的噪声需要被学习，并根据提取到的特征生成的新样本对于源模型同样是近边界数据。因此，在本文中我们设计了一种基于 DCGAN[5] 的特征提取器，提取近边界数据的特征之后作为近边界数据生成器并将近边界数据私有化。注意生成器以，$CW\text{-}L_2$ 生成的对抗性示例作为输入，并输出私有化后的近边界数据。

具体而言，DCGAN 的结构中包括一个判定器 D 和一个生成器 G，其本质上是一个博弈过程。生成器学习样本特征生成假数据，判定器判断生成器的结果。DCGAN 的目标函数如3.3所示，是一个生成网络和判别网络的互相对抗的过程，生成器尽可能生成逼真输入样本，判别器则尽可能去判别该样本是真实样本还是假样本。

$$\begin{aligned}\min_G \max_D V(D,G) = \min_G \max_D & E_{x\sim P_{data(x)}}[logD(x)] \\ & + E_{z\sim P_{z(z)}}[log(1-D(G(z)))]\end{aligned} \tag{3.3}$$

其中 x 表示真实数据样本，z 表示用于生成样本的随机噪声，GAN 对噪声 z 的分布没有特别要求，但是常用的有高斯分布，均匀分布。注意这里的优化过程是一个交替的过程。

我们希望 DCGAN 能够学习到足够多的近边界数据特征，尝试修改其判定器的目标函数，在保留梯度的情况下将其与源模型的结果相连，得到的结果在同样的生成规模下确实优于原始 DCGAN 的生成情况。然而，考虑到在两者的效率，实际情况下生成的结果并无较大区别。

尽管构建的近边界数据已经都位于目标分类边界附近，但我们仍希望近边

界数据最大程度上靠近目标分类边界。近边界数据与目标分类边界的距离越近，推断模型所有权成功的可能性就越大。此外，生成的近边界数据虽然只被模型所有者拥有，但对于一些功能易被泛化的模型，近边界的特性仍有可能被泛化。因此，本文提出使用近边界数据微调源模型的目标分类边界。具体而言，如3.4所示，$Loss_{FT}$ 是针对目标分类边界的损失函数，其中 n 是该目标分类边界的近边界数据的数量，x_i' 是生成的近边界数据，$g_t(\cdot)$ 和 $g_s(\cdot)$ 分别表示目标分类概率和源分类概率，$Loss_{FT}$ 本质是希望近边界数据更靠近目标分类边界 $Loss_{FM}$ 是源模型的损失函数，我们设计两者交替训练微调源模型，与 DCGAN 的过程相似，是一个博弈的过程。

$$Loss_{FT} = \frac{1}{n}\sum_{i=1}^{n}(g_t(x_i') - g_s(x_i'))^2 \tag{3.4}$$

微调目标分类边界使近边界数据与源模型之间的联系更加紧密。注意，我们只微调目标分类边界，且通过交替微调尽可能减少微调对源模型的影响，微调前后源模型的精度差不超过 5%，具体的测试结果在第五章第五节中。

第四节 本章小结

本章介绍了近边界数据的特征，并详细阐述了生成私有近边界数据的方法。首先通过 CW 方法，生成近边界对抗性样本，然后利用对抗生成网络的学习特征，将近边界数据私有化，最后通过自定义损失函数交替微调源模型，在几乎不损失模型精度的情况下，使其近边界数据更加靠近分类边界。

第四章　基于近边界数据的模型所有权推断方法研究

本章将从数据集推断引出近边界数据推断模型所有权的方法。

第一节　理论驱动

4.1.1　所有权验证局限性

现有的模型知识产权保护措施着重于被动的防御，只考虑针对模型修改的抗攻击性。模型所有者将水印嵌入训练好的模型或从其中提取抽象的模型知识作为指纹（称为源模型），当怀疑一个模型（称为可疑模型）的知识来自于源模型，模型所有者可以利用水印或指纹被动地从外部验证模型所有权。大多数工作基于这样的思路，设计不同的水印和指纹用于在源模型被盗窃后验证模型所有权，但这并不具有较强的鲁棒性。模型水印的缺陷例如对源模型性能和功能的影响，嵌入水印引起的额外代价都是研究水印工作的关键点。模型指纹目的是提取代表模型知识的固有特征，相较于水印指纹不会对源模型产生影响，但是指纹是脆弱的因为模型知识是易被修改的，所有的指纹方法都试图找到可以承受某些修改攻击的强鲁棒性指纹。

