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： 知识产权保护；所有权推断；近边界数据；深度神经网络；生成对抗网络

Abstract

This is the abstract.

Key Words: Intellectual property protection; Ownership inference; Near-boundary data; Deep neural network; Generative adversarial network

目录

第一章 绪论

第一节 研究背景与意义

近年来，科技迅速发展，计算资源日益丰富，计算能力得到显著提升，我们正在进入人工智能 (Artificial Intelligence, AI)[0] 的时代。随着互联网的的快速发展，产生了海量的数据，得益于深度神经网络 (Deep Neural Network, DNN)[0] 强大的数据处理能力，DNN 已经成为应用最广泛的人工智能方法之一。自 DNN 在计算机视觉[0]，语音识别[0]，自然语言处理[0] 等方面上突破性应用以来，使用 DNN 的应用数量呈爆炸式增长。这些 DNN 应用被应用到从自动驾驶[0] 到癌症检测[0] 到玩复杂的游戏[0] 等无数应用中。并且在许多这些领域中，DNN 已经能够超越人类的准确性。

DNN 在许多领域取得巨大成功，为人类社会生活带来极大便利的同时，也带来了非常严重的侵犯知识产权 (Intellectual Property, IP) 问题。训练一个大型的高性能的 DNN 模型都离不开该领域专家的专业知识、规模巨大的数据集以及大量的训练时间和强大的计算资源，具体体现在以下三个方面：

1) 人力资源，对于不同场景不同目的的 DNN 模型，需要不同领域的知识，包含对模型结构的设计分析、模型参数的调试校验等；
2) 大量的训练数据，模型所有者要在特定领域训练出一个高性能的模型，通常需要该领域大量的数据，并且需要覆盖到应用场景中的各种情况，这些数据的获取和整理本身就需要昂贵的价格，有的领域的数据还涉及到隐私性问题；
3) 昂贵的计算资源和大量的训练时间，DNN 模型的规模越来越大，层数越来越多，需要的训练时间也越多，并且训练过程中也需要越来越多的计算资源支持，才能对网络权重等进行精确的调整，这些都是巨额的经济成本。如 GPT-3[0]，包含了 1750 亿参数，仅训练成本需花费 460 万美元以上。

所以高性能 DNN 模型是模型所有者智慧的结晶，同时需要高额的经济开销，模型所有者享有 DNN 模型的知识产权[0]。

模型所有者出于学术目的将 DNN 模型放到开源社区上。或者，使用机器学习即服务 (Machine Learning as a Service, MLaaS)[0] 的商业模式，即 MLaaS 平台通过训练好的 DNN 模型来向用户提供应用程序接口 (Application Programming Interface, API)[0]，用户可以通过支付一定的费用来使用 API。或者，训练好的 DNN 模型将成为像我们日常商品一样的消费品，它们由不同的公司或个人进行训练，由不同的供应商分发，最终由用户消费。如图1.1所示，这样的方式极大的方便了科研工作者和一般的消费者，但是不法分子却可以以比模型所有者低很多的成本复制一个替代模型，用于自己盈利。

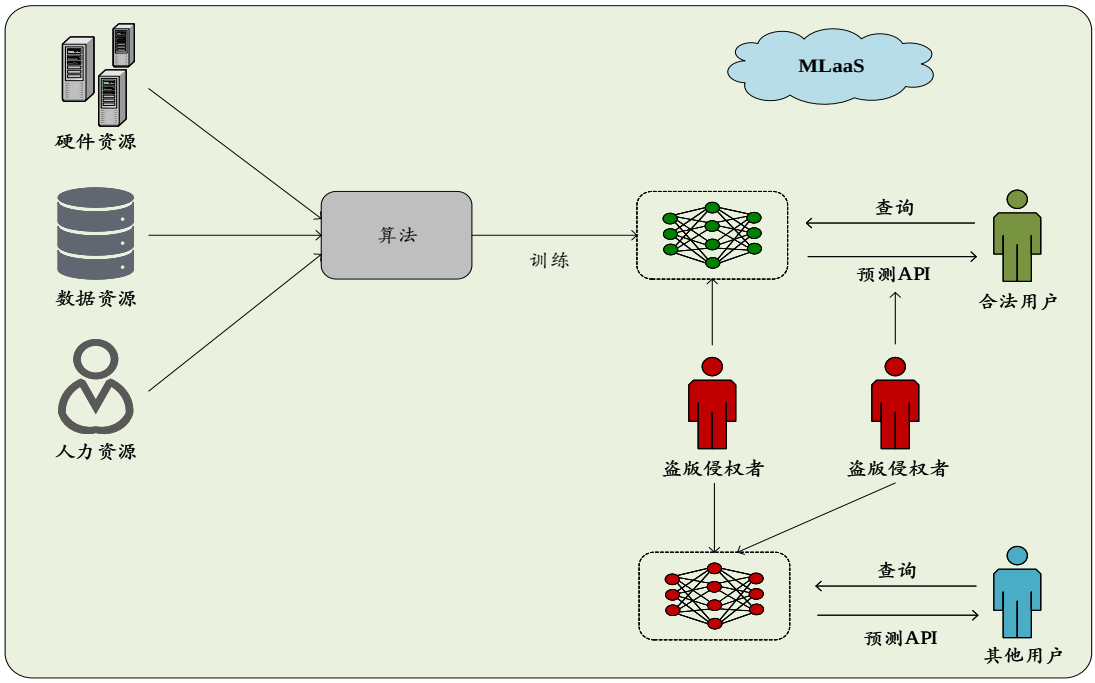

图 1.1 DNN 模型服务和盗窃示意图

所以如何在训练和部署时保护 DNN 模型所有者的知识产权是 AI 领域亟待解决的问题。

第二节 相关研究现状

作为一种数字产品，DNN 模型不仅凝结了设计者的智慧，还需要消耗大量的训练数据和昂贵的计算资源。近年来，拥有先进的的模型带来的工业优势已经被人们广泛认可，这开始激发一些不法分子窃取这些模型的攻击[0]。现在可以明确的是，DNN 模型将在未来的 IT 发展中发挥核心作用，因此保护这些模型

