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

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 本文主要架构

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

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。

原始样本

对抗性样本

图 3.1　原始样本与对抗性样本对比

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 $\tilde{x}$。

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

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

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，$clip$ 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。

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：$CW\text{-}L_0$，$CW\text{-}L_2$ 和 $CW\text{-}L_\infty$，不同的变体使用不同的方法来衡量噪声的大小，其中 $CW\text{-}L_2$ 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，$CW\text{-}L_2$ 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，$CW\text{-}L_2$ 牺牲了效率。

第三节　近边界数据私有化

第四节　本章小结

第四章　基于近边界数据的模型所有权推断方法研究

第一节　数据集推断

第二节　近边界数据推断模型所有权

第三节　本章小结

第五章　基于近边界数据的模型所有权推断方法分析

第一节　实验设置
第二节　生成初始近边界数据的算法选择
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
第五节　微调目标分类边界的影响
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.

致谢

感谢您使用本模板。

图索引

表索引

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.