本文的目标集中在水印和指纹另一个亟待解决的问题歧义攻击上，歧义攻击不关心如何去除水印和指纹以通过模型所有权验证，而是伪造额外的水印和指纹混淆所有权验证。具体来说，盗窃者对源模型嵌入新的水印或提取其他的指纹使本来的保护措施无效。歧义攻击对现有的深度神经网络模型的知识产权保护方法构成了严重威胁，在传统的数字水印领域中有研究表明，鲁棒性的水印可能不一定会验证所有权，除非水印方案是不可逆的[6]。在本文中，我们认为通过验证可疑模型是否具有源模型特定的水印或指纹来讨论盗窃行为是不充分的，特别是出现歧义攻击时，因此我们提出推断模型所有权而不是验证。这种方法的灵感来自于数据集推断[1] 提出的所有权决策，我们将在4.1.2中具体讨论。

4.1.2　利用数据推断模型所有权

数据集推断做了一个假设：源模型的知识来自于训练数据集。无论盗窃模型是直接攻击源模型还是其副产品，盗窃模型的知识是源模型中包含的知识。

如果原始训练数据集是私有的，模型所有者就比对手拥有强大优势，源模型在原始训练数据中的性能要远远优于其他数据集。因此，通过统计测试与估计多个数据点到决策边界的距离相结合，可以得到模型所有权归属。

源模型的知识被传播到盗窃模型使得所有盗窃模型都必须包含源模型训练数据集中的直接或间接信息。原始训练数据的私有性作为源模型的标识可以用来识别盗窃模型，只需要证明可疑模型和源模型都经过共同的私有数据集训练（不一定完全相同）。此过程和传统的验证模型所有权不同，通过私有数据集推断得到的是一个所有权决策，其中决策的最大者被认为拥有所有权。传统的模型所有权验证是从模型中提取水印或指纹进行匹配从而验证，这里涉及到了歧义攻击导致的验证冲突。可以发现数据集推断得到的是一个“最”的概念，因此可以有效避免歧义攻击。因此，我们指出推断所有权将会成为未来模型知识产权保护技术的主要方向。

我们的工作受到数据集推理验证模型所有权的启发，我们提出了数据驱动推断所有权代替验证所有权。我们认为所有权推断在有效证明所有权归属问题的同时，可以解决验证冲突问题。除此之外，数据驱动的推断所有权意味着只和输入输出相关，我们的方法既可以在白盒环境也可以在黑盒环境下工作。

但是数据集推断具有以下**局限性**：

1) 使用数据集推理的前提是原始训练数据不被盗窃者得到，公开数据集不能被用于训练源模型。然而，在大多数现实情况中，只有很少一部分工作会构造私有数据集用于训练模型，甚至这部分工作的应用点很狭窄，这意味着被盗窃的风险较小。因此，依赖于私有数据集的数据集推理方法在实际应用中使用范围很小，不能被大幅度推广使用；
2) 数据集推理方法的核心思想是源模型的功能在训练数据上的效果优于其他数据，但存在模型的功能可能相似，而结构和训练数据都不同的情况。因此该方法可能会导致误导。Li[7] 等人验证了此限制，结果表明该方法产生的结果值得怀疑。

我们指出，利用数据推断所有权的想法需要解决以上问题，因此我们提出构造私有化近边界数据作为推断依据，并利用近边界数据靠近决策边界的特性处理模型功能相似引起的误导。这是因为即使模型功能相似，但决策边界不可能完全相同。

第二节 近边界数据推断模型所有权

在本文中，我们提出了近边界数据，一种分布在分类边界附近的特殊样本。模型指纹[2] 使用对抗性样本抽象地反映模型分类边界，同一组对抗性样本的输入，其引起的决策模式的变化可以用于比较模型知识的相似性，但这种方法是脆弱的，对模型的任意操作都有可能破坏这种特性。因此，我们不直接比较决策模式的变化，它是不可信任的，而是比较对抗性样本与决策边界的距离。有意思的是大多数对抗性样本都是位于决策边界附近的，也就是说，它们与决策边界的距离很近。对抗性样本的这种性质被我们所利用并构造近边界数据，经过测试我们发现绝大多数的模型窃取方法都无法改变这种结果，即使样本分类被影响，其仍然位于分类边界附近。近边界数据背后的意义是如果被用于所有权验证如果两个模型的决策模式相似，参与训练的近边界数据一定可以反映出来。受到这个的启发，将近边界数据作为水印验证所有权是传统的思路，即使不会对模型的精度造成影响，然而这样的水印是脆弱的，很难抵御歧义攻击，因此我们提出由近边界数据驱动的所有权推断方法，其思想是构造私有的近边界数据，当验证一个模型的所有权时，模型所有者和盗窃者分别提供各自的私有近边界数据，距离分类边界最近的被推断获得所有权。这个方法的主要思想如图4.1所示。