的必要性显得更加突出。1994 年，Van Schyndel 等人[0] 第一次提出数字水印的概念，将标记隐蔽的嵌入到如音频、视频等数字内容中，来识别其所有权，具体来说，版权所有者通过显示此类标记的存在可以证明其对内容的所有权。DNN 模型也是一种数字产品，所以，许多研究者从数字媒体水印得到启发，从而设计模型水印和模型指纹用于解决 DNN 模型的所有权问题。

模型水印是解决 DNN 模型知识产权问题的主要方式之一，Uchida 等人[0] 在 2017 年首次提出了在 DNN 模型中嵌入水印的通用框架。该方法是一种白盒的模式，通过训练时使用正则化器，并且这种正则化在参数中引入了所需要的统计偏差来作为嵌入的水印。模型所有者清楚模型内部的细节，并且可以提取嵌入的水印，以此来作为模型所有权的依据[0]。Fan 等人[0] 提出了一种在 DNN 模型中嵌入数字护照的方案，嵌入数字护照的要点是设计和训练 DNN 模型，使得在伪造护照的情况下 DNN 的推断性能显著下降，而真正的护照可以通过查找预定签名来验证。Chen 等人[0] 提出了一种新颖的端到端框架，该框架同时依赖于用户和模型，它需要为每一个用户分配一个代码向量，并将该信息嵌入到可训练权重的的概率密度函数中，同时保持模型的准确性。不同于白盒的模式，另一种黑盒的模式，可以在不访问模型内部的情况下，通过特定的输入输出来验证模型的所有权。Le 等人[0] 提出了一种零比特水印算法，该算法标记模型的操作本身，稍微调整它的决策边界，来使特定的查询得到特定的输出。在减少模型性能损失的同时，该算法可以远程操作 DNN 或 API 服务，通过少量的查询提取水印。Zhang 等人[0] 提出了一种水印植入方法，将水印注入 DNN 模型。通过扩展 DNN 的内在泛化和记忆能力，使得模型能够在训练时学习特意制作的水印，然后在推断时激活预先指定的预测。Adi 等人[0] 提出了利用模型的后门机制当作 DNN 模型水印。后门通常是 DNN 将输入预测为错误的标签，虽然在大多数情况下这是不可取的，但是却可以将为 DNN 模型制作水印的任务转化为设计后门的任务。这些黑盒的方法利用对抗性样本作为触发集，或者使用一组特定的训练样本，然后根据特殊样本的输出来提取水印。因此黑盒的方法在所有权验证中不需要访问模型的权重参数。Rouhani 等人[0] 提出了一种端到端的 IP 保护框架 DeepSigns，可以在 DNN 模型中插入连贯的数字水印。DeepSigns 引入了一种通用水印方法，不同于直接将水印信息嵌入到模型的权重中，DeepSigns 将任意 N 位字符串嵌入到各层激活集的概率密度函数中，这意味着水印信息嵌入在 DNN 的动态内容中，并且只能通过特定的输入数据来触发，并且对权重矩阵

等静态属性没有影响。但是 DNN 模型水印的嵌入步骤总是会对原始进行修改。具体来说，白盒水印修改模型内部，比如模型权重，激活函数等，而黑盒水印通过特殊的训练调整模型来指定特定的输出。这些修改将会影响 DNN 模型在原始任务上的性能。

模型指纹是解决 DNN 模型知识产权问题的又一主流方法。不同与模型水印，模型指纹不需要对模型本身进行修改，而是利用模型本身来寻找和提取一些独特的特征作为模型指纹，一般来说，模型指纹不会影响模型的性能。Zhao 等人[0] 提出了一种新的 DNN 模型指纹技术，该技术旨在提取模型本身的固有特征，而不是嵌入额外的水印。具体来说，该方法选择一组专门设计的对抗性样本作为模型指纹特征，称为对抗性标记，相比于其他不相关的模型，它可以更好的转移到从原始模型派生出的模型上。与 Zhao 等人[0] 的方法类似，Lukas 等人[0] 提出了一种用于 DNN 分类器的指纹识别方法，该方法从源模型中提取一组输入，以便只有源模型的派生模型在此类输入的分类上与源模型一致。这些输入是可转移对抗性样本的一个子类，它们的目标标签会从源模型转移到其派生模型上。Cao 等人[0] 针对 DNN 分类器提出了一种名叫 IPGuard 的指纹方法，该方法的关键是 DNN 分类器可以由其分类边界唯一的表示。基于这一原理，IPGuard 在模型所有者的 DNN 分类器分类边界上提取了一些数据点，并使用它们对分类器进行指纹识别，如果 DNN 分类器对大多数指纹数据点预测相同的标签，那么该模型被认为是模型所有者分类器的盗版。Li 等人[0] 提出了一种适用于生成对抗网络 (Generative Adversarial Network, GAN)[0] 知识产权保护的指纹识别方案。该方案从目标 GAN 和分类器构建了一个复合深度学习模型，然后从该复合模型中生成隐蔽的指纹样本，并将其注册到分类器中进行有效的所有权验证。Dong 等人[0] 针对 DNN 水印和指纹容易受到最抗性训练攻击，不适用于多出口 DNN 模型的 IP 验证的问题，提出了一种根据推理时间而不是推理预测的结果来为多出口模型建立指纹的新方法。