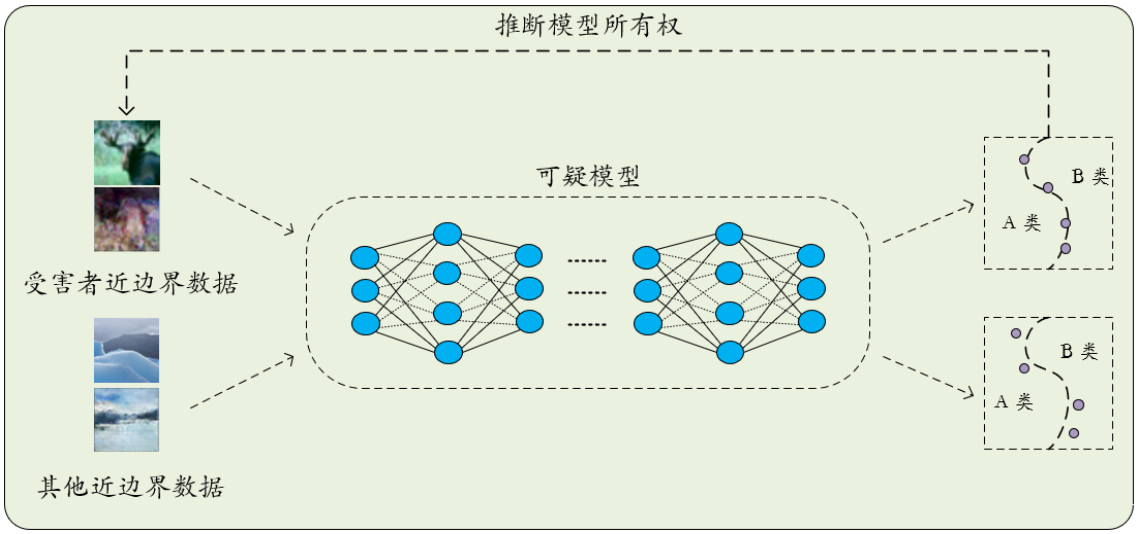

图 4.1 近边界数据推断所有权

4.2.1 设计目标

依据现有的工作，我们的方法在模型训练后进行部署，且在黑盒环境中推断所有权。我们的方法不关注模型盗窃的过程，目的是准确推断受害者所有权和识别可疑模型的盗窃行为。现在大多数所有权验证技术都是黑盒模型环境，因为模型所有者和攻击者通常不会提供完整模型。我们提出的方法仅利用模型提供的外部 API，获取近边界数据的决策结果，从而推断模型所有权。在通常的假设中，存在一个官方的仲裁机构，当对任一模型产生所有权怀疑时，受害者和可疑对手可以向机构提出申请并提供各自的私有化近边界数据，并通过我们的方法推断所有权。注意无论在白盒和黑盒的环境中，我们的方法均可以产生效果。

为了实现推断模型所有权，本文提出的方法的设计目标是：

1) **精确性:** 推断模型所有权的方法不应该影响模型的性能，模型的最大可接受测试精度下降不超过 5%。
2) **数据近边界性:** 如果可疑模型与源模型相同或来自源模型，则根据源模型构造的私有近边界数据在推断模型所有权中距离指定的分类边界最近。
3) **鲁棒性:** 近边界数据应该对常见的模型修改（如模型微调、剪枝和有损压缩）具有鲁棒性。
4) **不可见性:** 敌手无法获得私有的近边界数据，也无法在视觉上观察到近边界数据的部署。
5) **有效性:** 通过近边界数据推断模型所有权应能有效地计算距离边界数据，并通过对比全部近边界数据的决策结果确定可疑模型是盗窃模型。

4.2.2 方法概述

为了实现以上目标，本文提出了一种基于近边界数据的模型所有权推断方法。

问题定义：我们定义了一个深度神经网络（DNN）分类器 G 作为源模型，给定一个原始训练集 D，假设该源模型是一个 n-类的 DNN 分类器，分类器的输出层为 softmax 层或其他决策层，决策函数 $g_j(x)$ 表示数据样本 x 被分到第 j 类的概率，其中 $j = 1,2,..,n$。$Z_1,Z_2,..,Z_n$ 表示模型分类器的全部决策函数输出，其结果可作为分类边界的依据被我们使用，因此

$$g_j(x) = \frac{exp(Z_j(x))}{\sum_{i=1}^{n} exp(Z_i(x))} \tag{4.1}$$

其中，数据样本 x 的标签 y 被推断拥有最大概率的类别，例如 $y = arg\max_j g_j(x) = arg\max_j Z_j(x)$。

定义 2 分类边界。分类器的分类边界是一个抽象的概念，我们无法直接描述它。因此我们使用分类器的决策结果来反映分类边界。

通常来说，寻找位于分类边界上的数据点采用重复随机采样数据点的方法，具体地如果数据点满足上述定义则数据点在分类边界上。然而，简单的重复采样可能需要大量的时间消耗，甚至无法找到这样的数据点们。为了解决这样的问题，我们在第三章中讨论了如何构造位于分类边界上或其附近的的数据点，且将其私有化的过程。

基于第三章的讨论，我们提出构造近边界数据推断模型的所有权，而不是验证所有权。具体而言，如图4.2所示，我们的方法包括三个主要阶段：

1) 从数据集样本中生成对抗性样本；
2) 训练生成对抗模型生成私有化的近边界数据；
3) 加入近边界数据微调源模型。

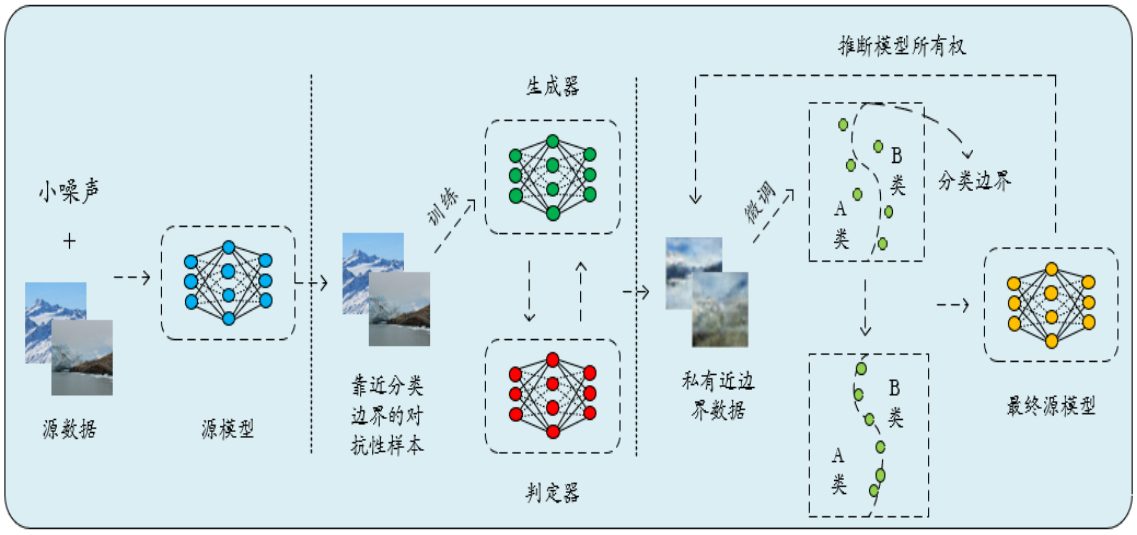

图 4.2 方法整体流程图

4.2.3 假设检验

根据第四章第一节讨论的结果，本文认为过去的验证模型所有权的思路具有较大的局限性，大多数研究无法抵御歧义攻击。因此，我们提出了推断模型所有权的想法，这是一种“最”的思路。在现实情况中，我们假设存在第三方仲裁机构，并约定目标分类边界，被盗窃者向第三方机构提出仲裁并提供近边界数据，盗窃者同样需要提供相应的近边界数据，第三方机构分别计算目标分类边界距离，本文认为持有最靠近目标分类边界的近边界数据所有者将获得模型所有权。注意由于近边界数据通常是一组数据，所以应该根据统计的结果来看。在实践中，我们计算了不同规模的近边界数据组在源模型, 盗窃模型和不相关模型上与分类边界的距离，并设计了一种基于假设检验的方法来表现推断置信度。