一般的模型窃取攻击涉及到模型的修改，主要包括模型微调，模型剪枝，模型压缩等。模型微调通常用于迁移学习，可以重新调整模型以更改模型参数，同时保持模型的性能。通过微调现有的模型，可以派生出许多功能相似的模型。模型剪枝是是部署 DNN 模型的常见方法，通过参数修剪来减少 DNN 的内存需求和计算开销，而盗窃者可能会使用修剪来删除水印或指纹。模型压缩中常见的是知识蒸馏，通过将大模型中的知识蒸馏到小模型中，可以显著降低模型的训

练成本，内存需求和计算开销，同时达到与大模型接近的性能。研究[0] 表明甚至不需要原始训练数据就可以直接利用 API 蒸馏模型，因此蒸馏常被用来派生模型。

虽然模型水印和模型指纹在保护模型知识产权方面已经取得了很大的进展，但是无论是水印还是指纹都容易受到歧义攻击[0]，歧义攻击是指通过为 DNN 模型伪造其他水印或指纹来对所有权验证产生干扰。直觉上，如果模型盗窃者可以在水印模型上嵌入第二个水印或者提取第二个指纹，那么该模型的知识产权归属存在巨大的歧义。

第三节　本文主要工作

为了解决 DNN 模型的知识产权问题，本文提出了近边界数据，一种分布在分类边界附近的特殊样本。模型指纹[0] 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。受这个特性的启发，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，这样的水印也是脆弱的，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。

本文的主要贡献如下：

1) 揭示了当前所有权验证方案的脆弱性并确认了数据驱动推断所有权的有效性。
2) 提出了利用对抗性样本构造近边界数据以抵御模型窃取攻击。
3) 设计了基于 DCGAN 的近边界数据生成器，并通过该生成器构造了私有化的近边界数据，提出了一种损失函数用以微调源模型的目标分类边界，

增加推断所有权的置信度。

4) 在 ResNet18 上进行了广泛的实验，实验结果证明了近边界数据在推断模型所有权上的显著效果。

第四节 本文组织架构

本文对模型的近边界数据进行了研究，并提出了生成私有近边界数据的方法以及一种基于近边界数据的模型所有权推断方法。全文共分为六个章节，每个章节的主要内容如下：

第一章：绪论。本章首先介绍了 DNN 模型在当今时代的广泛应用和训练的昂贵成本，引出了保护 DNN 模型知识产权的必要性和重大意义，然后介绍了模型水印和模型指纹两种保护方法的研究现状，并针对相关研究存在的问题提出了本文的研究内容，最后简要说明了各个章节的内容安排。

第二章：技术背景。本章主要介绍了深度神经网络的结构和相关概念，对抗性攻击和生成对抗网络的原理，常见的模型窃取攻击方式和模型知识产权保护方法。

第三章：基于生成对抗网络特征提取的近边界数据研究。本章首先给出了近边界数据的概念，然后详细介绍了选择 CW 算法生成对抗性样本的原因，最后给出了本文基于生成对抗网络生成私有化近边界数据的流程和微调源模型的设计。

第四章：基于近边界数据的模型所有权推断方法研究。本章首先阐述了所有权验证和数据集推断的局限性，然后提出了本文方法的设计目标，并详细说明了本文的方法流程，最后利用假设检验来比对结果。

第五章：基于近边界数据的模型所有权推断方法分析。本章在 resnet18 的基础上，对数据的近边界性，微调模型分类边界的影响，近边界推断模型所有权的有效性以及近边界数量的伸缩性做了详细的实验，证明了本文提出方法在推断模型所有权时的有效性。

第六章：总结与展望。本章总结了本文提出的方法，并针对可能的改进提出了未来工作的展望。

第二章　技术背景

引言

第一节　深度神经网络及相关术语

人工神经网络是一种类似于人类大脑生物神经系统的信息处理模型，它由许多相互连接的神经元（网络中的节点）组成，这些神经元都可以向其他神经元发送信号。一般的神经网络由输入层，隐藏层和输出层组成，如图2.1所示，如果一个神经网络有多个隐藏层，那么这个神经网络就被称为深度神经网络。DNN的隐藏层一般由卷积层，池化层，全连接层，Dropout 层和 Softmax 层构成，数据输入输入层后，会经过每一层，每层提取的抽象特征会作为下一层的输入，最终由输出层输出。

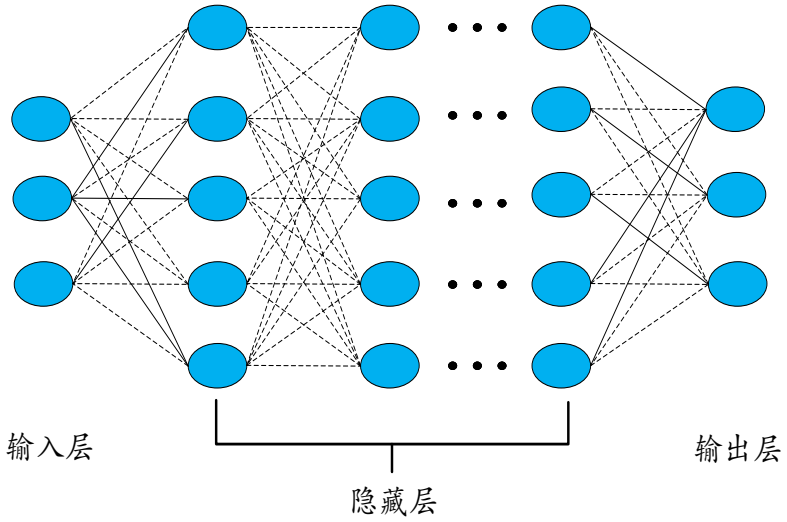

图 2.1　深度神经网络结构图

DNN 可以看作是将一组输入变量转化为一组输出变量的非线性数学函数。每个神经元都有对应的权重和偏置参数，控制着输入的精确转化，这些参数在反向传播的过程中，通过损失函数和梯度下降算法来更新。确定这些参数的过程称为 DNN 模型的学习或者训练，并且需要大量的计算资源，然而权重一旦确

定，DNN 模型就可以快速的处理相似类型的新数据，识别并提取海量数据中的复杂特征。

以下是本文中涉及到 DNN 知识产权保护领域中的相关术语：

1) 源模型。源模型也称作目标模型，是指模型所有者在私有或公共数据集上，消耗大量计算资源和人力资源训练出的高性能 DNN 模型，可能因学术研究放置在开源社区，或者作为商用给用户提供远程 API。
2) 可疑模型。可疑模型是指该模型可能是通过模型窃取攻击方法从源模型派生出来的模型，判断一个可疑模型是否是从源模型派生是模型知识产权保护领域的主要目标。
3) 白盒模式。白盒模式是指能够获得 DNN 模型的所有知识，包括训练集，训练方式，模型参数，模型结构等。
4) 黑盒模式。黑盒模式指不清楚模型内部参数，但可以通过模型提供的 API 获得指定输入的输出。

第二节 对抗性攻击

2.2.1 对抗性样本

对抗性样本的概念是 Szegedy 等人[0] 提出的。这篇文章中指出，通常情况下，一个良好性能的 DNN 模型具备很好的泛化能力，对输入的随机微小扰动具有鲁棒性，因此小扰动不应该改变图像的预测类别。然而，对图像添加特定的非随机扰动，使得损失函数的值增大，可以任意改变 DNN 模型的预测结果。这种人类肉眼上难以察觉但可以使模型输出错误类别的样本称为对抗性样本。

用 $f: R^m \rightarrow 1,2,...,n$ 表示将一张图片映射为 n 个标签的 DNN 分类器，对一个正常样本 $x \in R^m$ 以及一个错误标签 l，目标是找到一个最小的扰动 δ，使得分类器将样本 x 错误分类为 l，如式2.1所示：

$$\begin{aligned} &min \parallel \delta \parallel_2, \\ &s.t.\ f(x+\delta) = l,\ x+\delta \in [0,1]^m \end{aligned} \tag{2.1}$$

其中叠加了扰动的 $x+\delta$ 即为一个对抗性样本。2.1这种方式通常用在黑盒的场景下，仅根据 DNN 分类器的输出进行扰动 δ 的调整。

在白盒场景下，由于知道模型的所有知识，可以根据这些信息来寻找对抗性样本，通常利用 DNN 分类器的损失函数来寻找对抗性样本。

用 $f: R^m \to 1,2,...,n$ 表示将一张图片映射为 n 个标签的 DNN 分类器，对一个正常样本 $x \in R^m$ 以及它对应的正确标签 y，目标是找到一个足够小小的扰动 $\delta: \delta \leq \gamma$，使得加上扰动后的样本输入 DNN 模型后，损失函数 L 达到最大值，如式2.2所示：

$$\delta = arg\max_{\delta \leq \gamma} L(f(\theta, x+\delta), y) \tag{2.2}$$

其中 θ 是分类器 f 的参数，$x+\delta$ 是一个扰动后的对抗性样本。

2.2.2 对抗性攻击的类别

对抗性攻击技术是指生成对抗性样本的方法，不同的方法生成对抗性样本的效率，质量也不相同。根据方式的不同，可以分为以下几类：

1) 白盒攻击与黑盒攻击。白盒攻击指敌手知道 DNN 模型的参数和内部结构等信息，利用这些信息发起的攻击。黑盒攻击指敌手仅根据模型的输入输出来发起攻击。
2) 有目标攻击和无目标攻击。有目标攻击指对抗性样本的预测类别为敌手指定的类别，例如将一张牛的图片识别为羊，而不能是其他类别，常采取的方式是向各个方向搜索扰动来最大化 DNN 模型预测特定类上的可能性。无目标攻击指添加扰动来改变原始预测类别，对具体分类类别不做要求。通常来说有两种攻击方式，一种是最小化 DNN 模型预测正确类的可能性，一种是进行多次不同类别的的有目标攻击，然后在多个对抗性样本中选取扰动最小的。
3) 单步攻击和迭代攻击。单步攻击指通过一次添加扰动生成对抗性样本，迭代攻击指通过多次迭代添加微小扰动来生成对抗性样本。通常来说迭代攻击的成功率较高，但是相应的算法复杂度更高，效率较低。
4) 个体攻击和普适性攻击。个体攻击指针对每个样本都需要重新生成扰动，普适性攻击指找到一个通用的扰动，对数据集中的一类数据都叠加该扰动，普适性攻击效率较高，但是寻找通用扰动的难度较大。

第三节 生成对抗网络

Goodfellow 等人[0] 第一次提出了生成对抗网络 (Generative Adversarial Network, GAN)，是一种通过生成模型实现无监督学习的特殊方法。GAN 由一个生成器和一个判别器构成，训练是一个相互博弈的过程，如图2.2所示，首先随机