假设检验：我们假设事件 C 是模型所有者提供的私有近边界数据在怀疑模型上的计算结果，事件 C_S 表示盗窃者提供的近边界数据在怀疑模型上的计算结果，或模型所有者提供的私有近边界数据在无关模型上的计算结果。本文计算假设 $H_0 : \mu > \mu_S (H_1 : \mu \leq \mu_S)$ 的 p 值，以及差异大小 $\Delta\mu = \mu_S - \mu$，$\Delta\mu$ 越大，推断可信度越高。如果 p 值低于预定义的置信度评分 α，则拒绝 H_0，并称正在测试的模型是被盗模型。我们重复 30 次统计性实验以提高可信度。

第三节 本章小结

本章从所有权验证的局限性出发，引出了数据集推断，然后详细介绍了近边界数据推断模型所有权的设计目标以及具体的方法流程。

Algorithm 1 InitialDistribution

Input: $Nodes, kFrags, Set$

Output: $targetnodes$

$Nodes \leftarrow$ the neighboring online nodes
$kFrags \leftarrow$ the N re-encryption keys the node has generated
$Set \leftarrow$ the set of the nodes that have got the kFrag
$flag \leftarrow 0$
for $kFrag$ in $kFrags$ **do**
    SELECTNODE($Nodes, kFrag, Set, underload$)
    if $flag == 0$ **then**
        SELECTNODE($Nodes, kFrag, Set, normal$)
    end if
    if $flag == 0$ **then**
        SELECTNODE($Nodes, kFrag, Set, overload$)
    end if
end for

function SELECTNODE($Nodes, kFrag, Set, State$)
    for $node$ in $Nodes$ **do**
        if $node's\ state$ is $State$ and $node \notin Set$ **then**
            $Send(kFrag)$
            $Set = Set \cup node$
            if $Size(Set) == Size(Map)$ **then**
                $Clear(Set)$
            end if
            $flag \leftarrow 1$
            $Break$
        end if
    end for
end function

第五章 基于近边界数据的模型所有权推断方法分析

我们在开源数据集 CIFAR-10[8]，Heritage[9]，Intel_image[10] 上面进行实验，并选择 ResNet18 作为评估的源模型，VGG11 作为对照的无关模型。本文使用的模型均在开源的预训练模型上进行训练。

被盗模型：我们设置了常见的几种模型盗窃方法，包括模型微调，模型剪枝（不同的剪枝率）和模型蒸馏，并在源模型的基础上得到被盗模型。

第一节 实验设置

本文实验利用 CIFAR-10,Heritage 和 Intel_image 三种数据集训练 ResNet18, 训练过程中 Adam 优化器并将学习率 (Learning rate)，迭代轮次 (Epoch) 和每批次大小 (Batch size) 分别设置为 0.0001,200 和 64。蒸馏模型实验选择从 Resnet18 蒸馏至 VGG11，蒸馏时将蒸馏温度设置为 20 并且教师模型比例 α=0.7，训练轮次是 20。初始近边界数据生成采用 *CW*-L_2 算法，实验中选择有目标的生成方式，且学习率，迭代次数和二分搜索次数分别设置为 0.001,1000 和 6，其他参数为默认值。私有近边界数据生成器采用 DCGAN 的基础结构，训练过程使用 Adam 优化器且将学习率，训练轮次和每批次大小分别设置为 0.0002，8000 和 64。注意本发明最后微调源模型阶段需要交替使用源模型损失函数和微调目标边界的损失函数来微调源模型，具体设置为 10 个轮次交替一次且交替次数最多为 10 次。

第二节 生成初始近边界数据的算法选择

本小节将对第三章第二节中提出的 FGSM，IGSM，RFGSM 和 CW-L_2 进行测试，我们均使用原作者发布的实现。FGSM，IGSM，RFGSM 中均有一个用于界定噪声 ϵ 的参数，且 IGSM 和 RFGSM 还包含一个重要的参数 α 用来表示迭代次数。我们进行大量的实验探索选择合适的参数用于与 CW-L_2 进行比较。此外，CW-L_2 的实验设置如第一节所示。如表5.1所示，CW-L_2 生成的对抗性样例与目标分类边界的平均距离远比其他算法小。因此，本文使用该算法作为初始近边界数据生成算法。