噪声作为生成器的输入，生成器生成和真实图片维度一致的图像，使用原始图片和生成图片分别输入判定器，训练判定器区分它们的能力，再训练生成器，使之尽可能接近真实图片，通过迭代的交替训练，最终生成器生成的图片和原始图片在空间分布上基本一致，判定器判定生成图片和原始图片为真的概率均为$\frac{1}{2}$，也就是无法区分生成图片和原始图片。

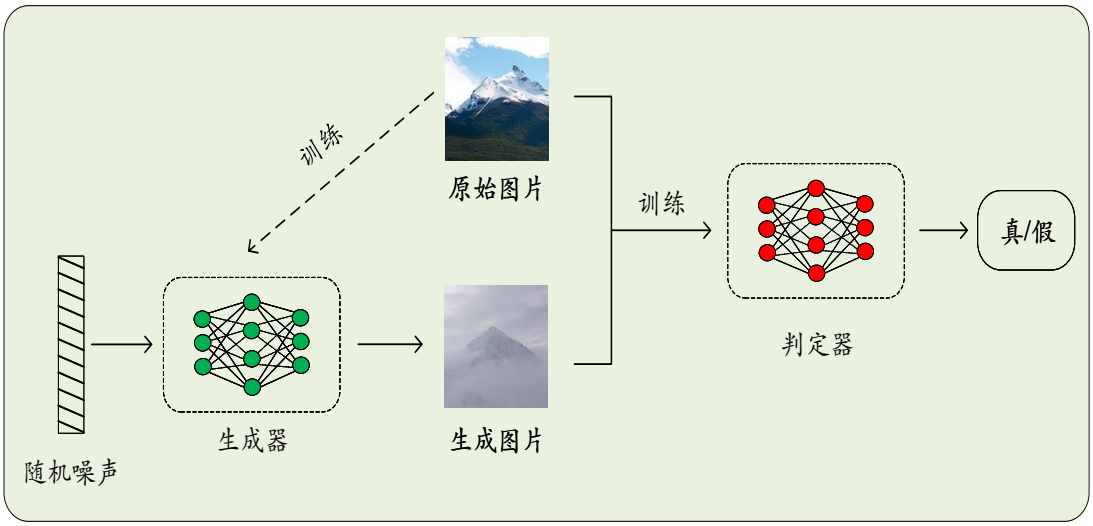

图 2.2 生成对抗网络结构图

具体而言，生成器 G 和判别器 D 可视为博弈中的双方，当训练 GAN 模型时，生成器 G 和判别器 D 通过更新各自的参数使损失达到最小，经过不断迭代优化，最后 G 和 D 达到纳什均衡。GAN 的目标函数如式2.3所示，对于原始图片 x，判别器希望 $D(x)$ 越大越好，对应于式中的 $maxD$，对于生成图片 $G(T)$，生成器希望 $D(G(T))$ 越大越好，即 $log(1-D(G(T))$ 越小，对应于式中的 $minG$，所以 GAN 的目标函数由两个目标构成。

$$\begin{aligned}\min_G \max_D V(D,G) = \min_G \max_D E_{x\sim P_{data}(x)}[logD(x)] \\ + E_{T\sim P_T(T)}[log(1-D(G(T)))]\end{aligned} \tag{2.3}$$

其中 x 表示原始图片，T 表示用于生成样本的随机噪声，GAN 对噪声 T 的分布没有特别要求，但是常用的有高斯分布，均匀分布，E 表示数学期望。

第四节 深度神经网络的模型窃取攻击

自 DNN 在各个领域取得巨大成功以来，针对 DNN 模型的攻击就层出不穷，按照攻击方式的不同，可以分为以下三类[0]：(1) 模型修改攻击。指常见的模型修改，主要包括包括模型微调，模型剪枝，模型压缩，模型再训练等方式。(2) 删除攻击。指攻击者试图逃避水印或指纹的检测，主要包括删除攻击，篡改攻击，逆向工程攻击等方式。(3) 主动攻击。指攻击者主动攻击和强攻击，主要包括歧义攻击，水印和指纹覆盖攻击，查询修改攻击等方式。

2.4.1 模型修改攻击

模型盗窃者在盗窃 DNN 模型后，通常会对 DNN 模型进行修改或者压缩，然后部署模型作为 MLaas 来非法盈利。模型修改主要包括：

1) 模型微调。微调通常用于迁移学习中，包括在源模型的基础上，根据自己定制的任务，继续训练模型，使得 DNN 模型在保持性能的同时修改内部的参数。模型微调可以从源模型派生出非常多的模型。由于内部参数发生改变，水印等可能也会随之变化，因此这对水印的鲁棒性是一个考验。
2) 模型剪枝。由于 DNN 模型通常内存占用多，计算开销大，因此模型剪枝是在小型设备上部署 DNN 模型的常用方法。但是模型盗窃者可能会利用剪枝来删除水印，因此有效的水印技术应该能够抵御由模型剪枝引起的参数变化。
3) 模型压缩。模型压缩可以显著降低 DNN 模型的内存需求和计算开销，常用的方法是知识蒸馏，通过将大型模型包含的知识转移到小模型上来达到模型压缩的目的。
4) 模型再训练[0]。模型再训练是一种很直接的方法，这样能尽可能的去除或者减少原有水印的影响，相应的，这种攻击方式成本也比较高。

2.4.2 删除攻击

目前大部分 DNN 模型的知识产权保护工作专注于水印对 DNN 模型被修改时的鲁棒性，而很少考虑水印或指纹本身受到的攻击。删除攻击主要包括：

1) 删除攻击[0]。攻击者试图修改模型以删除原有的水印。

2) 篡改攻击。攻击者知道 DNN 模型中存在水印，试图窜改模型来删除原有的水印和指纹特征。
3) 逆向工程攻击[0]。如果攻击者知道并可以获得原始训练数据，可能会直接对内部参数进行逆向工程。

Shafieinejad 等人[0] 研究了 DNN 中基于后门的水印方法的移除攻击，表明攻击者可以仅依靠公共数据集删除水印，而不用访问训练集和模型参数。还提出了一种检验水印的方法，表明基于后门的水印不够安全，无法保持水印的隐藏。

2.4.3 主动攻击

除了被动的攻击方式，攻击者还可能对 DNN 模型发动更强的主动攻击。主动攻击主要包括：

1) 歧义攻击。歧义攻击指在 DNN 模型上伪造额外的水印来混淆所有权的验证。研究表明，除非采取不可逆的水印方案，否者即使是鲁棒性的水印，也不一定能验证模型的所有权[0]。
2) 水印覆盖攻击[0]。即使攻击者不知道具体的私有水印信息，但他知道模型水印嵌入的方法，就可能通过在 DNN 模型中嵌入新的水印来覆盖原有的水印，从未破坏原有的水印使其不可读。
3) 查询修改攻击。攻击者修改查询结果来使得水印验证过程无效。一个典型的方式是攻击者获得 DNN 模型并部署为 MLaaS 后，会主动检测一个查询是否为水印验证查询，从而修改或者屏蔽该查询，使水印验证无效。

第五节 深度神经网络模型的知识产权保护

训练一个高性能 DNN 模型需要该领域专家的先验知识来设计模型结构，大量的训练数据和昂贵的计算资源和漫长的训练时间，因此，训练后的 DNN 模型属于模型所有者的知识产权。得益于 DNN 模型在各个领域的高效应用，许多不法分子开始偷盗，复制和修改这些模型来提供服务盈利。为了保护 DNN 模型的知识产权，许多学者受数字水印的启发，使用模型水印和模型指纹来验证 DNN 模型知识产权。

2.5.1 模型水印

模型水印是第一种被提出的保护 DNN 模型知识产权的方法，根据水印嵌入方式和提取方式的不同，主要分为白盒水印和黑盒水印。

在白盒场景下，模型所有者可以利用模型的全部知识构造水印，这些知识包括训练数据集，训练方法，模型内部权重参数和结构。Kuribayashi 等人[kuribayashi2020deepwatermark] 提出一种基于全连接层权重的可量化水印嵌入方法，通过在训练中改变参数，可以量化水印的影响，从而保证嵌入水印引起的变化较小。不同于基于权重的方法，基于内部结构的水印方法抵抗模型修改的鲁棒性更强。可以在 DNN 模型中添加一个额外的护照[0]，比如在模型卷积层后面添加一个额外的护照层，以此来作为数字签名，这种方式还可以解决模型受到歧义攻击的问题。

在黑盒场景下，模型所有者不知道可疑模型的内部结构和权重参数等，只能通过 API 进行访问。一般而言，是通过构造特殊的触发集来实现的，主要有以下几种方式：

1) 通过更改样本标签构造触发集，将原始样本标签更改为模型所有者指定的与原始内容不符合的标签，这样仅修改标签不做任何其他修改的水印方法称为零位水印。
2) 通过在原始样本中嵌入额外水印信息和更改标签构造触发集，这样可以在模型输出中嵌入模型所有者的版权信息。
3) 通过添加新的样本构造触发集，这样的方式对模型的精度影响较大，一般通过模型微调最大限度的减少新样本对模型决策的影响。

2.5.2 模型指纹

模型指纹一般是利用模型本身来寻找和提取一些固有的特征来作为指纹。相较于模型水印的方法，模型指纹一般不对模型进行修改，因此不会影响模型的精度。一般来说，可以选择靠近决策边界的对抗性样本作为模型的指纹特征，来验证模型所有权。模型指纹分为指纹生成和指纹验证两个阶段。

(1) 模型指纹生成

(2) 模型指纹验证

第六节 本章小结

小结

第三章 基于生成对抗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**????**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[0]，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[0] 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，如式3.1，3.2所示，即可生成样本 x 的对抗性样本 $\tilde{x}$。

(a) 原始样本　　(b) 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**????**中。

Carlini and Wagner's methods(CW) :CW[0] 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[0] 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。

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

其中 x 表示真实数据样本，T 表示用于生成样本的随机噪声，GAN 对噪声 T 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，如表3.1所示，微调前后源模型的精度差不超过 5%，因此，微调对于源模型的性能影响十分微小，甚至可以被忽略，但却有效提高了最后的所有权推断效果。更多的准确度测试结果可以在**????**中找到。

表 3.1 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

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能验证所有权，除非水印方案是不可逆的[0]。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[0] 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 等人[0] 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[0] 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。