表 5.1 不同对抗性样本生成算法生成的数据与目标分类边界的平均距离

数据集	FGSM	IGSM	RFGSM	CW-L_2
CIFAR-10	0.557	0.430	0.418	0.066
	0.461	0.419	0.373	0.103
	0.586	0.369	0.356	0.112
Heritage	0.347	0.356	0.314	0.014
	0.277	0.340	0.281	0.016
	0.348	0.332	0.276	0.010
Intel_image	0.522	0.447	0.353	0.088
	0.475	0.506	0.387	0.122
	0.468	0.402	0.428	0.127

第三节 数据近边界特性的评估与扩展

第四节 推断模型所有权

表 5.2 推断模型所有权

数据集	攻击方法	分类边界 1		分类边界 2		分类边界 3		分类边界 4		分类边界 5	
		$\Delta\mu$	p 值	$\Delta\mu$	p 值	$\Delta\mu$	p 值	$\Delta\mu$	p 值	$\Delta\mu$	p 值
CIFAR-10	源模型	0.913	10^{-6}	0.954	10^{-6}	0.927	10^{-5}	0.967	10^{-5}	0.958	10^{-5}
	模型微调	0.718	10^{-5}	0.745	10^{-6}	0.698	10^{-5}	0.692	10^{-4}	0.729	10^{-5}
	剪枝 10%	0.572	10^{-5}	0.487	10^{-5}	0.458	10^{-5}	0.533	10^{-4}	0.512	10^{-4}
	剪枝 30%	0.537	10^{-4}	0.497	10^{-4}	0.401	10^{-3}	0.428	10^{-4}	0.587	10^{-4}
	剪枝 50%	0.545	10^{-4}	0.614	10^{-4}	0.506	10^{-3}	0.570	10^{-4}	0.484	10^{-3}
	知识蒸馏	0.372	10^{-3}	0.297	10^{-3}	0.288	10^{-3}	0.308	10^{-3}	0.340	10^{-3}
Heritage	源模型	0.876	10^{-5}	0.845	10^{-5}	0.859	10^{-4}	0.801	10^{-4}	0.837	10^{-5}
	模型微调	0.815	10^{-5}	0.792	10^{-4}	0.824	10^{-4}	0.833	10^{-4}	0.784	10^{-4}
	剪枝 10%	0.530	10^{-4}	0.535	10^{-3}	0.508	10^{-4}	0.486	10^{-3}	0.471	10^{-3}
	剪枝 30%	0.491	10^{-3}	0.452	10^{-3}	0.469	10^{-4}	0.470	10^{-3}	0.427	10^{-4}
	剪枝 50%	0.502	10^{-3}	0.517	10^{-3}	0.434	10^{-3}	0.451	10^{-3}	0.490	10^{-3}
	知识蒸馏	0.329	10^{-3}	0.365	10^{-2}	0.238	10^{-3}	0.310	10^{-3}	0.274	10^{-3}
Intel_image	源模型	0.859	10^{-5}	0.896	10^{-4}	0.872	10^{-4}	0.899	10^{-4}	0.914	10^{-4}
	模型微调	0.717	10^{-5}	0.784	10^{-4}	0.752	10^{-4}	0.791	10^{-3}	0.709	10^{-4}
	剪枝 10%	0.451	10^{-4}	0.522	10^{-4}	0.539	10^{-3}	0.472	10^{-3}	0.438	10^{-4}
	剪枝 30%	0.407	10^{-4}	0.415	10^{-4}	0.346	10^{-3}	0.382	10^{-3}	0.395	10^{-3}
	剪枝 50%	0.370	10^{-3}	0.395	10^{-3}	0.327	10^{-3}	0.360	10^{-3}	0.458	10^{-3}
	知识蒸馏	0.336	10^{-2}	0.395	10^{-3}	0.360	10^{-2}	0.308	10^{-3}	0.287	10^{-2}

第五节 微调目标分类边界的影响

表 5.3 微调分类边界对模型的影响

数据集	微调前准确率	微调后准确率
CIFAR-10	0.886	0.873
Heritage	0.879	0.856
Intel_image	0.794	0.786