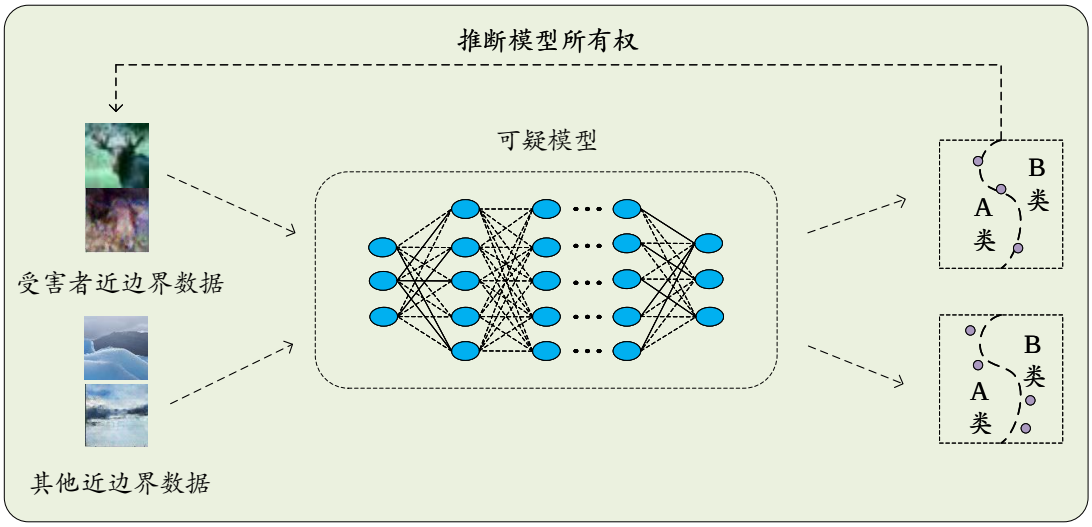

图 4.1 近边界数据推断所有权

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。

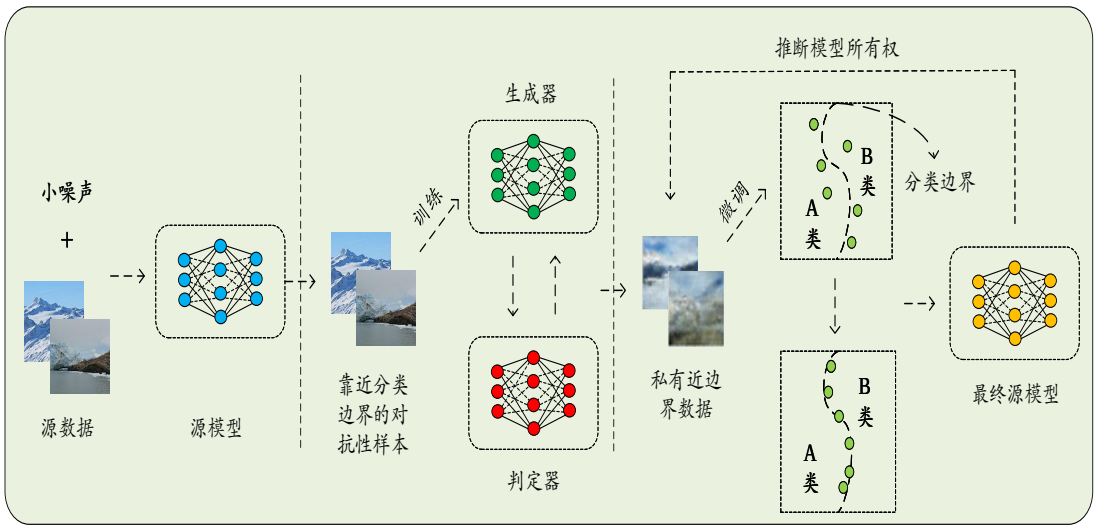

图 4.2 方法整体流程图

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 总结与展望

第一节 工作总结

在本文中，我们讨论了以往研究中验证模型所有权的局限性，提出了用推断模型所有权代替验证。我们认为可以从数据驱动的角度抵御模型盗窃，即如果数据在源模型上存在一种可衡量的特性，那么这种特性也会被被盗模型所继承。因此，我们构建了一种有趣的近边界数据用以推断所有权，并设计了使用 CW-L_2 迭代添加小噪声的方法生成对抗性样本，这是初始的近边界数据。我们训练了一种基于 DCGAN 的近边界生数据成器用以将近边界数据私有化和扩展，实验测试了生成器能够显著地学习近边界数据的特征并生成新的数据。最后，我们设计了新的损失函数微调源模型的分类边界，得到了最终版本的源模型和近边界数据。我们在 CIFAR-10，Heritage 和 Intel_image 数据集上进行评估，实验证明了我们的方法可以高置信度地推断模型所有权。

第二节 工作展望

参考文献

[0] WINSTON P H. Artificial intelligence. [M]. [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 1984.

[0] SZE V, CHEN Y.-H, YANG T.-J, et al. Efficient processing of deep neural networks: A tutorial and survey. [J]. Proceedings of the IEEE, 2017, 105 (12): 2295–2329.

[0] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks. [C] // Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer. [S.l.]: [s.n.], 2016: 630–645.

[0] CORTES C, LAWARENCE N, LEE D, et al. Advances in neural information processing systems 28. [C] // Proceedings of the 29th Annual Conference on Neural Information Processing Systems. [S.l.]: [s.n.], 2015.

[0] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. [J]. ArXiv preprint arXiv:1409.1556, 2014.

[0] NASSIF A B, SHAHIN I, ATTILI I, et al. Speech recognition using deep neural networks: A systematic review. [J]. IEEE access, 2019, 7: 19143–19165.

[0] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch. [J]. Journal of machine learning research, 2011, 12 (ARTICLE): 2493–2537.

[0] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. [J]. ArXiv preprint arXiv:1609.08144, 2016.

[0] XIONG W, DROPPO J, HUANG X, et al. Achieving human parity in conversational speech recognition. [J]. ArXiv preprint arXiv:1610.05256, 2016.