表 5.4 CIFAR-10 上不同对抗性样本生成算法生成的数据与目标分类边界的平均距离

数据集 (准确率)	分类边界	数据数量	微调后准确率
CIFAR-10 (0.886)	分类边界 1	64	0.873
		128	0.862
		256	0.862
		512	0.854
	分类边界 2	64	0.871
		128	0.870
		256	0.860
		512	0.844
	分类边界 3	64	0.871
		128	0.868
		256	0.858
		512	0.856
	分类边界 4	64	0.873
		128	0.873
		256	0.866
		512	0.862
	分类边界 5	64	0.876
		128	0.866
		256	0.868
		512	0.861

表 5.5 Heritage 上不同对抗性样本生成算法生成的数据与目标分类边界的平均距离

数据集 (准确率)	分类边界	数据数量	微调后准确率
Heritage (0.879)	分类边界 1	64	0.856
		128	0.825
		256	0.830
		512	0.797
	分类边界 2	64	0.823
		128	0.839
		256	0.841
		512	0.779
	分类边界 3	64	0.848
		128	0.826
		256	0.779
		512	0.791
	分类边界 4	64	0.819
		128	0.795
		256	0.803
		512	0.783
	分类边界 5	64	0.843
		128	0.851
		256	0.776
		512	0.774

表 5.6 Intel_image 上不同对抗性样本生成算法生成的数据与目标分类边界的平均距离

数据集 (准确率)	分类边界	数据数量	微调后准确率
Intel_image (0.794)	分类边界 1	64	0.755
		128	0.769
		256	0.756
		512	0.779
	分类边界 2	64	0.770
		128	0.741
		256	0.768
		512	0.777
	分类边界 3	64	0.781
		128	0.753
		256	0.764
		512	0.752
	分类边界 4	64	0.787
		128	0.751
		256	0.762
		512	0.747
	分类边界 5	64	0.786
		128	0.764
		256	0.761
		512	0.765

第六节 可伸缩性扩展

第七节 本章小结

第六章　总结与展望

第一节　工作总结

第二节　工作展望

参考文献

[1] MAINI P, YAGHINI M, PAPERNOT N. Dataset inference: Ownership resolution in machine learning. [J]. ArXiv preprint arXiv:2104.10706, 2021.

[2] CAO X, JIA J, GONG N Z. IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. [C] // Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. [S.l.]: [s.n.], 2021: 14–25.

[3] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples. [J]. ArXiv preprint arXiv:1412.6572, 2014.

[4] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks. [C] // 2017 ieee symposium on security and privacy (sp). Ieee. [S.l.]: [s.n.], 2017: 39–57.

[5] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks. [J]. ArXiv preprint arXiv:1511.06434, 2015.

[6] FAN L, NG K W, CHAN C S. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. [J]. Advances in neural information processing systems, 2019, 32.

[7] LAO Y, ZHAO W, YANG P, et al. Deepauth: A dnn authentication framework by model-unique and fragile signature embedding. [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. 9. [S.l.]: [s.n.], 2022: 9595–9603.

[8] KRIZHEVSKY A, HINTON G, et al. Learning multiple layers of features from tiny images. [J]. 2009.

[9] LLAMAS J. Architectural Heritage Elements image Dataset. [EB/OL]. 2017, Feb 20. `https://datahub.io/dataset/architectural-heritage-elements-image-dataset`.

[10] BANSAL P. Intel Image Classification. [EB/OL]. 2019. `https://www.kaggle.com/datasets/puneet6060/intel-image-classification`.

图索引

3.1 原始样本与对抗性样本对比 4

4.1 近边界数据推断所有权 9
4.2 方法整体流程图 11

表索引

5.1 不同对抗性样本生成算法生成的数据与目标分类边界的平均距离 . 15
5.2 推断模型所有权 16
5.3 微调分类边界对模型的影响 16
5.4 CIFAR-10 上不同对抗性样本生成算法生成的数据与目标分类边界的平均距离 17
5.5 Heritage 上不同对抗性样本生成算法生成的数据与目标分类边界的平均距离 18

致谢

感谢您使用本模板。

个人简历

xxx，出生于 yyyy 年 mm 月 dd 日。在 20yy 年毕业于 xx 大学 XX 专业并获得 xx 士学位。于 20xx 年至今在南开大学就读 xxx 研究生。

研究生期间发表论文:

- 周恩来. 周恩来选集 [M]. 人民出版社, 1980.
- 周恩来. 周恩来外交文选 [M]. 中央文献出版社, 1990.