[0] CHEN C, SEFF A, KORNHAUSER A, et al. Deepdriving: Learning affordance for direct perception in autonomous driving. [C] // Proceedings of the IEEE international conference on computer vision. [S.l.]: [s.n.], 2015: 2722–2730.

[0] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks. [J]. Nature, 2017, 542 (7639): 115–118.

[0] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge. [J]. Nature, 2017, 550 (7676): 354–359.

[0] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners. [J]. Advances in neural information processing systems, 2020, 33: 1877–1901.

[0] CHEN H, ROUHANI B D, FAN X, et al. Performance comparison of contemporary DNN watermarking techniques. [J]. ArXiv preprint arXiv:1811.03713, 2018.

[0] DARVISH ROUHANI B, CHEN H, KOUSHANFAR F. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. [C] // Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. [S.l.]: [s.n.], 2019: 485–497.

[0] RIBEIRO M, GROLINGER K, CAPRETZ M A. Mlaas: Machine learning as a service. [C] // 2015 IEEE 14th international conference on machine learning and applications (ICMLA). IEEE. [S.l.]: [s.n.], 2015: 896–902.

[0] OFOEDA J, BOATENG R, EFFAH J. Application programming interface (API) research: A review of the past to inform the future. [J]. International Journal of Enterprise Information Systems (IJEIS), 2019, 15 (3): 76–95.

[0] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing Machine Learning Models via Prediction APIs. [C] // USENIX security symposium. Vol. 16. [S.l.]: [s.n.], 2016: 601–618.

[0] DUDDU V, SAMANTA D, RAO D V, et al. Stealing neural networks via timing side channels. [J]. ArXiv preprint arXiv:1812.11720, 2018.

[0] VAN SCHYNDEL R G, TIRKEL A Z, OSBORNE C F. A digital watermark. [C] // Proceedings of 1st international conference on image processing. Vol. 2. IEEE. [S.l.]: [s.n.], 1994: 86–90.

[0] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks. [C] // Proceedings of the 2017 ACM on international conference on multimedia retrieval. [S.l.]: [s.n.], 2017: 269–277.

[0] NAGAI Y, UCHIDA Y, SAKAZAWA S, et al. Digital watermarking for deep neural networks. [J]. International Journal of Multimedia Information Retrieval, 2018, 7: 3–16.

[0] 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.

[0] CHEN H, ROHANI B D, KOUSHANFAR F. Deepmarks: A digital fingerprinting framework for deep neural networks. [J]. ArXiv preprint arXiv:1804.03648, 2018.

[0] LE MERRER E, PEREZ P, TRÉDAN G. Adversarial frontier stitching for remote neural network watermarking. [J]. Neural Computing and Applications, 2020, 32: 9233–9244.

[0] ZHANG J, GU Z, JANG J, et al. Protecting intellectual property of deep neural networks with watermarking. [C] // Proceedings of the 2018 on Asia Conference on Computer and Communications Security. [S.l.]: [s.n.], 2018: 159–172.

[0] ADI Y, BAUM C, CISSE M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. [C] // 27th {USENIX} Security Symposium ({USENIX} Security 18). [S.l.]: [s.n.], 2018: 1615–1631.

[0] ROUHANI B D, CHEN H, KOUSHANFAR F. Deepsigns: A generic watermarking framework for ip protection of deep learning models. [J]. ArXiv preprint arXiv:1804.00750, 2018.

[0] ZHAO J, HU Q, LIU G, et al. AFA: Adversarial fingerprinting authentication for deep neural networks. [J]. Computer Communications, 2020, 150: 488–497.

[0] LUKAS N, ZHANG Y, KERSCHBAUM F. Deep neural network fingerprinting by conferrable adversarial examples. [J]. ArXiv preprint arXiv:1912.00888, 2019.

[0] 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.

[0] LI G, XU G, QIU H, et al. A Novel Verifiable Fingerprinting Scheme for Generative Adversarial Networks. [J]. ArXiv preprint arXiv:2106.11760, 2021.

[0] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks. [J]. Communications of the ACM, 2020, 63 (11): 139–144.

[0] DONG T, QIU H, ZHANG T, et al. Fingerprinting Multi-exit Deep Neural Network Models via Inference Time. [J]. ArXiv preprint arXiv:2110.03175, 2021.

[0] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network. [J]. ArXiv preprint arXiv:1503.02531, 2015.

[0] LI H, WENGER E, SHAN S, et al. Piracy resistant watermarks for deep neural networks. [J]. ArXiv preprint arXiv:1910.01226, 2019.

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

[0] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets. [J]. Stat, 2014, 1050: 10.

[0] XUE M, ZHANG Y, WANG J, et al. Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations. [J]. IEEE Transactions on Artificial Intelligence, 2021, 3 (6): 908–923.

[0] NAMBA R, SAKUMA J. Robust watermarking of neural network with exponential weighting. [C] // Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security. [S.l.]: [s.n.], 2019: 228–240.

[0] SHAFIEINEJAD M, LUKAS N, WANG J, et al. On the robustness of backdoor-based watermarking in deep neural networks. [C] // Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. [S.l.]: [s.n.], 2021: 177–188.

[0] CHEN H, ROUHANI B D, KOUSHANFAR F. Blackmarks: Blackbox multibit watermarking for deep neural networks. [J]. ArXiv preprint arXiv:1904.00344, 2019.

[0] CHEN H, ROUHANI B D, FU C, et al. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models. [C] // Proceedings of the 2019 on International Conference on Multimedia Retrieval. [S.l.]: [s.n.], 2019: 105–113.

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

[0] 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.

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

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

[0] 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.

图索引

表索引

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.